LÍVIA GOMES TORRES

OPTIMIZING GWS AND GWAS IN CROP BREEDING

Thesis submitted to the Universidade Federal de Viçosa, in partial fulfilment of the requirements of the Graduate Program in Genetics and Breeding for degree of *Doctor Scientiae*.

Adviser: Marcos Deon Vilela de Resende

Co-adviser: Camila Ferreira Azevedo
Eder Jorge de Oliveira

VIÇOSA – MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
T693o
2021

Torres, Lívia Gomes, 1990-
Optimizing GWS and GWAS in crop breeding / Lívia Gomes Torres. – Viçosa, MG, 2021.
118 f. : il. (algumas color.) ; 29 cm.

Orientador: Marcos Deon Vilela de Resende.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Mandioca - Melhoramento genético. 2. Soja - Melhoramento genético. 3. Genômica. 4. Predição. 5. Mapeamento genômico vegetal. I. Universidade Federal de Viçosa. Departamento de Agronomia. Programa de Pós-Graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 631.52

LÍVIA GOMES TORRES

OPTIMIZING GWS AND GWAS IN CROP BREEDING

Thesis submitted to the Universidade Federal de Viçosa, in partial fulfilment of the requirements of the Graduate Program in Genetics and Breeding for degree of *Doctor Scientiae*.

APPROVED: January 14th, 2021.

Assent:

Lívia Gomes Torres
Author

MResende
Marcos Deon Vilela Resende
Adviser

To my grandmother Adalgisa (in memoriam), my parents Robledo and Elmira, Robledo Filho, Vanelle and Cecília.

ACKNOWLEDGEMENTS

I thank God for the protection, renewal of hope and for always allowing me to meet good people wherever I go.

I would like to thank the Universidade Federal de Viçosa and the Graduate Program of Genetics and Breeding for the opportunity, contribution to my training and impact in my life.

To CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for the financial support to carry out this work.

To Embrapa Mandioca e Fruticultura and the Nextgen Cassava project, through a grant to Cornell University by the UK's Foreign, Commonwealth & Development Office (FCDO) and the Bill & Melinda Gates Foundation.

To my advisor Dr. Marcos Deon Vilela de Resende, for his supervision, support, opportunities, friendship, the professional example he is, discussions about our studies and conversations about life in general.

To Dr. Camila Ferreira Azevedo, for her co-supervision, support, contribution, friendship, suggestions and teaching.

To Dr. Eder Jorge de Oliveira for his co-supervision, opportunities, discussions and suggestions about our studies.

To Dr. Fabyano Fonseca e Silva, for his support, suggestions, teaching, contribution, friendship and encouragement.

To Dr. François Belzile and all members of his laboratory at the Institut de Biologie Intégrative et des Systèmes at Université Laval for the support, discussions about science, for making me feel welcome in their research group and for the good moments shared.

To Dr. Lukas Mueller and all members of his laboratory at the Boyce Thompson Institute at Cornell University, in my short stay in Ithaca, for the support, kindness and for making me feel welcome in their research group.

To Dr. Guillaume Bauchet and Alex Ogbonna for the support, contribution, discussions about our study and friendship.

To Dr. Ana Carolina Campana Nascimento, Dr. Camila Ferreira Azevedo and Dr. Moysés Nascimento, for their friendship, teaching and the opportunity to participate in their study group LICAE (Laboratory of Computational Intelligence and Statistical Learning).

To Dr. Rodrigo Oliveira de Lima for his contribution to my training and friendship.

To Dr. Andrei Caíque Pires Nunes**, Dr. Camila Ferreira Azevedo, Dr. Fabyano Fonseca e Silva** and Dr. Felipe Lopes da Silva, members of the evaluation committee, for having accepted to participate, the discussions and suggestions.

I am also thankful for all teachers/professors with whom I learned from childhood to graduate studies, especially those who inspired and encouraged me to always continue studying.

To my beloved parents Robledo and Elmira, for being my greatest supporters, examples and best friends; for filling my life with love and this love transforming the way I see the world every day. Words are not enough to translate my feelings and how much I admire them, each in their own unique way.

To my brother Robledo Filho, for his friendship, support, kindness, niceness, advice, encouragement and example.

To my sister-in-law Vanelle, for her friendship, support and encouragement.

To my niece/goddaughter Cecília, for all the love and joy, the smile in her eyes brings light to the grayest of days.

To my grandmother Adalgisa (*in memoriam*), for everything she represents to me and for the example of life and wisdom. Eternally grateful for all of our moments together, her lessons of kindness and strength.

To all my family members, aunts, uncles and cousins, for their care and support.

To the staff of the Graduate Program in Genetics and Breeding, Marco Tulio and Odilon, for their niceness and for always doing their best to help the graduate students.

To my friends since childhood, Coluni, undergraduate and graduate studies, laboratories/internships: MIPD-UFV, Bioinformática, Biocafé, Programa Milho-UFV, LICAE (Laboratory of Computational Intelligence and Statistical Learning), and those I did during my academic exchanges through Guelph, Quebec City and Ithaca, I thank you for the all the moments, conversations and ideas shared.

To everyone that, in some way, supported and encouraged me on this journey.

THANK YOU!

BIOGRAPHY

Lívia Gomes Torres was born on June 26, 1990, in Viçosa, Minas Gerais, Brazil.

In March 2006, she started her studies at the Universidade Federal de Viçosa, located in Viçosa, Minas Gerais, Brazil, attending high school at Colégio de Aplicação – Coluni, graduating in December 2008.

In March 2009, she began her undergraduate studies in Agronomy at the Universidade Federal de Viçosa, where she graduated in July 2015. Since the beginning of her Bachelor's degree, she was involved in research activities, and she was granted by FAPEMIG funding agency with undergraduate research scholarships from 2010 to 2013. During her undergraduate degree, she awarded a scholarship and participated of the Science Without Borders Program (CNPq funding agency), studying at the University of Guelph, Ontario, Canada, from September 2013 to December 2014.

In August 2015, she started her Master's degree (CNPq scholarship) studies in Genetics and Breeding at the Universidade Federal de Viçosa, undergoing the final dissertation defense exams to obtain the title of *Magister Scientiae* in February 2017, entitled: "Characterization of tropical maize inbred lines for root morphology and nutritional efficiency in contrasting nitrogen conditions".

In March 2017, she started her Doctoral degree (CNPq scholarship) studies in Genetics and Breeding at the Universidade Federal de Viçosa. During her doctorate, she participated in the PDSE Program (CAPES funding agency) and did part of her training at Université Laval, Quebec, Canada, from November 2018 to November 2019. She is a member of the LICAE (Laboratory of Computational Intelligence and Statistical Learning), a study group from the Statistics Department – UFV, since 2017. Her final thesis defense exams to obtain the title of *Doctor Scientiae* is set to January 2021, entitled: "Optimizing GWS and GWAS in Crop Breeding".

RESUMO

TORRES, Lívia Gomes, D.Sc., Universidade Federal de Viçosa, janeiro de 2021. **Otimização da GWS e da GWAS no melhoramento de plantas.** Orientador: Marcos Deon Vilela de Resende. Coorientadores: Camila Ferreira Azevedo e Eder Jorge de Oliveira.

Em um cenário de avanço dos métodos de genotipagem, juntamente com a redução dos custos por amostra genotipada, as informações obtidas em nível molecular têm sido cada vez mais utilizadas em análises genéticas para acelerar e tornar mais efetiva a seleção de materiais genéticos superiores, por meio da predição genômica (GP). A GP conduz a predições mais acuradas do mérito de indivíduos para caracteres de interesse. Estudos de associação genômica ampla (GWAS) também tem sido muito utilizados para a identificação de variantes causais, visando a detecção de associações significativas entre marcadores genéticos e caracteres de interesse. No primeiro capítulo desta tese, predições baseadas em informações de pedigree, marcadores moleculares e fenótipos para caracteres produtivos de clones de mandioca provenientes de cruzamentos biparentais foram realizadas separando-se o conjunto de dados em número de estádios de um programa de melhoramento (avaliação clonal – CET, ensaios preliminares de produtividade – PYT, ensaios avançados de produtividade – AYT, e ensaios uniformes de produtividade – UYT): um estádio (CET), dois estádios (CET e PYT), três estádios (CET, PYT e AYT) e quatro estádios (CET, PYT, AYT e UYT). Os resultados indicaram potencial satisfatório para utilização da predição genômica em mandioca, especialmente para seleção precoce, sendo assim, poupando-se recursos. No segundo capítulo, foi avaliada a viabilidade da utilização de predições genômicas entre países (Brasil e Nigéria) como uma estratégia para a escolha de clones a serem envolvidos em intercâmbio de germoplasma, em um estudo de caso com o caractere teor de cianeto de hidrogênio (HCN) em mandioca. Além disso, objetivos adicionais foram fornecer uma avaliação da estrutura da população para o conjunto de dados envolvendo os dois países e estimar parâmetros genéticos baseados nos polimorfismos de nucleotídeo único (SNPs) e em uma abordagem de haplótipos. As comparações das estimativas dos GEBVs foram feitas considerando a situação hipotética de não haver caracterização fenotípica para um conjunto de clones para um determinado instituto de pesquisa/país e poder ser necessário utilizar-se apenas dos efeitos dos marcadores que foram estimados com dados de outro instituto de pesquisa/país para estimar o GEBV de seus clones. As predições envolvendo o conjunto de dados dos dois países apresentaram maiores acurácias do que as de cada fonte individualmente. A predição genômica entre países parece ter uso potencial, ao menos sob

o cenário do presente estudo. A correlação entre GEBVs preditos com população de treinamento da Embrapa e IITA foi de 0,55 para o germoplasma da Embrapa, enquanto para o IITA foi de 0,10. No terceiro capítulo, também foi proposto um estudo de caso, porém com mapeamento associativo em melhoramento de soja, para os caracteres conteúdo de proteína e óleo. Investigou-se os benefícios do aumento da cobertura e densidade de marcadores em análises de GWAS. O objetivo deste estudo foi explorar essa questão comparando-se os resultados de GWAS para conteúdo de proteína e óleo em soja realizado com um painel de 17K SNPs derivados de GBS (*genotyping by sequencing*) e o de um painel derivado de 2.18M SNPs obtidos por meio de uma abordagem conjunta de genotipagem por GBS e WGS (*whole genome sequencing*), para o mesmo conjunto de linhagens. Nossos resultados revelaram que as análises de GWAS com um número limitado de marcadores foram bem sucedidas na identificação de regiões relevantes associadas a conteúdo de proteína e óleo em soja; os resultados deste capítulo também proporcionaram novos "insights" sobre a aplicação de imputação com dados de WGS em melhoramento de soja. Esses trabalhos procuraram abordar questões relacionadas a escolha de população de treinamento, tanto por meio de estádios de um programa de melhoramento, quanto entre programas de melhoramento (de diferentes países) com o objetivo de obter um maior entendimento sobre a possibilidade do encurtamento do longo ciclo de melhoramento para mandioca, bem como uma forma de utilizar informações genômicas para orientar a troca de germoplasma quando não há informações fenotípicas para clones de determinada instituição/país. Nas análises de associação para soja, objetivou-se verificar até que ponto se justifica o trabalho e custo extra envolvido com imputações de dados de WGS, pois, para alguns pesquisadores, o grau necessário de resolução para seleção por marcadores já seria suprido por GBS, mas isso certamente dependeria dos objetivos da pesquisa, espécie e herdabilidade do caractere.

Palavras-chave: *Manihot esculenta*. Predição genômica. População de treinamento. Genotipagem por sequenciamento. Soja. GWAS. mapeamento associativo.

ABSTRACT

TORRES, Lívia Gomes, D.Sc., Universidade Federal de Viçosa, January, 2021. **Optimizing GWS and GWAS in crop breeding.** Adviser: Marcos Deon Vilela de Resende. Co-advisers: Camila Ferreira Azevedo and Eder Jorge de Oliveira.

In a scenario of advances in genotyping technologies, with the decreasing cost per sample, the information obtained at the molecular level has been increasingly used in genetic analyses to accelerate the effective selection of clones, through genomic prediction (GP). GP leads to more accurate predictions of the individual's merit for the traits of interest. Genome-wide association studies (GWAS) have also been widely used, it aims to identify causal variants to detect significant associations between genetic markers and traits of interest. In the first chapter, predictions based on pedigree information, molecular markers and cassava productive traits were done with clones from biparental crosses by separating the full dataset by the number of stages of a breeding program (clonal evaluation trial – CET, preliminary yield trial – PYT, advanced yield trial – AYT, and uniform yield trial – UYT): one stage (CET), two stages (CET and PYT), three stages (CET, PYT and AYT) and four stages (CET, PYT, AYT and UYT). The results indicated a satisfactory potential for using genomic prediction in cassava, especially for early selection, thus saving resources. The second chapter was a case study with hydrogen cyanide content (HCN) in cassava, to evaluate the feasibility of using cross-country genomic predictions, with datasets from Brazil and Nigeria, as a strategy to assist clones' selection in germplasm exchange. In addition, it also provided an assessment of the population structure for the joint dataset with the two countries as well as genetic parameters estimations based on single nucleotide polymorphisms (SNPs) and in a haplotype approach. Comparisons on GEBVs' estimation were made considering the hypothetical situation of not having the phenotypic characterization for a set of clones for a certain research institute/country and might need to use the markers' effects that were trained with data from other research institute/country's germplasm to estimate their clones' GEBV. The joint dataset provided an improved accuracy compared to the prediction accuracy of either germplasm' sources individually. Cross-country genomic predictions proved to have potential use under the present study's scenario, the correlation between GEBVs predicted with TP from Embrapa and IITA was 0.55 for Embrapa's germplasm, whereas for IITA's it was 0.10. In the third chapter, a case study was proposed as well, it was related to association mapping in soybean breeding, for protein and oil content. The benefits of increasing the marker coverage in GWAS analyses were investigated. This study

aimed to explore this issue by comparing the results of a genome-wide association study (GWAS) for seed oil and protein content performed with either 17K GBS-derived SNPs or 2.18M GBS-/WGS-derived SNPs on the same set of soybean accessions. Our results revealed that comparatively to GWAS with WGS-imputed SNPs, GWAS analyses with a very limited number of markers have been successful in identifying relevant regions associated with protein and oil content in soybean. The results of this chapter also provided new insights into the application of imputation with WGS data in soybean breeding. These chapters sought to address issues related to the choice of training population, through stages of a breeding program, as well as between breeding programs (from different countries) in order to either have a better understanding of the possibility of shortening the long cassava breeding cycle, as well as of the use of genomic information to guide germplasm's exchange when there is no phenotypic information for clones from a given research institute/country. In the genome association analyses for soybeans, the objective was to verify the extent to which the work and extra cost involved with WGS-imputation is worthy, because, for some researchers, the level of resolution required for marker-assisted selection would already be supplied by GBS, but that would certainly depend on the research objectives, species and trait heritability.

Keywords: *Manihot esculenta*. Genomic prediction. Training population. genotyping-by-sequencing. Soybean. GWAS. Association mapping.

SUMMARY

1 GENERAL INTRODUCTION

Advances in genotyping technologies and the decreasing cost per sample, followed by the development of new tools and statistical methods applied to plant genetics are making it possible to obtain vastly improved datasets for genetic analysis (CROSSA et al., 2017). This in-depth characterization of genetic variation provides exceptional data for genotype-phenotype studies, such as in genomic predictions (GP) as well as in the study of key trait's genetic architecture through genome-wide association analysis (GWAS) (TORKAMANEH et al., 2018).

Moreover, more efficient selection methods need to be developed and tested in order to fastener genetic progress to meet the increasing food demand worldwide. Among molecular markers, SNPs stand out for their wide coverage of the genome, they occur as single nucleotide differences between individuals. Implementation of GP/selection schemes was recently boosted by the advent of low-cost SNPs marker platforms such as genotyping-by-sequencing (GBS) (ELSHIRE et al., 2011). GP has been applied to many breeding crops and has shown great potential of genetic gain and early selection for some traits, with a lot of research effort concentrated in a short period of time (DE LOS CAMPOS et al., 2009; CROSSA et al., 2010; BURGUEÑO et al., 2012; CROSSA et al., 2013; CROSSA et al., 2014; JARQUÍN et al., 2014; JARQUÍN et al., 2017; ALLIER et al., 2020; DIAS et al., 2020).

The use of genetic markers for genetic selection and breeding is based on the genetic linkage between such markers and the loci that govern the quantitative trait loci (QTLs) of interest. The linkage disequilibrium (LD) between markers and QTLs is crucial for the detection of QTL, marker-assisted selection (MAS) and genomic predictions (RESENDE et al., 2014).

Meuwissen et al. (2001) proposed genomic selection (GS) in which individual candidates for selection could be identified through molecular markers that are in linkage disequilibrium with loci associated with quantitative traits of interest. Thus, GS combines molecular and phenotypic data in a training population to predict genomic breeding values (GEBV) of individuals in a tested population that has only been genotyped (CROSSA et al., 2017). GS emphasizes the simultaneous prediction (without the use of significance tests for individual markers) of the genetic effects of thousands of SNPs throughout the genome, to capture the effects of all loci and to better explain the genetic variation of a quantitative trait.

For GWAS, it is important to test the significance of the association between loci and phenotypic trait at the population level, for each SNP, by means of hypothesis tests. The study of genetic variability between individuals through GWAS contributes to a better understanding of traits' genetic control, since it allows the identification of specific allelic variants and their regions in the genome that are related to the trait of interest. Once one has the phenotypes and the SNP data for a group of accessions, GWAS enables the detection of genetic differences accounting for the phenotypic variation (COPLEY et al., 2018), thus identifying markers that are significantly associated with the trait of interest. The idea of genome-wide association studies is that, given the sufficiently dense coverage of the genome, the functional alleles will be in LD with at least one marker; according to Joiret et al. (2019), the success of GWAS in part relies on LD as a population concept.

Cassava (*Manihot esculenta* Crantz) is an important staple root crop, a food source for millions of people mostly in developing countries and used for several industrial applications (CEBALLOS et al., 2015). In general, for some traits, the improvement rate in cassava using traditional breeding strategies is slow due to the combination of several problems related to the biology of the species, such as poor flowering, a long breeding cycle, limited genetic diversity and a low multiplication rate of planting materials (KAYONDO et al., 2018). Moreover, traditional cassava genetic improvement strategies are still very demanding in terms of financial resources; and, historically, cassava is a crop that receives less investment than other commodities crops (BART; TAYLOR, 2017).

Recently, promising results regarding trait improvement by means of genomic predictions were achieved in cassava breeding (WOLFE et al., 2017, ELIAS et al., 2018, IKEOGU et al., 2019, SOMO et al., 2020, YONIS et al., 2020). Despite being a trivial practice for some plant breeding programs, and especially for animal breeding programs, some plant breeding programs are still facing the transition's challenges from conventional plant breeding to a present scenario where conventional breeding is still essential but now it can also be supported by genome-assisted breeding evidence to drive decision making, by adding relevant information to speed up the breeding pipeline.

One of the essential factors to be considered when working with genomic predictions is how the training sets are composed; they can be constituted of landraces and/or modern improved lines, clones from different breeding programs, across breeding stages and/or even from different countries (SOMO et al., 2020). In the first chapter of this study, predictions based on pedigree information and molecular markers for cassava productive traits were done with clones prevenient to biparental crosses by separating the full dataset by the number

of stages of a breeding program (clonal evaluation trial – CET, preliminary yield trial – PYT, advanced yield trial – AYT, and uniform yield trial – UYT): one stage (CET), two stages (CET and PYT), three stages (CET, PYT and AYT) and four stages (CET, PYT, AYT and UYT). The second chapter evaluated the feasibility of using cross-country genomic predictions, with datasets from Brazil and Nigeria, as a strategy to assist clones' selection in germplasm exchange. It consisted of a case study with hydrogen cyanide (HCN) in cassava. Comparisons on GEBVs' estimation were made considering the hypothetical situation of not having the phenotypic characterization for a set of clones for a certain research institute/country and might need to use the markers' effects that were trained with data from other research institute/country's germplasm to estimate their clones' GEBV.

The third chapter addressed how and to what extent enriched SNP datasets improve the results of association studies compared to commonly used datasets composed of tens of thousands of markers. The results of a GWAS for seed oil and protein content performed with either 17K GBS-derived SNPs or 2.18M GBS-/WGS-derived SNPs on the same set of short-season soybean accessions were compared.

2 REFERENCES

ALLIER, A.; TEYSSÈDRE, S.; LEHERMEIER, C.; CHARCOSSET, A.; MOREAU, L. Genomic prediction with a maize collaborative panel: identification of genetic resources to enrich elite breeding programs. **Theoretical and Applied Genetics**, v. 133, n. 1, p. 201-215, 2020.

BART, R.; TAYLOR, N. New opportunities and challenges to engineer disease resistance in cassava, a staple food of African small-holder farmers. **PLoS Pathog**, v. 13, n. 5, p. e1006287, 2017. https://doi.org/10.1371/journal.ppat.1006287.

BURGUEÑO, J.; de los CAMPOS, G.; WEIGEL, K.; CROSSA, J. Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. **Crop Science**, v. 52, 707-719, 2012.

CEBALLOS, H.; KAWUKI, R. S.; GRACEN, V. E.; YENCHO, G. C.; HERSHEY, C. H. Conventional breeding, marker-assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. **Theoretical and Applied Genetics**, v. 128, n. 9, p. 1647-1667, 2015.

COPLEY, T. R.; DUCEPPE, M.; O'DONOUGHUE, L. S Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines. **BMC Genomics**, v. 19, p. 167, 2018.

CROSSA, J.; de los CAMPOS, G.; PÉREZ, P.; GIANOLA, D.; BURGUEÑO, J.; ARAUS, J. L.; MAKUMBI, D.; SINGH, R. P.; DREISIGACKER, S.; YAN, J.; ARIEF, V.; BAZINGER, M.; BRAUN, H. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, 186(2), 713-724, 2010. https://doi.org/10.1534/ genetics.110.118521.

CROSSA, J.; BEYENE, Y.; KASSA, S.; PÉREZ, P.; HICKEY, J. M.; CHEN, C.; de los CAMPOS, G.; BURGUEÑO, J.; WINDHAUSEN, V. S.; BUCKLER, E.; JANNICK, J.; CRUZ, M. A. L.; BABU, R. Genomic prediction in maize breeding populations with genotyping-by-sequencing. **G3: Genes, Genomes, Genetics**, v. 3, n. 11, p. 1903-1926, 2013. https://doi.org/10.1534/g3.113.008227.

CROSSA, J.; PÉREZ, P.; HICKEY, J. et al. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112, 48-60, 2014. https://doi.org/10.1038/hdy. 2013.16.

CROSSA, J.; PÉREZ-RODRÍGUEZ, P.; CUEVAS, J.; MONTESINOS-LÓPEZ, O.; JARQUÍN, D.; de los CAMPOS, G.; BURGUEÑO, J.; GONZÁLEZ-CAMACHO, J. M.; PÉREZ-ELIZALDE, S.; BEYENE, Y.; DREISIGACKER, S.; SINGH, R.; ZHANG, X.; GOWDA, M.; ROORKIWAL, M.; RUTKOSKI, J.; VARSHNEY, R. K. Genomic selection in plant breeding: Methods, model, and perspectives. **Trends in Plant Science**, v. 22, n. 11, p. 961-975, 2017.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting quantitative traits with regression

models for dense molecular markers and pedigree. **Genetics**, v. 182, n. 1, p. 375-385, 2009. https://doi.org/10.1534/ genetics.109.101501.

DIAS, K. O. G.; PIEPHO, H. P.; GUIMARÃES, L. J. M.; GUIMARÃES, P. E. O.; PARENTONI, S. N.; PINTO, M. D. O.; NODA, R. W.; MAGALHÃES, J. V.; GUIMARÃES, C. T.; GARCIA, A. A. F.; PASTINA, M. M. Novel strategies for genomic prediction of untested single-cross maize hybrids using unbalanced historical data. **Theoretical and Applied Genetics,** v. 133, n. 2, p. 443-455, 2020. https://doi.org/10. 1007/s00122-019-03475-1.

ELIAS, A.; RABBI, I.; KULAKOW, P.; JANNICK, J. Improving genomic prediction in Cassava field experiments using spatial analysis. **G3: Genes, Genomes, Genetics**, v. 8, n. 1, p. 53-62, 2018. https://doi.org/10.1534/g3.117.300323.

ELSHIRE, R.; GLAUBITZ, J.; SUN, Q.; POLAND, J.; KAWAMOTO, K.; BUCKLER, E.; MITCHELL, S. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS ONE**, v. 6, p. 1-10, 2011.

IKEOGU, U.; AKDEMIR, D.; WOLFE, M.; OKEKE, U.; CHINEDOZI, A.; JANNICK J.; EGESI, C. N. Genetic correlation, genome-wide association and genomic prediction of portable NIRS predicted carotenoids in Cassava roots. **Front. Plant Sci**., v. 10, p. 1570, 2019. https://doi: 10.3389/fpls.2019.01570.

JARQUÍN, D.; KOCAK, K.; POSADAS, L.; HYMA, K.; JEDLICKA, J.; GRAEF, G.; LORENZ, A. Genotyping by sequencing for genomic prediction in a soybean breeding population. **BMC Genomics**, v. 15, n. 1, p. 740, 2014.

JARQUÍN, D.; LEMES DA SILVA, C.; GAYNOR, R. C.; POLAND, J.; FRITZ, A.; HOWARD, R.; BATTENFIELD, S.; CROSSA, J. Increasing genomic-enabled prediction accuracy by modeling genotype x environment interactions in Kansas wheat. **The Plant Genome**, v. 10, n. 2, p. 1-15, 2017.

JOIRET, M.; MAHACHIE JOHN, J. M.; GUSAREVA, E. S.; van STEEN, K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. **BioData Mining**, v. 12, n. 11, 2019. https://doi.org/10.1186/s13040-019-0199-7.

KAYONDO, S. I.; CARPIO, D. P. D.; LOZANO, R.; OZIMATI, A.; WOLFE, M.; BAGUMA, Y.; GRACEN, V.; SAMUEL, O.; FERGUSON, M.; KAWUKI, R.; JANNINK, J. Genome-wide association mapping and genomic prediction unravels CBSD resistance in a 2 Manihot esculenta breeding population. **Scientific Reports,** v. 8, p. 1549, 2018. doi:10.1038/s41598-018-19696-1.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, p. 1819, 2001.

RESENDE, M. D. V.; SILVA, F. F.; AZEVEDO, C. F. Estatística matemática, biométrica e computacional: Modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência. Visconde do Rio Branco, MG: Suprema Editora, 2014.

SOMO, M.; KULEMBEKA, H.; MTUNDA, K.; MREMA, E.; SALUM, K.; WOLFE, M. Genomic prediction and quantitative trait locus discovery in a cassava training population constructed from multiple breeding stages. **Crop Science**, v. 60, p. 896-913, 2020. doi:10.1002/csc2.20003

TORKAMANEH, D.; BOYLE, B.; BELZILE, F. Efficient genome-wide genotyping strategies and data integration in crop plants. **Theoretical and Applied Genetics**, v. 131, p. 499-511, 2018.

WOLFE, M. D.; DEL CARPIO, D. P; ALABI, O.; EZENWAKA, L. C.; IKEOGU, U. N.; KAYONDO, I. S.; LOZANO, R.; OKEKE, U. G.; OZIMATI, A. A.; WILLIAMS, E.; EGESI, C.; KAWUKI, R. S.; KULAHOW, P.; RABBI, I. Y.; JANNIC, J.-L. Prospects for genomic selection in cassava breeding. Plant Genome, v. 10, 2017. https://doi.org/10.3835/plantgenome2017.03.0015

YONIS, B.; DEL CARPIO, D.; WOLFE, M.; JANNICK, J.; KULAKOW, P.; RABBI, I. Improving root characterization for genomic prediction in cassava. **Scientific Reports**, v. 10, 2020.

3. CHAPTER I: GENOMIC SELECTION FOR PRODUCTIVE TRAITS IN BIPARENTAL CASSAVA BREEDING POPULATIONS

This chapter was published in *Plos One* (2019 – DOI: 10.1371/journal.pone.0220245).

Abstract - Cassava improvement using traditional breeding strategies is slow due to the species' long breeding cycle. However, the use of genomic selection can lead to a shorter breeding cycle. This study aimed to estimate genetic parameters for productive traits based on pedigree (pedigree and phenotypic information) and genomic (markers and phenotypic information) analyses using biparental crosses at different stages of selection. A total of 290 clones were genotyped and phenotyped for fresh root yield (FRY), dry matter content (DMC), dry yield (DY), fresh shoot yield (FSY) and harvest index (HI). The clones were evaluated in clonal evaluation trials (CET), preliminary yield trials (PYT), advanced yield trials (AYT) and uniform yield trials (UYT), from 2013 to 2018 in ten locations. The breeding stages were analyzed as follows: one stage (CET), two stages (CET and PYT), three stages (CET, PYT and AYT) and four stages (CET, PYT, AYT and UYT). The genomic predictions were analyzed via *k*-fold cross-validation based on the genomic best linear unbiased prediction (GBLUP) considering a model with genetic additive effects and genotype × location interactions. Genomic and pedigree accuracies were moderate to high (0.56–0.72 and 0.62–0.78, respectively) for important starch-related traits such as DY and FRY; when considering one breeding stage (CET) with the aim of early selection, the genomic accuracies ranged from 0.60 (DMC) to 0.71 (HI). Moreover, the correlations between the genomic estimation breeding values of one-stage genomic analysis and the estimated breeding values of the four-stage (full data set) pedigree analysis were high for all traits as well as for a selection index including all traits. The results indicate great possibilities for genomic selection in cassava, especially for selection early in the breeding cycle (saving time and effort).

Keywords: *Manihot esculenta*; genomic prediction; selection stages; breeding cycle; dry yield.

3.1. INTRODUCTION

Cassava (*Manihot esculenta* Crantz) is an important starchy root crop grown predominantly in the tropics that serves as a source of raw material for industrial applications and a food source for millions of people, mainly in tropical and developing regions [1,2]. Moreover, it is the world's second most important source of starch [3], as its underground roots contain more than 80 percent starch by total dry weight [4]. Cassava is a rustic crop; it is believed to have an acclimation mechanism since it has no great environmental requirements and can grow in a wide range of climates and regions [5]. However, the average root yield is far below its potential due to its inadequate management and the use of unimproved varieties.

The use of improved varieties is one option to increase the production of the crop without increasing the planting area. Cassava breeding is based on recurrent phenotypic selection, taking advantage of vegetative propagation [6], and includes several stages of selection [7].

In general, for some traits, the improvement rate in cassava using traditional breeding strategies is slow due to the combination of several problems related to the biology of the species, such as poor flowering, a long breeding cycle, limited genetic diversity and a low multiplication rate of planting materials [8]. By virtue of the low multiplication rate of cassava, it takes several years to obtain enough planting material to conduct replicated multi-location evaluations [9]. Usually, the selection cycle requires one to two years to produce botanical seeds of the clones to be tested and four to six years of field evaluations, commonly divided into selection stages such as clonal evaluation trials, preliminary yield trials, advanced yield trials and uniform yield trials [6,10]. Selection decisions are made during this process to reduce the number of genotypes to be evaluated in replicated multi-location trials. However, traditional breeding strategies are still very demanding in terms of financial resources because cassava cultivation is highly labor-intensive and field costs can account for up to ninety percent of total costs [11].

When made possible by advancements in genotyping methods together with reduced costs per sample, information obtained at the molecular level has been used in genetic analyses to accelerate the effective selection of clones. Among molecular markers, single nucleotide polymorphisms (SNPs) stand out for broad genome coverage. SNPs occur as single nucleotide differences between individuals, and thousands of SNP markers are widely used in genome analysis [12]. According to [13], obtaining a catalog of SNPs segregating

within farmers' varieties can be used to accelerate breeding programs to improve cassava's quality, addressing nutrition and plant disease concerns.

[14] have proposed genomic selection (GS) to identify individual candidates for selection through molecular markers that are in linkage disequilibrium with loci associated with quantitative traits of interest. Thus, GS combines molecular and phenotypic data in a training population to predict the genomic estimated breeding values (GEBV) of individuals in a testing population that has only been genotyped [15]. The main advantage of genomic selection for cassava breeding programs is the ability to predict the agronomic potential of clones at an early stage and with high accuracy, thus reducing the generation interval [16]. This can accelerate the recurrent selection method [2]. Although the results are promising for some traits, we are currently in the early stages of genomic selection for cassava [10].

This study aims to (i) estimate genetic parameters (variance components, heritability and accuracy) from genomic (phenotypes and marker information) and pedigree (phenotypes and pedigree information) predictions for agronomic traits related to starch production, at different stages of a cassava breeding program; (ii) integrate clones' breeding values into a selection index; (iii) obtain the clones' genomic estimated breeding values through a genetic additive effect model with genotype × location interactions via Genomic Best Linear Unbiased Prediction (GBLUP); and (iv) compare the one-stage genomic analysis and the four-stage pedigree analysis to verify the feasibility of performing selection at the early stages of the breeding program using GS predictions.

3.2. MATERIALS AND METHODS

3.2.1. Plant material

A total of 290 cassava clones from the Cassava Breeding Program (CBP) of Embrapa Mandioca e Fruticultura (Cruz das Almas, Brazil) were genotyped and phenotyped for some important agronomic traits. The set of clones consisted of populations of full- and half-siblings obtained from 30 biparental crosses; commercial clones were used as parents and control trials.

3.2.2. Experimental design and phenotypical evaluation

Cassava clones were evaluated at different selection stages: clonal evaluation trials (CET), preliminary yield trials (PYT), advanced yield trials (AYT) and uniform yield trials (UYT), in field trials carried out from 2013 to 2018 in ten locations within Bahia State that included Cruz das Almas, Santo Amaro and Laje (Table 1).

Table 1. Experimental areas, cities and geographical coordinates.

Experimental Area	City	Geographical Coordinates
CNPMF	Cruz das Almas	12°40'42.4"S 39°05'27.8"W
Capela	Laje	13°10'37.8"S 39°14'22.4"W
Novo Horizonte	Laje	13°06'38.4"S 39°16'20.4"W
Novo Rumo	Laje	13°15'33.5"S 39°14'12.8"W
Rio de Areia1	Laje	13°08'51.1"S 39°17'59.7"W
Rio de Areia2	Laje	13°07'37.5"S 39°19'02.1"W
Santo Amaro	Santo Amaro	12°32'58.0"S 38°48'16.1"W
São Jorge	Laje	13°07'01.3"S 39°18'08.6"W
Sombra Verde	Laje	13°13'08.4"S 39°19'27.8"W
UFRB	Cruz das Almas	12°39'14.3"S 39°04'49.0"W

The CET plots each consisted of a single row with 5–8 plants per plot, in an augmented block design of 6–19 blocks, with 3–11 checks as a control (with the exception of one field trial that was a complete randomized block design with 2 reps). PYT plots contained two rows with 20 plants per plot, in a complete randomized block design with 2 reps and 4–8 checks as a control. AYT plots contained four rows with 36 plants per plot, in a complete randomized block design with 3 reps and 3–5 checks as a control. Finally, UYT plots included four rows with 80 plants per plot, in a complete randomized block design with 3 reps and 3–6 checks as a control.

Planting was performed from May to July (during the rainy season). Spacings between rows and plants were 0.90 and 0.80 meters, respectively. All trait management was performed, whenever necessary, in accordance with the technical recommendations and standard agricultural practices for cassava [17]. Plants were harvested 11–12 months after planting. All plants in the plots were evaluated.

The traits evaluated were as follows: fresh root yield (FRY, in t ha^{-1}), dry matter content by specific gravity method (DMC, %), measured using the gravimetric method described by [18] and expressed in %; dry yield (DY, in t ha^{-1}), determined by multiplying FRY by DMC; fresh shoot yield (FSY, in t ha^{-1}) and harvest index (HI, %). HI is the proportion of fresh root weight in total biomass.

3.2.3. Genotyping and data quality control

The genotyping of the 290 clones was performed at Cornell University. The DNA was extracted from the leaves of cassava accessions using cetyltrimethylammonium bromide according to the protocol of [19]. Then, the DNA concentration was adjusted to 60 ng/µl, and the samples were sent to the Genomic Diversity laboratory for library preparation, sequencing and bioinformatics analyses. The genotyping by sequencing (GBS) protocol is described in detail by [20]. The DNA was digested with an ApeKI enzyme, a type II restriction endonuclease that recognizes a degenerate five-base sequence and creates a 5' overhang (3 bp). The linkage between ApeKI-cut genomic DNA and adapter was made after the digestion of the samples and the 192-multiplex samples for sequencing. The *Genome Analyser 2000* (Illumina, Inc., San Diego, CA, USA) was used to perform the GBS.

Data quality control was performed for markers; indels and markers with minor allele frequency (MAF) less than 0.0415—calculated by the formula $\frac{1}{\sqrt{2N}}$, where N is the number of clones [21]—and a call rate of less than 0.90 were removed from the analyses. The GBS coverage was approx. 72%. Excluding the indels, the dataset consisted of 209,316 SNP markers. After quality control, the final dataset used to perform the genomic analyses consisted of 51,259 SNPs, with mean heterozygosity of 0.10 and mean missingness of 3.34%. The missing data were imputed by the heterozygote [22].

3.2.4. Pedigree and genomic statistical analyses

Pedigree and genomic analyses were performed. The pedigree analyses were based on the pedigree information, whereas for the genomic analyses, a genomic kinship matrix was constructed based on the information provided by the 51,259 SNP markers. The clones were evaluated at the four selection stages (CET, PYT, AYT and UYT) for FRY, DMC, DY, FSY and HI. Statistical analyses were performed by splitting the dataset by the stages of the breeding program and analyzing subsets of data with different numbers of stages. The breeding stages were analyzed as follows: one stage (CET), two stages (CET and PYT), three stages (CET, PYT and AYT) and four stages (CET, PYT, AYT and UYT) (Table 2). This division was made to assess the differences in the genetic parameters that appeared with the addition of new trials and therefore new combinations of year, location and block within location, taking into consideration the unbalance common in this type of system. The overall

aim was to verify the possibility of performing selection at an earlier stage of a breeding program (CET).

Table 2. (A) Selection stages, locations, years, number of observations and number of clones; and (B) Datasets established by the number of stages included in the analysis, locations, years, number of observations and number of clones.

(A)

Stages	Locations	Years	Number of observations	Number of clones
Clonal evaluation trials (CET)	CNPMF, Novo Horizonte, Novo Rumo and RiodeAreia1	2013, 2014, 2015, 2016 and 2018	964	290
Preliminary yield trials (PYT)	Capela and Novo Horizonte	2014 and 2017	288	136
Advanced yield trials (AYT)	Novo Horizonte, RiodeAreia1, Santo Amaro and UFRB	2015, 2016 and 2018	388	72
Uniform yield trials (UYT)	Novo Horizonte, RiodeAreia1, RiodeAreia2, Santo Amaro, São Jorge, Sombra Verde and UFRB	2016, 2017 and 2018	315	21

(B)

1 stage: CET	CNPMF, Novo Horizonte, Novo Rumo and RiodeAreia1	2013, 2014, 2015, 2016 and 2018	964	290
2 stages: CET and PYT	Capela, CNPMF, Novo Horizonte, Novo Rumo and RiodeAreia1	2013, 2014, 2015, 2016, 2017 and 2018	1252	290
3 stages: CET, PYT and AYT	Capela, CNPMF, Novo Horizonte, Novo Rumo, RiodeAreia1, Santo Amaro and UFRB	2013, 2014, 2015, 2016, 2017 and 2018	1640	290
4 stages: CET, PYT, AYT and UYT	Capela, CNPMF, Novo Horizonte, Novo Rumo, RiodeAreia1, RiodeAreia2, Santo Amaro, São Jorge, Sombra Verde and UFRB	2013, 2014, 2015, 2016, 2017 and 2018	1955	290

The models used for the pedigree (P) and genomic analyses through markers (M) are as follows, respectively:

$$\mathbf{y} = \mathbf{X}\boldsymbol{b} + \mathbf{Z_1}\boldsymbol{a_P} + \mathbf{Z_2}\boldsymbol{a_{LP}} + \boldsymbol{\varepsilon} \qquad (1)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{b} + \mathbf{Z_1}\boldsymbol{a_M} + \mathbf{Z_2}\boldsymbol{a_{LM}} + \boldsymbol{\varepsilon} \qquad (2)$$

where $\mathbf{y}$ is the vector of phenotypic observations and $\boldsymbol{b}$ is the vector of fixed effects considering the combination of year, location and block within location added to the mean. The vectors $\boldsymbol{a_P}$ and $\boldsymbol{a_M}$ refer to the random additive genotype effects, $\boldsymbol{a_P} \sim N(\mathbf{0}, \sigma^2_{aP}\mathbf{A})$ in the pedigree analysis and $\boldsymbol{a_M} \sim N(\mathbf{0}, \sigma^2_{aM}\mathbf{G})$ in the genomic analysis, where $\mathbf{A}$ is the pedigree

relationship matrix, **G** is the additive genomic relationship matrix, σ^2_{aP} is the genetic variance estimate associated with $\boldsymbol{a_P}$ and σ^2_{aM} is the genetic variance associated with $\boldsymbol{a_M}$. The random effect vectors $\boldsymbol{a_{LP}}$ and $\boldsymbol{a_{LM}}$ refer to the genotype × location interaction effect, $\boldsymbol{a_{LP}} \sim N(\mathbf{0}, \sigma^2_{alP}\mathbf{A_L})$ in the pedigree analysis, and $\boldsymbol{a_{LM}} \sim N(\mathbf{0}, \sigma^2_{alM}\mathbf{G_L})$ in the genomic analysis, where $\mathbf{A_L}$ is the block-diagonal of each location pedigree relationship matrix, $\mathbf{G_L}$ is the block-diagonal of each location genomic relationship matrix, σ^2_{alP} is the variance estimate due to genotype × location interactions with phenotype and σ^2_{alM} is the variance estimate due to genotype × location interactions with marker information. **X**, $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are the incidence matrices for fixed effects and the random effects of genotypes and genotype × location interactions, respectively. **ε** represents the residual vector ($\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2_e\mathbf{I})$), where σ^2_e is the residual variance estimate. We estimated heritability as the ratio of the genetic variance to the sum of the genetic variance, the variance due to the genotype × location interaction and the variance of the residuals (total variance) in each model; the interaction's coefficient of determination was estimated by the ratio of the variance due to the genotype × location interaction and the total variance.

The mixed model equations for best linear unbiased prediction (BLUP) (1) are as follows:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z_1} & \mathbf{X'Z_2} \\ \mathbf{Z_1'X} & \mathbf{Z_1'Z_1} + \mathbf{A^{-1}}\frac{\sigma^2_e}{\sigma^2_{aP}} & \mathbf{Z_1'Z_2} \\ \mathbf{Z_2'X} & \mathbf{Z_2'Z_1} & \mathbf{Z_2'Z_2} + \mathbf{A_L^{-1}}\frac{\sigma^2_e}{\sigma^2_{alP}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{a}}_{\mathbf{P}} \\ \hat{\boldsymbol{a}}_{\mathbf{LP}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \end{bmatrix}$$

where **A** is the pedigree relationship matrix and $\mathbf{A_L}$ is the block-diagonal of $\mathbf{An_i}$ matrices of pedigree relationship matrix in each location:

$$\boldsymbol{A_L} = \begin{bmatrix} \boldsymbol{A_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{An_i} \end{bmatrix}$$

where n_i = [4, 5, 7, 10] is the number of locations and i = [1, 2, 3, 4] is the number of stages of evaluation.

The size of **A** is 290 × 290, regardless of the number of stages included in the analysis. The size of the $\mathbf{A_L}$ matrix is 457 × 457, 581 × 581, 698 × 698 and 757 × 757 when one, two, three and four stages are included in the analysis, respectively.

The equations of mixed genomic best linear unbiased prediction (GBLUP) models (2) [23] are given as follows:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z_1} & \mathbf{X'Z_2} \\ \mathbf{Z_1'X} & \mathbf{Z_1'Z_1} + \mathbf{G^{-1}}\frac{\sigma_e^2}{\sigma_{aM}^2} & \mathbf{Z_1'Z_2} \\ \mathbf{Z_2'X} & \mathbf{Z_2'Z_1} & \mathbf{Z_2'Z_2} + \mathbf{G_L^{-1}}\frac{\sigma_e^2}{\sigma_{alM}^2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{a}}_{\boldsymbol{M}} \\ \hat{\boldsymbol{a}}_{\mathbf{LM}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \end{bmatrix}$$

where $\mathbf{G} = \frac{\mathbf{MM'}}{\sum_{i=1}^{n} 2p_i q_i}$, and **G** is the genomic relationship matrix, p_i and q_i are the allele frequencies of marker i, and **M** is the markers incidence matrix centered by the mean $2p_i$. $\boldsymbol{G_L}$ is a block-diagonal matrix of $\mathbf{Gn_i}$ matrices in each location:

$$\boldsymbol{G_L} = \begin{bmatrix} \boldsymbol{G_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{Gn_i} \end{bmatrix}$$

where n_i = [4, 5, 7, 10] is the number of locations and i = [1, 2, 3, 4] is the number of stages of evaluation.

The size of **G** is 290 × 290 regardless of the number of stages included in the analysis. The size of the $\mathbf{G_L}$ matrix is 457 × 457, 581 × 581, 698 × 698 and 757 × 757 when one, two, three and four stages are included in the analysis, respectively.

The SNP effects ($\widehat{\boldsymbol{m}}$) are calculated as follows:

$$\widehat{\boldsymbol{m}} = (\boldsymbol{MM'})^{-1}\boldsymbol{M'}\widehat{\boldsymbol{a}}_{\boldsymbol{M}}$$

The pedigree and genomic analyses were performed using the methodology of mixed models via REML/BLUP, which combines the genetic values of BLUP (best non-biased linear prediction; in genomic analysis it is called GBLUP because it depends on the G matrix) prediction procedure with the REML (residual or restricted maximum likelihood) estimation of variance components. A Likelihood Ratio Test (LRT) and deviance analysis

were undertaken to test the random effects, which were tested using the chi-square statistic [24].

The cross-validation method used was the ten-fold, and the clones were randomly assigned to each fold. The training set, composed of 9 of the 10 subsets, was used to estimate marker effects and the remaining subset was the validation set. These marker effects estimates were used to predict the genomic breeding values of the validation set individuals. This process was repeated until all 10 subsets had served as the validation population once. Trait estimates of predictive ability, accuracy and bias were calculated from cross-validation with the training set for pedigree and genomic analyses. The selection efficiency measures were estimated at each fold, and the value presented in this study is the mean of the folds.

The predictive ability was given by the correlation coefficient between predicted genetic values (PGVs) and predicted genetic values added to the mean of the residuals in the validation population. PGVs were the estimated breeding values (EBVs) for the pedigree analysis and the genomic estimated breeding values (GEBVs) for the genomic analysis. Predicted genetic values added to the mean of the residuals for each clone can be interpreted as pseudo-phenotypes (PPs). We adopted this type of cross-validation due to the considerable imbalance of the dataset. For example, the control clones (some of which were parents of other clones to be tested) were replicated at several locations and through the years, whereas many other clones to be tested were replicated only a few times (due to the limited planting material).

The accuracy, one of the main measures to compare models and methods in genomic selection, was calculated as the ratio between the predictive ability and the square root of the phenotypic trait heritability. This measure indicates how accurate the model is in estimating genetic values.

The PPs were linearly regressed on the PGVs, and the regression coefficient $\hat{b}$PP,PGV was used to measure the degree of bias of the PGV prediction. The bias relates to the size of the absolute differences between clones' predicted genetic values and their pseudo-phenotypes. The estimated magnitude of these differences can be quantified by the $\hat{b}$PP,PGV regression coefficient and can be overestimated ($\hat{b}$PP,PGV < 1) or underestimated ($\hat{b}$PP,PGV > 1). A regression coefficient equal to one indicates no bias. Then, here we will represent bias as one unit minus the regression coefficient $\hat{b}$PP,PGV (Bias = 1-$\hat{b}$PP,PGV). Statistical analyses were performed using the package sommer [25] in the software R [26].

3.2.5. Selection Index

After obtaining the EBVs and GEBVs of the clones, we calculated an empirical selection index (SI) that integrates five relevant traits, assigning them weights based on the breeding program's objective:

SI = (FRY*10) + (DMC*10) + (DY*10) + (FSY*5) + (HI*3)

3.3. RESULTS

3.3.1. Pedigree analyses

The set of clones presented genetic variability for all evaluated traits ($p \leq 0.01$). In relation to the genotype × location interaction effect (GxL), when only one stage was evaluated, the clones presented different behaviors in relation to environmental changes for FRY, DY and FSY ($p \leq 0.01$); however, when four stages were included in the analysis, the additive GxL interaction effect was significant ($p \leq 0.01$) for all of the traits (Table 3).

Table 3. Pedigree analyses: comparison of models by Likelihood Ratio Test (LRT).

One stage					
	LRT [/3]				
Model	**FRY**	**DMC**	**DY**	**FSY**	**HI**
Complete model x Restricted model (without G[/1])	81.71**	144.46**	78.61**	61.00**	77.45**
Complete model x Restricted model (without GxL[/2])	14.46**	0.00	13.83**	51.03**	-0.25
Four stages					
	LRT				
Model	**FRY**	**DMC**	**DY**	**FSY**	**HI**
Complete model x Restricted model (without G)	160.18**	306.91**	148.09**	155.56**	180.85**
Complete model x Restricted model (without GxL)	38.20**	11.08**	38.33**	132.03**	39.68**

For each trait, three models were compared based on the Likelihood Ratio Test (LRT); [/1] G: genotype effects; [/2] GxL: genotype by location interaction effects; [/3] LRT = 2*(complete model log-likelihood - restricted model log-likelihood), ** significant at 0.01 and * significant at 0.05 by the likelihood ratio test.

Trait heritabilities ranged from 0.42 to 0.73, i.e., from moderate to high. DMC and HI were the traits with the highest heritabilities (ranging from 0.69 to 0.73 and from 0.56 to 0.60, respectively), while heritability of the other traits varied across the analyzed data sets (Table 4). The variation of the genetic parameters with the increment of selection stages occurs due to the addition of phenotypic observations, and therefore information that comes

from different experimental designs. Successive stages of selection gradually reduce the number of genotypes as plot-size and locations increase [6]. Trait heritabilities were higher when the full data set (four-stage) was analyzed, with values of 0.52, 0.50, 0.52 and 0.60 for FRY, DY, FSY and HI, respectively, with the exception of DMC, which exhibited higher heritability when one stage was analyzed (0.73). DMC and FSY were the traits with the lowest and highest coefficients of determination of the interaction, respectively. The pedigree means varied from 29.24–31.81 t ha^{-1}, 33.41–33.54%, 9.63–10.61 t ha^{-1}, 12.25–13.65 t ha^{-1} and 70.62–73.86% for FRY, DMC, DY, FSY and HI, respectively, depending on the number of stages included in the analysis.

Table 4. Estimates of heritability, mean, coefficient of variation (CVe), predictive ability, accuracy and bias for pedigree analyses.

Traits[/1]	Heritability	Interaction's Coefficient of Determination	Mean	CVe	Predictive Ability	Accuracy	Bias
One stage							
FRY	0.49	0.07	29.24	0.36	0.44±0.18	0.63±0.25	-0.03±0.51
DMC	0.73	0.00	33.49	0.05	0.54±0.09	0.64±0.11	0.01±0.27
DY	0.47	0.07	9.63	0.39	0.42±0.19	0.62±0.27	-0.03±0.51
FSY	0.48	0.12	12.25	0.82	0.45±0.15	0.64±0.22	0.01±0.40
HI	0.58	0.00	70.62	0.13	0.52±0.22	0.69±0.29	0.08±0.34
Two stages							
FRY	0.44	0.08	31.34	0.31	0.51±0.12	0.76±0.18	-0.01±0.21
DMC	0.70	0.00	33.48	0.05	0.61±0.08	0.73±0.10	-0.03±0.17
DY	0.42	0.08	10.53	0.33	0.51±0.13	0.78±0.20	-0.03±0.22
FSY	0.42	0.17	13.65	0.67	0.45±0.17	0.70±0.26	0.05±0.34
HI	0.56	0.06	72.84	0.11	0.50±0.16	0.67±0.21	0.09±0.31
Three stages							
FRY	0.46	0.08	31.81	0.29	0.49±0.15	0.72±0.23	0.07±0.27
DMC	0.69	0.01	33.41	0.04	0.63±0.08	0.75±0.10	-0.06±0.11
DY	0.45	0.09	10.61	0.31	0.50±0.15	0.74±0.22	0.05±0.26
FSY	0.42	0.19	12.51	0.68	0.46±0.16	0.70±0.24	0.07±0.40
HI	0.58	0.06	73.86	0.11	0.53±0.13	0.69±0.17	0.08±0.27
Four stages							
FRY	0.52	0.08	31.36	0.27	0.53±0.13	0.73±0.18	0.02±0.23
DMC	0.70	0.04	33.54	0.04	0.61±0.10	0.73±0.12	-0.02±0.12
DY	0.50	0.09	10.48	0.29	0.53±0.13	0.75±0.19	0.00±0.23
FSY	0.52	0.17	12.66	0.62	0.55±0.13	0.77±0.19	-0.02±0.33
HI	0.60	0.07	73.86	0.10	0.53±0.12	0.69±0.16	0.08±0.22

[/1] FRY – fresh root yield (t ha^{-1}); DMC – dry matter content (%); DY – dry yield (t ha^{-1}); FSY – fresh shoot yield (t ha^{-1}); and HI – harvest index (%).

The coefficient of residual variation (CVe), which is a measure of dispersion and relative variability, ranged from 0.04 to 0.82 across traits and datasets (number of stages included), being higher for FSY (0.62–0.82) and lower for DMC (0.04–0.05). In general, for all traits, the CVe decreased when a greater number of stages was included in the analysis.

Trait predictive abilities ranged from 0.42–0.54, 0.45–0.61, 0.46–0.63 and 0.53–0.61 when one, two, three and four stages were included in the analyses, respectively. For FRY, the predictive ability was 0.44, 0.51, 0.49 and 0.53, when one, two, three and four stages were included in the analyses. For DY, the predictive ability also increased continuously (except for a small reduction in the three-stage analysis), with values of 0.42, 0.51, 0.50 and 0.53 when one, two, three and four stages were included in the analyses. That increasing pattern occurred for all traits with the exception of HI, for which the predictive ability did not alter much, ranging from 0.50 to 0.53.

Trait accuracies ranged from 0.62–0.69, 0.67–0.78, 0.69–0.75 and 0.69–0.77 when one, two, three and four stages were included in the analyses, respectively (Table 4). FRY, DMC, DY and FSY presented accuracies greater than 0.70 when more than one stage was included in the analyses (two-, three- and four-stage datasets). Accuracies above 0.70 are very suitable from the selection practice point of view. According to [27], accuracy values between 0.70 and 0.90 are classified as high precision and values above 0.90 as very high precision. The increase in accuracy and predictive ability may reflect a greater experimental reliability of the stages included in the analyses. When we think of the one-stage analysis, with CET only, this field trial was performed with single rows with 6–8 plants per plot, differing from the subsequent stages, which included replications and were performed with 2–4 rows and a greater number of plants per plot. FRY and DY presented higher accuracies when four stages were included in the analysis compared to when only one was evaluated. For FRY, these values were 0.63 and 0.73, and for DY, they were 0.62 and 0.75. For the other traits, those differences in accuracy were of approximately the same amplitude, except for HI, for which the accuracy was 0.69 whether one or four stages were included in the analyses.

The bias ranged from -0.06 (DMC, three stages) to 0.09 (HI, two stages), with the lowest amplitude (highest bias value minus lowest bias value) occurring when four stages were evaluated. In general, the traits presented little bias, ranging from -0.03 to 0.07 for FRY, from -0.06 to 0.01 for DMC, from -0.03 to 0.05 for DY, from -0.02 to 0.07 for FSY and from 0.08 to 0.09 for HI. DY presented zero bias when four stages were included in the analysis. HI was the trait that presented the highest values of bias; therefore, the estimates of the genetic values for HI were slightly overestimated.

Indirect selection is an option when there is a linear relationship between two traits and one trait either has low heritability or is difficult to measure compared to the other trait. For example, FSY is relatively easier to measure than root-related traits, and it had moderate positive correlations with FRY (0.41–0.48) and DY (0.43–0.49) and negative and moderate

to high correlations with HI ((-0.59)–(-0.62)). FRY was highly and positively correlated with DY (0.99) (Table 5). HI was negatively correlated with DMC (-0.31 and -0.19 in the one-stage and four-stage analyses, respectively). The correlations between vectors of estimated breeding values obtained through the different analyses (with a different number of stages of selection—one and four stages) are presented in the diagonal of Table 5. The high correlations, ranging from 0.84 to 0.88, indicate good correspondence between the one-stage and four-stage EBVs, which might allow early selection.

Table 5. Trait correlation estimates between vectors of EBVs from pedigree analysis, above the diagonal considering the four stages of evaluation, and below the diagonal considering only one stage of evaluation; along the diagonal is the correlation between vectors of genetic values through the different analyses of stages of selection.

Traits[/1]	FRY[/2]	DMC	DY	FSY	HI
FRY	0.88**	-0.04^{ns}	0.99**	0.48**	0.26**
DMC	-0.10^{ns}	0.84**	0.11^{ns}	0.09^{ns}	-0.19**
DY	0.99**	0.04^{ns}	0.88**	0.49**	0.24**
FSY	0.41**	0.17**	0.43**	0.87**	-0.62**
HI	0.35**	-0.31**	0.32**	-0.59**	0.87**

[/1] FRY – fresh root yield (t ha^{-1}); DMC – dry matter content (%); DY – dry yield (t ha^{-1}); FSY – fresh shoot yield (t ha^{-1}); and HI – harvest index (%); [2/]** Significant at $p \leq 0.01$, * Significant at $p \leq 0.05$, and [ns] not significant.

3.3.2. Genomic analyses

The set of clones presented genetic variability for all evaluated traits ($p \leq 0.01$), as was shown for the pedigree analyses. The GxL interaction effect was significant ($p \leq 0.01$) for all traits with the exception of DMC (Table 6).

Table 6. Genomic analyses: comparison of models by Likelihood Ratio Test (LRT).

One stage					
	LRT [/3]				
Model	**FRY**	**DMC**	**DY**	**FSY**	**HI**
Complete model x Restricted model (without G[/1])	108.60**	159.53**	106.11**	96.91**	105.42**
Complete model x Restricted model (without GxL[/2])	18.38**	0.00	19.09**	47.19**	6.65**
Four stages					
	LRT				
Model	**FRY**	**DMC**	**DY**	**FSY**	**HI**
Complete model x Restricted model (without G)	151.58**	381.14**	149.84**	194.97**	315.53**
Complete model x Restricted model (without GxL)	60.64**	1.55	61.52**	139.10**	19.85**

For each trait, three models were compared based on the Likelihood Ratio Test (LRT); [/1] G: genotype effects; [/2] GxL: genotype by location interaction effects; [/3] LRT = 2*(complete model log-likelihood - restricted model log-likelihood), ** significant at $p \leq 0.01$ and * significant at $p \leq 0.05$ by the likelihood ratio test.

The traits presented moderate to high heritability, ranging from 0.38–0.48, 0.66–0.69, 0.36–0.46, 0.41–0.51 and 0.47–0.60 for FRY, DMC, DY, FSY and HI, respectively (Table 7). In general, heritability estimates obtained in the genomic analyses were lower than those obtained by the pedigree analyses. However, the differences in heritabilities estimated through pedigree and genomic analyses were smaller when a larger number of stages of selection was included in the analysis. DMC was the trait that presented the highest heritabilities, whereas DY presented the lowest heritabilities, regardless of the dataset evaluated. As was found in the pedigree analyses, in the genomic analyses DMC and FSY were the traits with the lowest and highest coefficients of determination of the interaction, respectively.

Table 7. Estimates of heritability, mean, coefficient of variation (CVe), predictive ability, accuracy and bias for genomic analyses.

One stage							
Traits[/1]	**Heritability**	**Interaction's Coefficient of Determination**	**Mean**	**CVe**	**Predictive Ability**	**Accuracy**	**Bias**
FRY	0.39	0.06	29.67	0.36	0.43±0.13	0.61±0.19	0.04±0.40
DMC	0.66	0.00	33.43	0.05	0.51±0.09	0.60±0.10	0.05±0.27
DY	0.36	0.06	9.87	0.38	0.43±0.14	0.63±0.20	0.03±0.40
FSY	0.45	0.09	14.15	0.71	0.48±0.11	0.69±0.17	-0.02±0.30
HI	0.47	0.04	72.76	0.13	0.54±0.20	0.71±0.26	0.01±0.28
Two stages							
FRY	0.38	0.08	29.61	0.33	0.46±0.11	0.69±0.16	0.06±0.23
DMC	0.67	0.00	33.42	0.05	0.57±0.12	0.68±0.14	0.00±0.22
DY	0.37	0.08	9.85	0.35	0.47±0.11	0.72±0.17	0.01±0.21
FSY	0.41	0.13	15.80	0.58	0.44±0.21	0.68±0.32	0.11±0.40
HI	0.49	0.06	70.19	0.12	0.42±0.21	0.56±0.28	0.27±0.38
Three stages							
FRY	0.42	0.08	29.15	0.32	0.44±0.14	0.65±0.21	0.10±0.28
DMC	0.68	0.00	33.22	0.05	0.58±0.10	0.70±0.12	-0.02±0.15
DY	0.39	0.08	9.65	0.34	0.45±0.13	0.67±0.20	0.07±0.24
FSY	0.43	0.12	12.47	0.70	0.47±0.17	0.72±0.26	0.10±0.34
HI	0.56	0.03	73.74	0.11	0.47±0.19	0.62±0.25	0.22±0.33
Four stages							
FRY	0.48	0.08	31.17	0.27	0.50±0.12	0.69±0.17	0.03±0.24
DMC	0.69	0.01	33.23	0.04	0.59±0.10	0.70±0.12	-0.01±0.14
DY	0.46	0.08	10.37	0.29	0.51±0.11	0.72±0.16	-0.01±0.21
FSY	0.51	0.12	14.18	0.56	0.50±0.16	0.70±0.23	0.07±0.29
HI	0.60	0.03	73.58	0.11	0.48±0.17	0.61±0.22	0.23±0.28

[/1] FRY – fresh root yield (t ha^{-1}); DMC – dry matter content (%); DY – dry yield (t ha^{-1}); FSY – fresh shoot yield (t ha^{-1}); and HI – harvest index (%).

In general, the predictive ability was higher when four stages were included in the analysis, except for HI, for which it decreased. For FRY, DMC and DY, the predictive abilities were higher in the four-stage analyses (0.50, 0.59 and 0.51, respectively) and lower

in the one-stage analyses (0.43, 0.51 and 0.43, respectively); for FSY, the predictive ability was higher in the four-stage analysis (0.50) but lower in the two-stages analysis (0.44) compared to one stage. For HI, predictive ability was also higher for the one-stage analysis (0.54) than for the two-stage analysis (0.42).

The accuracies ranged from 0.61–0.69, 0.60–0.70, 0.63–0.72, 0.68–0.72 and 0.56–0.71 for FRY, DMC, DY, FSY and HI, respectively (Table 7). Across the analyses based on different datasets (with a different number of stages included), accuracy estimates varied (but they were at least greater than 0.56). As an example, for DY, the accuracy values were 0.63, 0.72, 0.67 and 0.72 when one, two, three and four stages were included in the analyses. For HI, the scenario was quite different, with a large reduction in accuracy from one to two stages, with values of 0.71, 0.56, 0.62 and 0.61 when one, two, three and four stages were included. The inclusion of more information and the imbalance common for this type of system (more stages – selection included – fewer clones – more locations in the latter stages) may have affected the traits differently in relation to parameters of genomic selection precision.

In general, the traits presented no or very little bias, with the exception of HI when two, three or four stages were included in the analyses. As shown in the pedigree analysis, the genetic values for HI were overestimated. The bias in the genomic analyses varied slightly more, in terms of amplitude, when compared to the biases from the pedigree analyses (they ranged from -0.02 to 0.27 for genomic analyses and from -0.06 to 0.09 for pedigree analyses).

The estimates of the correlations based on the GEBVs (Table 8) were similar to the estimates of the correlations based on the EBVs (Table 5). The most expressive differences of trait correlations presented in Tables 5 and 8 were between the trait pairs FRY and FSY, FRY and HI, DY and FSY, DY and HI, and DMC and HI. FRY and DY were highly and positively correlated (0.99 with one-stage and four-stage datasets). HI and FSY were negatively correlated, with values of -0.62 and -0.71 when one and four stages were included in the analyses, respectively. The differences in correlation with respect to the stage of the breeding program may be explained by the fact that the relationships among traits change through the breeding program due to selection [6]. The diagonal of Table 8 shows the correlation between vectors of GEBVs through the different analyses (with different numbers of stages of selection—one and four stages). These values (ranging from 0.80 to 0.86) indicate good correspondence between the vectors of GEBVs from one- and four-stage analyses, which may be evidence in favor of early selection.

Table 8. Trait correlation estimates between vectors of GEBVs from genomic analysis, above the diagonal considering the four stages of evaluation, and below the diagonal considering only one stage of evaluation; along the diagonal is the correlation between vectors of genomic values through the different analyses of stages of selection.

Traits[/1]	FRY[/2]	DMC	DY	FSY	HI
FRY	0.81**	-0.05^{ns}	0.99**	0.40**	0.37**
DMC	-0.05^{ns}	0.86**	0.11^{ns}	0.11^{ns}	-0.20**
DY	0.99**	0.07^{ns}	0.80**	0.41**	0.34**
FSY	0.35**	0.10^{ns}	0.35**	0.86**	-0.61**
HI	0.41**	-0.21**	0.40**	-0.62**	0.85**

[/1] FRY – fresh root yield (t ha^{-1}); DMC – dry matter content (%); DY – dry yield (t ha^{-1}); FSY – fresh shoot yield (t ha^{-1}); and HI – harvest index (%); [2/]** Significant at $p \leq 0.01$, * Significant at $p \leq 0.05$, and ^{ns} not significant.

The SNP effects for the five traits, considering one and four stages of evaluation, are shown in Fig 1. Comparing the one-stage with the four-stage evaluation graphs, one can see differences in terms of which chromosome the higher-effect SNPs were located on and also in the magnitude of the effects resulting from the addition of extra information from one stage to four stages of evaluation. These differences could be explained by the fact that in the analyses that included only one stage, the clones were evaluated at four locations, while in the analyses that included the four stages of selection, the clones were evaluated in ten locations and the amount of individual information doubled (Table 2).

3.3.3. Comparisons

If the results of the one-stage and four-stage evaluations are comparable, it would verify the possibility of using the results from genomic analysis to select clones at the initial stage of a cassava breeding program. To perform this comparison, we compared the rankings of the top 10 clones based on the one-stage genomic analysis (aiming at a reduction of the selection interval) and the four-stage pedigree analysis (complete data set) for each trait as well as for the selection index. We also calculated the correlations between GEBVs (one-stage analysis) and EBVs (four-stage analysis), which were high for all traits, ranging from 0.82 (FRY and DY) to 0.87 (FSY).

For FRY, all clones in the top 10 ranking based on the one-stage genomic analysis were from the cross of Fécula Branca and BRS Formosa, indicating that their progeny represent an important full-siblings family for this trait. There was a good correspondence between the GEBVs from the one-stage genomic analysis and the EBVs from the four-stage pedigree analysis, with a correlation of 0.82 (S1 Table). Many of the clones in the two top-

10 rankings coincided (eight out of ten, from the 'Fécula Branca x BRS Formosa' crossing). In the top 10 based on the four-stage pedigree analysis, a clone from the cross between BGM-1683 and Fécula Branca also stood out.

There was a good correspondence between GEBVs and EBVs for DMC (correlation of 0.83), and the one-stage and four-stage top 10 lists shared five clones. Progeny from the cross between Fécula Branca and BRS Formosa also stood out for this trait, as well as crossings between the clones Equador72 and BGM-0728, Equador72 and BGM-1124, Cascuda and BGM-0728, BGM-1662 and Fécula Branca and the BRS Mulatinha self-crossing (S2 Table).

For DY, all of the clones in the rankings originated from the cross between Fécula Branca and BRS Formosa, and there was a good correspondence between the GEBVs and the EBVs (correlation of 0.82) (S3 Table). There was a high coincidence between the clones presented in the top 10 rankings for the traits FRY and DY (nine out of ten clones), which was already expected since those traits were highly correlated (0.99).

Fresh shoot yield was the trait that presented the highest amplitude in the top 10 ranking list. The amplitude of GEBVs (one-stage genomic analysis) for the top 10 clones was 20.90; for EBVs (four-stage pedigree analysis), the amplitude was 8.62. This trait also presented a high correlation between the GEBVs and the EBVs (0.87) (S4 Table).

The same pattern was observed for HI, which presented a correlation of 0.84 between GEBVs and EBVs. Moreover, for this trait, all clones in both rankings were from the crossing of Fécula Branca and BRS Formosa (S5 Table).

Assuming that it is a good idea to select clones based on their GEBVs obtained from the one-stage genotypic analysis, the genetic gains from selecting the top 10% of clones (29 clones) were calculated for each trait. The genetic gains were as follows: 15.91% for FRY, 5.15% for DMC, 14.76% for DY, 43.11% for FSY, and 8.47% for HI.

For the selection index, the correlation between GEBVs and EBVs was 0.83 (Table 9). Comparing the top 10% of the clones based on the rankings from these analyses, it was observed that there was a coincidence of 72.41% (21 out of 29 clones). Several clones from Fécula Branca × BRS Formosa crosses presented good performance, as previously noted for individual traits. However, clones from a considerable number of other crosses were also highlighted, such as Equador 72 × Eucalipto, Equador 72 × BGM-0728, BGM-1683 × Fécula Branca, and BRS Formosa × BRS Jari. The genetic gains made by selecting the top 10% of clones based on SI (calculated from the GEBVs of the one-stage analysis) were estimated as

follows: 15.13% for FRY, 0.39% for DMC, 14.38% for DY, 24.97% for FSY, and 1.85% for HI.

Table 9. Comparison of top 10% rankings between selection index (SI) values based on genomic estimated breeding values (one evaluation stage) or on estimated breeding values (four stages).

Correlation between GEBVs (one stage genomic analysis) and EBVs (four stages pedigree analysis) = 0.83							
Genomic analysis – One stage				**Pedigree analysis – Four stages**			
Clone	GEBV	Male genitor	Female genitor	Clone	EBV	Male genitor	Female genitor
2012_108_043	1459.09	Fécula Branca	BRS Formosa	2012_108_043	1442.70	Fécula Branca	BRS Formosa
2012_108_208	1402.01	Fécula Branca	BRS Formosa	2012_108_208	1423.61	Fécula Branca	BRS Formosa
2012_108_108	1333.69	Fécula Branca	BRS Formosa	2012_108_108	1413.39	Fécula Branca	BRS Formosa
2012_108_215	1310.99	Fécula Branca	BRS Formosa	2012_108_060	1341.44	Fécula Branca	BRS Formosa
2012_108_060	1308.46	Fécula Branca	BRS Formosa	2012_108_188	1338.13	Fécula Branca	BRS Formosa
2012_108_155	1285.94	Fécula Branca	BRS Formosa	2012_108_155	1310.21	Fécula Branca	BRS Formosa
2012_108_188	1285.12	Fécula Branca	BRS Formosa	2012_108_046	1283.49	Fécula Branca	BRS Formosa
2012_108_046	1267.61	Fécula Branca	BRS Formosa	2014_013_28	1239.57	Equador72	BGM-0728
2012_108_035	1254.20	Fécula Branca	BRS Formosa	2012_108_036	1227.49	Fécula Branca	BRS Formosa
2012_108_036	1250.92	Fécula Branca	BRS Formosa	2012_108_143	1226.44	Fécula Branca	BRS Formosa
BRS Kiriris	1238.41	-	-	2014_018_39	1224.53	Equador72	Eucalipto
2014_018_06	1207.62	Equador72	Eucalipto	2014_013_07	1220.31	Equador72	BGM-0728
2012_108_185	1201.35	Fécula Branca	BRS Formosa	2012_108_185	1218.95	Fécula Branca	BRS Formosa
2012_108_143	1198.61	Fécula Branca	BRS Formosa	2012_108_035	1216.20	Fécula Branca	BRS Formosa
2014_001_34	1192.00	BGM-1332	Fécula Branca	2012_108_041	1205.53	Fécula Branca	BRS Formosa
2014_006_48	1191.36	BGM-1683	Fécula Branca	2014_006_33	1202.38	BGM-1683	Fécula Branca
2012_108_041	1182.19	Fécula Branca	BRS Formosa	2012_108_186	1197.24	Fécula Branca	BRS Formosa
2012_108_129	1178.80	Fécula Branca	BRS Formosa	2012_108_215	1196.28	Fécula Branca	BRS Formosa
2012_108_050	1171.34	Fécula Branca	BRS Formosa	2012_108_050	1187.90	Fécula Branca	BRS Formosa
2012_108_062	1170.54	Fécula Branca	BRS Formosa	2014_006_48	1184.81	BGM-1683	Fécula Branca
2011_34_41	1160.19	Irará	BRS Kiriris	2014_025_42	1184.05	BGM-1662	Fécula Branca
2012_108_038	1159.94	Fécula Branca	BRS Formosa	2014_013_30	1183.75	Equador72	BGM-0728
2012_108_071	1157.50	Fécula Branca	BRS Formosa	2012_108_062	1176.65	Fécula Branca	BRS Formosa
2012_107_002	1156.17	BRS Formosa	BRS Jari	2012_108_129	1174.30	Fécula Branca	BRS Formosa
2014_013_28	1153.43	Equador72	BGM-0728	BRS Kiriris	1166.00	-	-
2012_108_092	1148.86	Fécula Branca	BRS Formosa	2012_108_071	1163.47	Fécula Branca	BRS Formosa
2014_001_08	1148.75	BGM-1332	Fécula Branca	2012_108_092	1161.13	Fécula Branca	BRS Formosa
2011_24_156	1147.67	Lagoa	Mani-Branca	2012_108_132	1160.68	Fécula Branca	BRS Formosa
2012_108_206	1145.19	Fécula Branca	BRS Formosa	2014_012_13	1155.49	Cascuda	BGM-0728

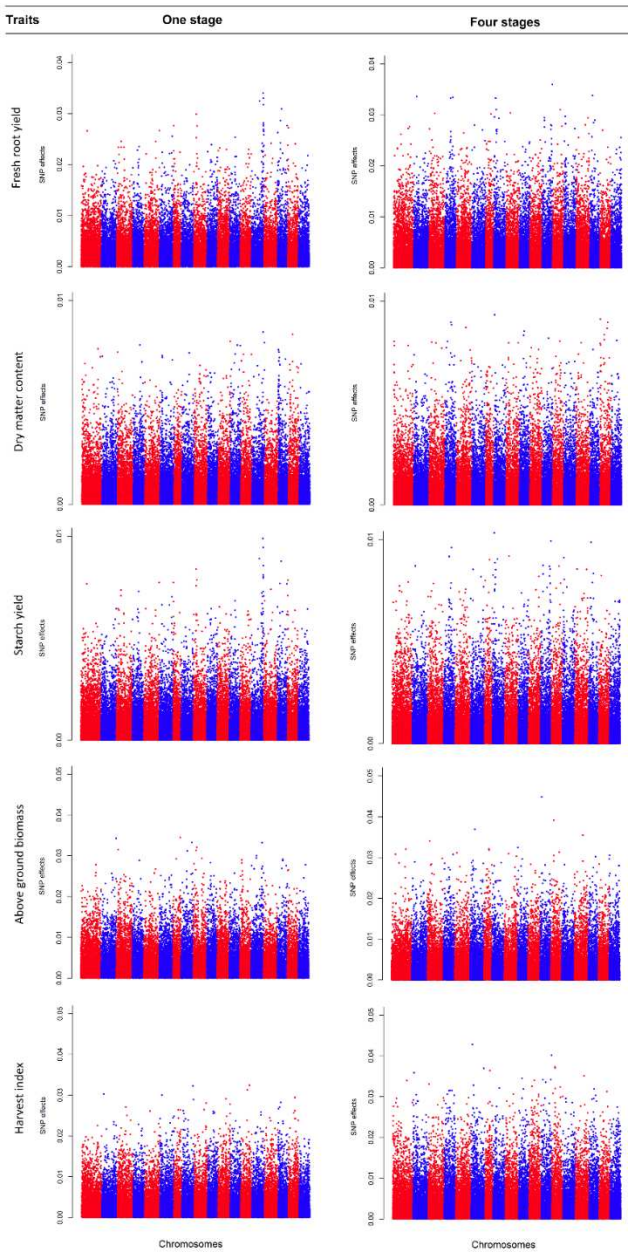


Fig 1. SNP effects for the traits FRY (fresh root yield (t ha^{-1})), DMC (dry matter content (%)), DY (dry yield (t ha^{-1})), FSY (fresh shoot yield (t ha^{-1})) and HI (harvest index (%)), considering only one and four stages of evaluation.

3.4. DISCUSSION

Cassava breeding is usually based on a recurrent phenotypic selection with several stages. For some traits, the improvement rate in cassava is slow due to a combination of several problems related to the biology of cassava. Genomic selection could be one promising approach to reduce the selection interval by predicting the agronomic potential of clones earlier in the process [16, 15, 10].

In our study, we evaluated 290 cassava clones from biparental crosses, which were genotyped and phenotyped for productive traits. Our major objective was to characterize the set of clones, determining genetic parameters relying on pedigree and marker information, as well as to compare the results of one and four selection stages to verify the feasibility of genomic selection at the program's early stage. We found substantial genetic variation among the 290 cassava clones for all traits evaluated. Moreover, they showed intermediate to high heritability, and consequently, good progress can be made in selecting clones for those traits. The majority of traits presented significant genotype × location interaction effects ($p \leq 0.01$), with the exception of DMC (one-stage pedigree analysis and one- and four-stage genomic analyses) and HI (one-stage pedigree analysis).

The coefficient of determination of the interaction represents how much of the total variation is explained by the genotype × location interaction. According to our results, DMC and HI appear less impacted by genotype × location interactions and their impacts than the other traits (FRY, DY and FSY).

FRY and DMC are important agronomic traits for cassava, and the mean trait values found in the present work were consistent with the results of [28]. They evaluated 627 genotypes in a clonal trial, and their estimates for FRY ranged from 9.64 to 63.66 t ha^{-1} (mean of 34.37 t ha^{-1}) and for DMC from 20.65 to 41.25% (mean of 33.47%). In the study by [6], they performed successive stages of phenotypic assessment of cassava adapted to the sub-humid environment, and the observed means for FRY, DMC and HI were 24.00 t ha^{-1}, 33.57% and 58%, respectively. In the present study, when only the one-stage pedigree analysis (CET) was assessed, we found means of 29.24 t ha^{-1}, 33.49% and 70.62%, for FRY, DMC and HI, respectively (Table 4).

Dry matter content is usually what determines the price paid to cassava producers since it is directly related to the starch yield and, consequently, its several derivative products [29, 30]. Despite its importance for starch yield (SY), the correlations between DMC and DY (which is a redundant trait for SY since the correlation between SY and DY is 0.99 (data not shown)) in the present work were low, ranging from 0.04 to 0.11 (Tables 5 and 8), which differs from the values reported by [30], where the genetic and phenotypic correlations for DMC and SY were 0.36 and 0.42, respectively.

Harvest index is commonly considered an indirect cassava yield indicator based on reports of high correlation between HI in single row trials and FRY performance in replicated plots in multi-location trials [31, 3]. In our work, the correlations between HI and FRY were moderate, ranging from 0.26 to 0.41 (Tables 5 and 8). HI and FSY are negatively correlated

traits. It is important to be aware that plants with high HI and low FSY, even with high-yielding roots, are undesirable because they produce little propagating material. In this case, a high HI may not reflect a high-yielding root, but a low production of fresh shoot yield. It is important to seek a balance between the production of roots and shoot [32].

The total dataset consisted of clones from biparental crosses and commercial clones (used as parents and controls); therefore, there were half-sibling and full-sibling populations (clones sharing one or two parents, respectively). According to [15], using GS to achieve a shorter interval cycle is predicted to be favorable within full-sibling families because biparental populations have high linkage disequilibrium (LD) between marker alleles and trait alleles with no group structure.

All traits presented moderate to high heritability, which shows great potential for obtaining genetic gains. Traits with higher heritability tend to show higher accuracy, but this was not the case in some situations for the present work. For example, in one-stage genomic analysis, FSY and HI presented the highest accuracies (0.69 and 0.71, respectively), but the heritability of DMC (0.66) was higher than those of FSY and HI (0.45 and 0.47, respectively; Table 7). As reported by [33], those discrepancies can presumably be caused by the proximity of SNP markers to the QTL (quantitative trait loci) and the genetic architecture. In pedigree and genotypic analyses, DMC was the trait that presented the highest heritability, which might be related to the fact that DMC was the only trait that did not present a significant ($p \leq 0.01$) genotype × location interaction effect (Table 6). According to [33], the genotype × location effect and the relatedness among genotypes affect the precision of genomic selection estimates.

Using only the most informative SNPs, [16] observed an increase in the accuracy for the traits FRY, DMC and SY. In their work, 358 cassava accession were evaluated using 390 marker SNPs, and broad heritabilities (obtained by REML analysis of phenotypic data) of 0.72, 0.67 and 0.69, respectively, were found. For FRY, DMC and SY, they obtained accuracies of 0.31, 0.20 and 0.33, respectively. When they selected markers, the accuracies were 0.76, 0.67 and 0.77 for FRY, DMC and SY, respectively, with predictive abilities of 0.38, 0.30 and 0.39. In the present work, in our four-stage genomic analysis, we observed predictive abilities for FRY, DMC and DY of 0.50, 0.59 and 0.51, respectively (Table 7), without selecting markers, taking into consideration 51,259 SNPs. The heritabilities for FRY, DMC, FSY and HI were considerably higher (ranging from 0.46 to 0.69; four-stage, Table 7) than those reported by [33], which ranged from 0.11 to 0.28. [33] adopted three different cross-validation (CV) approaches: CV without close relatives, random fivefold CV

and CV with close relatives. Accuracies varied from 0.46 to 0.48 for DMC, from 0.43 to 0.48 for HI, from 0.36 to 0.41 for FRY, and from 0.23 to 0.30 for FSY; the lowest accuracy value was obtained using CV without close relatives, and the highest was obtained by random fivefold CV. The genomic accuracies that we have presented in this study varied from 0.60–0.70, 0.56–0.71, 0.61–0.69 and 0.68–0.72 for DMC, HI, FRY and FSY, respectively. It is possible that the values presented here were higher because we adopted a random CV approach and because we had a considerable number of related clones.

[10] analyzed data from three African cassava research institutions and assessed the accuracy of seven prediction models for seven traits. Those authors reported that some trait-dataset combinations exhibited better predictive accuracies than others. For example, for DMC, the estimated accuracies varied from 0.27 to 0.66 across the evaluated breeding programs. According to [34] and [15], when a large number of loci controls the trait, genomic accuracy depends on several factors, among them the training population size, trait heritability and genetic diversity and its relationship with the test population.

The comparisons between the top 10 clone rankings of the one-stage genomic analysis and the four-stage pedigree analysis provided insights about the feasibility of using the clonal evaluation trials' GEBVs for early selection. This approach would reduce the selection interval and consequently minimize the extensive, costly and time-consuming phenotypic evaluations that are usually conducted in a traditional cassava breeding program [35]. Among the traits, the correlation between GEBVs (one-stage genomic analysis) and EBVs (four-stage pedigree analysis) ranged from 0.82 to 0.87, and between 50 and 90% of their top 10 rankings coincided. The clones that originated from the 'Fécula Branca x BRS Formosa' crosses stood out for all of the traits as well as for the selection index. This indicates that this specific combination is favorable for the improvement of cassava productive traits. The feasibility of using genomic selection at the early stages of evaluation is further supported by the fact that the accuracy estimates obtained for the genomic analysis that included only the clonal evaluation trial (one stage) ranged from 0.60 (DMC) to 0.71 (HI). The harvest index, which was the trait that presented the highest accuracy in the one-stage genomic analysis, is an important trait for cassava breeding programs as it is usually used as an indirect indicator of cassava yield.

The genetic gains, calculated separately for each trait (based on the selection of the top 10% of clones by GEBVs obtained by means of the one-stage genotypic analysis), represent great perspectives for cassava breeding (genetic gains of 15.91% for FRY, 5.15% for DMC, 14.76% for DY, 43.11% for FSY, and 8.47% for HI, respectively). [30] calculated the

genetic gains for DMC and SY, selecting 30 out of 471 clones. They obtained genetic gains of 10.75 and 5.50% for DMC and 74.62 and 49.95% for SY when comparing the overall averages of the experiments and controls, respectively. On the other hand, the genetic gains when selecting the best 10% of clones based on the proposed selection index (calculated with GEBVs from the one-stage genotypic analysis) were as follows: 15.13% for FRY, 0.39% for DMC, 14.38% for DY, 24.97% for FSY, and 1.85% for HI. Although the selection gains were lower when based on the selection index compared to the selection by trait individually, with the selection index, it is possible to improve key traits simultaneously [6].

It is worth mentioning that although many of the clones with high selection index values come from the Fécula Branca × BRS Formosa crossing, many others came from other crossings. According to [7], this variability weakens the identity of families and supports the idea that outstanding hybrids can be obtained from essentially every family. This idea arises from the large within-family segregations (due to the fact that progenitors are generally heterozygous) that cassava breeders observe in their evaluation fields.

Traditional assessments based solely on phenotypic information are still very effective and work well for cassava breeding since most of the traits have high heritability. However, here, we wanted to present genomic selection as an option and a great tool to obtain accurate estimates by assessing clones' genomic estimated breeding value earlier in the cassava breeding cycle, thus saving resources such as time, planting area, and direct and indirect costs. The evaluated clones that stood out in the present work could be new varieties to be released in the future and/or could be used as a germplasm source for the development of new varieties with high allele frequencies for starch-related productive traits.

3.5. CONCLUSIONS

The heritability and accuracy values obtained from genomic and pedigree analyses were moderate to high for important traits related to starch production. The results indicate excellent potential for breeding and genomic selection in these cassava populations aimed at increasing starch production for commercial use. In addition, they indicate an optimal potential for selection at early stages of a cassava breeding program (CET), since the correlations between the GEBVs of one-stage genomic analysis and the EBVs of the four-stage pedigree analysis were high for all traits. Moreover, the GEBV-based and EBV-based rankings of the top 10 best clones overlapped by 5 to 9 clones for individual traits, and the rankings of the 10% best clones exhibited 72% overlap for the selection index. The estimated

genetic gains obtained by using the 10% best clones identified by the selection index with the GEBVs from the one-stage genomic analysis were 15.13%, 0.39%, 14.38%, 24.95% and 1.84% for FRY, DMC, DY, FSY and HI, respectively. Thus, genomic selection enables effective early selection in the first stage of clonal evaluation.

3.6. ACKNOWLEDGMENTS

The authors thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), FAPESB (Fundação de Amparo à Pesquisa do Estado da Bahia), and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for financial support. This work was also partially supported by the NEXTGEN Cassava project, through a grant to Cornell University by the Bill & Melinda Gates Foundation and the UK Department for International Development.

3.7. REFERENCES

1. FAO. Cassava for food and energy security. FAO Newsroom. 2008. http://www.fao.org/newsroom/en/news/2008/1000899/index. html. Accessed in 29 Mai 2018.

2. Ceballos H, Kawuki RS, Gracen VE, Yencho GC, Hershey CH. Conventional breeding, marker assisted selection, genomic selection and inbreeding in clonally propagated crops: a case study for cassava. Theoretical and Applied Genetics. 2015; 9:1647. doi: 10.1007/s00122-015-2555-4.

3. Karlström A, Calle F, Salazar S, Morante N, Dufour D, Ceballos H. Biological Implications in Cassava for the Production of Amylose-Free Starch: Impact on Root Yield and Related Traits. Frontiers in Plant Science. 2016; 7: 604.

4. El-Sharkawy M.A. Cassava Biology and Physiology. Plant Molecular Biology. 2004; 56, 481-501.

5. Saithong T, Saerue S, Kalapanilak S, Sojikul P, Narangajavana J, Bhumiratana S. Gene Co-Expression Analysis Inferring the Crosstalk of Ethylene and Gibberellin in Modulating the Transcriptional Acclimation of Cassava Root Growth in Different Seasons. Plos One. 2015; 10(9):e0137602.

6. Barandica OJ, Pérez JC, Lenis JI, Calle F, Morante N, Pino L, Hershey C.H, Ceballos H. Cassava Breeding II: Phenotyphic Correlations through the Different Stages of Selection. Frontiers in Plant Science. 2016; 7:1649. doi: 10.3389/fpls.2016.01649.

7. Ceballos H, Pérez JC, Barandica OJ, Lenis JI, Morante N, Calle F, Pino L, Hershey CH. Cassava Breeding I: The Value of Breeding Value. Frontiers in Plant Science. 2016; 7:1227. doi: 10.3389/fpls.2016.01227.

8. Kayondo SI, Carpio DPD, Lozano R, Ozimati A, Wolfe M, Baguma Y, Gracen V, Samuel O, Ferguson M, Kawuki R, Jannink, J. Genome-wide association mapping and genomic prediction unravels CBSD resistance in a 2 Manihot esculenta breeding population. Scientific Reports. 2018; 8:1549. doi:10.1038/s41598-018-19696-1.

9. Jennings DL and Iglesias CA. Breeding for crop improvement in Cassava: Biology, Production and Utilization. Wallingford: CABI Publishing; 2002.

10. Wolfe MD, Carpio DPD, Alabi O, Ezenwaka LC, Ikeogu UN, Kayondo IS, Lozano R, Okeke UG, Ozimati AA, Williams E, Egesi C, Kawuki RS, Kulakow P, Rabbi IY, Jannink J. Prospects for Genomic Selection in Cassava Breeding. The Plant Genome. 2017; 10:1-19.

11. Sanginga N. Root and Tuber Crops (Cassava, Yam, Potato and Sweet Potato). Feeding Africa. 2015. https://www.afdb.org/fileadmin/uploads/afdb/Documents/Events/DakAgri 2015/Root_and_Tuber_Crops__Cassava__Yam__Potato_and_Sweet_Potato_.pdf. Accessed in 27 Sept 2018.

12. Sakurai T, Mochida K, Yoshida T, Akiyama K, Ishitani M, Seki M. et al. Genome-Wide Discovery and Information Resource Development of DNA Polymorphisms in Cassava. Plos One. 2013; 8(9): e74056.

13. Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, Rokhsar DS, Rounsley S. The Cassava Genome: Current Progress, Future Directions. Tropical Plant Biology. 2012; 5:88-94.

14. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157:1819.

15. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK. Genomic Selection in Plant Breeding: Methods, Model, and Perspectives. Trends in Plant Science. 2017; 22(11):961-975. doi: 10.1016/j.tplants.2017.08.011.

16. Oliveira EJ, Resende MDV, Santos VS, Ferreira CF, Oliveira GAF, da Silva MS, de Oliveira LA, Aguilar-Vildoso CI. Genome-wide selection in cassava. Euphytica. 2012; 187:263.

17. Souza LS, Farias AR, Mattos PLP, Fukuda WMG. Aspectos socioeconômicos e agronômicos da mandioca. Embrapa Mandioca e Fruticultura Tropical; 2006.

18. Kawano K, Fukuda WMG, Cenpukdee U. Genetic and environmental effects on dry matter content of cassava root. Crop Science. 1987; 27:69-74.

19. Doyle JJ and Doyle JL. Isolation of plant DNA from fresh tissue. Focus. 1990; 12, 13-15.

20. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. Plos One. 2011; 6, 1-10.

21. Resende MDV, Azevedo CF, Silva FF, Gois IB. Atualidades da biometria no melhoramento de plantas perenes. In: Desafios biométricos no melhoramento genético. eDOC Brasil; 2017. pp 67-84.

22. Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannick, J. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. The Plant Genome. 2012; 5:103:113.

23. VanRaden PM. Efficient methods to compute genomic predictions. Journal of Dairy Science. 2008; 91:4414.

24. Resende MDV. Selegen-REML/BLUP: Sistema estatístico e seleção genética computadorizada via modelos lineares mistos. Embrapa Florestas; 2007.

25. Covarrubias-Pazaran G. Genome assisted prediction of quantitative traits using the R package sommer. Plos One. 2016; 11(6):1-15. doi:10.1371/journal.pone.0156744.

26. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available in: http://www.R-project.org.

27. Resende MDV, Duarte JB. Precisão e controle de qualidade em experimentos de avaliação de cultivares. Pesquisa Agropecuária Tropical. 2007; 37(3):182-194.

28. Ojulong H, Labuschangne MT, Fregene M, Herselman L. A cassava clonal evaluation trial based on a new cassava breeding scheme. Euphytica. 2008; 160:119-129.

29. Carvalho PRN, Mezette TF, Valle TL, Carvalho CRL, Feltran JC. Avaliação da exatidão, precisão e robustez do método de análise do teor de matéria seca de mandioca (Manihot esculenta, Crantz) por meio da determinação do peso específico (balança hidrostática). Revista raízes e amido tropical. 2007; 3:1-4.

30. Oliveira EJ, Santana FA, Oliveira LA, Santos VS. Genetic parameters and prediction of genotypic values for root quality traits in cassava using REML/BLUP. Genetics and Molecular Research. 2014; 13 (3): 6683-6700.

31. Kawano K, Daza P, Amaya A, Ríos M, Gonçalvez MF. Evaluation of cassava germplasm for productivity. *Crop Science.* 1978; 18:377-380.

32. Farias ARN, Souza LS, Mattos PLP, Fukuda WMG. Aspectos Socioeconômicos e Agronômicos da Mandioca. Embrapa Mandioca e Fruticultura; 2006.

33. Ly D, Hamblin M, Rabbi I, Melaku G, Bakare M, Gauch Jr HG, Okechukwu R, Dixon AGO, Kulakow P, Jannick J. Relatedness and Genotype x Environment Interaction Affect Prediction Accuracies in Genomic Selection: A Study in Cassava. Crop Science. 2013; 53:1312-1325.

34. Resende MDV. Genética Quantitativa e de Populações. Editora Suprema; 2015.

35. Brito AC, Oliveira SAS, Oliveira EJ. Genome-wide association study for resistance to cassava root rot. Journal of Agricultural Science. 2017; 155:1424-1441.

3.8. SUPPLEMENTARY MATERIAL

S1 Table. Comparison of top 10 rankings based on genomic estimated breeding value (one evaluation stage) or on estimated breeding value (four stages) for fresh root yield (FRY, in t ha^{-1}).

Correlation between GEBVs (one stage genomic analysis) and EBVs (four stages pedigree analysis) = 0.82							
Genomic analysis – One stage				**Pedigree analysis – Four stages**			
Clone	**GEBV**	**Male genitor**	**Female genitor**	**Clone**	**EBV**	**Male genitor**	**Female genitor**
2012_108_043	56.52	Fécula Branca	BRS Formosa	2012_108_043	55.55	Fécula Branca	BRS Formosa
2012_108_208	53.67			2012_108_208	53.75		
2012_108_060	48.57			2012_108_108	52.34		
2012_108_108	48.07			2012_108_060	50.88		
2012_108_046	47.64			2012_108_188	48.98		
2012_108_215	46.32			2012_108_046	48.48		
2012_108_188	45.97			2012_108_143	45.21		
2012_108_035	44.95			2012_108_155	43.36		
2012_108_143	43.12			2012_108_035	42.68		
2012_108_036	43.07			2014_006_33	42.45	BGM-1683	Fécula Branca

S2 Table. Comparison of top 10 rankings based on genomic estimated breeding value (one evaluation stage) or on estimated breeding value (four stages) for dry matter content (DMC, in %).

Correlation between GEBVs (One stage genomic analysis) and EBVs (Four stages pedigree analysis) = 0.83							
Genomic analysis – One stage				**Pedigree analysis – Four stages**			
Clone	**GEBV**	**Male genitor**	**Female genitor**	**Clone**	**EBV**	**Male genitor**	**Female genitor**
2014_013_41	38.32	Equador72	BGM-0728	2012_108_155	37.21	Fécula Branca	BRS Formosa
2014_008_19	38.04	Equador72	BGM-1124	2012_108_125	36.90	Fécula Branca	BRS Formosa
2012_108_155	37.38	Fécula Branca	BRS Formosa	2014_008_19	36.82	Equador72	BGM-1124
2014_012_13	36.68	Cascuda	BGM-0728	2012_108_049	36.68	Fécula Branca	BRS Formosa
2012_108_206	36.45	Fécula Branca	BRS Formosa	2012_108_061	36.57	Fécula Branca	BRS Formosa
2014_013_24	36.44	Equador72	BGM-0728	2014_025_42	36.47	BGM-1662	Fécula Branca
2012_108_030	36.34	Fécula Branca	BRS Formosa	2012_108_030	36.43	Fécula Branca	BRS Formosa
2011_52_01	36.33	BRS Mulatinha	BRS Mulatinha	2012_108_098	36.39	Fécula Branca	BRS Formosa
2014_008_11	36.32	Equador72	BGM-1124	2012_108_206	36.32	Fécula Branca	BRS Formosa
2012_108_125	36.25	Fécula Branca	BRS Formosa	2012_108_190	36.20	Fécula Branca	BRS Formosa

S3 Table. Comparison of top 10 rankings based on genomic estimated breeding value (one evaluation stage) or on estimated breeding value (four stages) for dry yield (DY, in t ha^{-1}).

Correlation between GEBVs (one stage genomic analysis) and EBVs (four stages pedigree analysis) = 0.82							
Genomic analysis – One stage				Pedigree analysis – Four stages			
Clone	GEBV	Male genitor	Female genitor	Clone	EBV	Male genitor	Female genitor
2012_108_043	18.77	Fécula Branca	BRS Formosa	2012_108_043	18.49	Fécula Branca	BRS Formosa
2012_108_208	17.54			2012_108_108	17.72		
2012_108_215	16.39			2012_108_208	17.51		
2012_108_108	16.00			2012_108_060	16.43		
2012_108_060	15.62			2012_108_188	15.94		
2012_108_035	15.50			2012_108_155	15.81		
2012_108_155	15.17			2012_108_046	15.46		
2012_108_046	14.98			2012_108_035	14.88		
2012_108_188	14.95			2012_108_036	14.45		
2012_108_036	14.67			2012_108_143	14.27		

S4 Table. Comparison of top 10 rankings based on genomic estimated breeding value (one evaluation stage) or on estimated breeding value (four stages) for fresh shoot yield (FSY, in t ha^{-1}).

Correlation between GEBVs (one stage genomic analysis) and EBVs (four stages pedigree analysis) = 0.87							
Genomic analysis – One stage				Pedigree analysis – Four stages			
Clone	GEBV	Male genitor	Female genitor	Clone	EBV	Male genitor	Female genitor
2014_018_06	49.60	Equador72	Eucalipto	2012_108_185	38.43	Fécula Branca	BRS Formosa
2012_108_222	37.21	Fécula Branca	BRS Formosa	2012_108_208	34.85	Fécula Branca	BRS Formosa
2014_013_28	36.02	Equador72	BGM-0728	2014_013_28	34.81	Equador72	BGM-0728
2012_108_185	34.42	Fécula Branca	BRS Formosa	2012_108_222	32.18	Fécula Branca	BRS Formosa
2012_108_224	31.46	Fécula Branca	BRS Formosa	2014_018_06	31.70	Equador72	Eucalipto
2014_013_07	29.59	Equador72	BGM-0728	2012_108_188	31.68	Fécula Branca	BRS Formosa
2012_108_208	29.42	Fécula Branca	BRS Formosa	2012_108_224	30.21	Fécula Branca	BRS Formosa
2012_108_043	29.25	Fécula Branca	BRS Formosa	2014_025_42	30.11	BGM-1662	Fécula Branca
2014_013_30	29.01	Equador72	BGM-0728	2014_026_33	29.88	BGM-1683	Fécula Branca
2012_106_160	28.70	Fécula Branca	BRS Jari	2012_108_108	29.81	Fécula Branca	BRS Formosa

S5 Table. Comparison of top 10 rankings based on genomic estimated breeding value (one evaluation stage) or on estimated breeding value (four stages) for harvest index (HI, in %).

Correlation between GEBVs (one stage genomic analysis) and EBVs (four stages pedigree analysis) = 0.84							
Genomic analysis – One stage				**Pedigree analysis – Four stages**			
Clone	**GEBV**	**Male genitor**	**Female genitor**	**Clone**	**EBV**	**Male genitor**	**Female genitor**
2012_108_034	89.61	Fécula Branca	BRS Formosa	2012_108_179	88.56	Fécula Branca	BRS Formosa
2012_108_179	88.78			2012_108_021	88.24		
2012_108_205	88.69			2012_108_034	86.49		
2012_108_149	88.38			2012_108_149	86.44		
2012_108_021	88.12			2012_108_205	86.25		
2012_108_144	88.00			2012_108_029	86.15		
2012_108_187	87.46			2012_108_162	86.12		
2012_108_006	87.20			2012_108_143	85.14		
2012_108_035	86.92			2012_108_006	84.87		
2012_108_162	86.91			2012_108_161	84.85		

4. CHAPTER II: CAN CROSS-COUNTRY GENOMIC PREDICTIONS BE A REASONABLE STRATEGY TO SUPPORT GERMPLASM EXCHANGE? – A CASE STUDY WITH HYDROGEN CYANIDE IN CASSAVA

Abstract - Genomic prediction (GP) offers great opportunities for accelerated genetic gains by optimizing the breeding pipeline. One of the key factors to be considered is how the training populations (TP) are composed in terms of genetic improvement, kinship/origin and their impacts on GP. Hydrogen cyanide content (HCN) is a determinant trait to guide cassava's products usage and processing. This work aimed to: i) evaluate the feasibility of using cross-country (CC) GP between germplasms's accessions of Embrapa Mandioca e Fruticultura (Embrapa, Brazil) and The International Institute of Tropical Agriculture (IITA, Nigeria) for HCN; ii) provide an assessment of population structure for the joint dataset; iii) estimate the genetic parameters based on single nucleotide polymorphisms (SNPs) and on a haplotype-approach. Datasets of HCN from Embrapa and IITA breeding programs were analyzed, separately and jointly, with 1,230, 590 and 1,820 clones, respectively. After quality control, ~14K SNPs were used for GP. The genomic estimated breeding values (GEBVs) were predicted based on SNP effects from analyses with TP composed by: i) Embrapa genotypic and phenotypic data, ii) IITA genotypic and phenotypic data, and iii) the joint datasets. Comparisons on GEBVs' estimation were made considering the hypothetical situation of not having the phenotypic characterization for a set of clones for a certain research institute/country and might need to use the markers' effects that were trained with data from other research institute/country's germplasm to estimate their clones' GEBV. Fixation index (F_{ST}) among the genetic groups identified within the joint dataset ranged from 0.002 to 0.091. The joint dataset provided an improved accuracy (0.80-0.85) compared to the prediction accuracy of either germplasm' sources individually (0.51-0.67). CC GP proved to have potential use under the present study's scenario, the correlation between GEBVs predicted with TP from Embrapa and IITA was 0.55 for Embrapa's germplasm, whereas for IITA's it was 0.10. This seems to be among the first attempts to evaluate the CC GP in plants. As such, a lot of useful new information was provided on the subject, which can guide new research on this very important and emerging field.

Keywords: *Manihot esculenta*. Cross-country genomic prediction. Genomic population structure. haplotype prediction. HCN.

4.1. INTRODUCTION

Cassava (*Manihot esculenta* Crantz) is an important staple root crop, food source for millions of people mostly in developing countries and used for several industrial applications. Subsistence farmers rely on cassava as a vital source of dietary energy, as it can be planted and harvested throughout the year, tolerates periods of drought and grows on marginal soils (HILLOCKS et al., 2002). It has been suggested that cassava can be highly resilient to in a climate change scenario; providing adaptation opportunities that are not offered by any other basic food crops (JARVIS et al., 2012).

Despite of exciting crop's future prospects, cassava roots, peels and leaves contain two cyanogenic glucosides, linamarin and lotaustralin, mainly linamarin. Linamarin and lotaustralin, in the acids and enzymes' presence, go under hydrolysis that leads to CN^- production and, consequently, hydrogen cyanide (HCN) (CEREDA; MATTOS, 1996). Although cyanogenic glycosides play important roles such as defense against herbivores and transport forms of reduced nitrogen in some plant species (POULTON, 1990; WHITE et al., 1998), the high content of cyanogenic glycosides (consequently liberated HCN), together with the rapid tuber perishability, represents one of the main shortcomings associated with cassava consumption.

HCN is a toxic compound that may be harmful to human and animal health, the potential toxicity depends on amount, route and duration of the exposure. The toxicological effects of cyanide absorption due to intake of improperly processed cassava occur in acute and chronic form, it can affect the nervous, respiratory, cardiovascular and endocrine systems (WHO, 2004), causing symptoms like headache, dizziness, nausea, vomiting, weakness, among others.

Cassava varieties are classified in bitter and sweet depending on the content of HCN in the roots. This component varies substantially with plant variety, environment conditions, time of harvest and post-harvest practices (OJO et al., 2013). Those with concentration of cyanogenic glycosides greater than 50 mg per kg HCN on fresh weight basis are classified as bitter cassava, whereas sweet varieties contain less than 50 mg per kg (OBI et al., 2019). Bitter varieties require more efforts in terms of processing. Processing is effective to ensure reduction of cyanogenic compound content in cassava food products, as examples of processing steps: peeling, grating, soaking, boiling, cooking, among others, the combination of procedures within the production system essentially varies depending on the intended end-product, according to the Code of Practice for the Reduction of Hydrocyanic Acid (HCN) in

Cassava and Cassava Products (CAC/RCP 73-2013). Besides the most obvious relevance of the identification of sweet cassava varieties, which is to ensure safe recommendations for food consumption, Obueh and Kolawole (2016) described in comparative study, that sweet varieties are also associated with high nutritional quality (Ca, Mg, P, Zn, Cu and amino acid contents) and low anti-nutrients content (oxalate, tannin, phytate, alkaloid, lignin and saponin).

Traditional cassava genetic improvement strategies are still very demanding in terms of financial resources; and, historically, cassava is a crop that receives less investment than other commodities crops (BART; TAYLOR 2017). However, this scenario is changing, possibly encouraged by the following reasons: i) progress constraints of conventional breeding due the crop biology – e.g. long breeding cycle, slow multiplication rate (CEBALLOS et al. 2020) –, ii) the importance of cassava for food security and poverty alleviation, iii) the increase of challenges in the face of biotic stresses and, also, iv) the increased commercial interest in cassava, due to the better starch properties than cereals. Promising results with a recently sequenced genome (PROCHNIK et al., 2012; BREDESON et al., 2016) and SNP-based genetic linkage maps (INTERNATIONAL CASSAVA GENETIC MAP CONSORTIUM, 2014), was obtained in a short period of time, opening a path to the study of key traits' genetic architecture through genome wide association studies (GWAS) (ESUMA et al., 2016; WOLFE et al., 2016; BRITO et al., 2017; RABBI et al., 2017; KAYONDO et al., 2018; SOMO et al., 2020; OGBONNA et al., 2020a) and trait improvement by means of genomic predictions (GP) and selection (OLIVEIRA et al., 2012; OKEKE et al., 2017; WOLFE et al., 2017; ELIAS et al., 2018; ANDRADE et al., 2019; IKEOGU et al., 2019; TORRES et al., 2019; SOMO et al., 2020; YONIS et al., 2020).

Among the methodologies currently used to accelerate genetic gain and to shorten the breeding cycle interval, genomic selection proposed by Meuwissen et al. (2001) offers great opportunities. Besides the reduction on the selection interval, Werner et al. (2020) pointed two other additional key advantages: the increased selection accuracy and the possibility of increasing the selection intensity. Genomic prediction can also be extremely useful when applied for a trait which is expensive and/or hard to measure, restricting the phenotypic data's availability to a subset of the population (DE HAAS et al., 2012).

Implementation of genomic prediction/selection schemes were recently boosted by the advent of low-cost SNPs marker platforms such as GBS (ELSHIRE et al., 2011). Genotyping must be sufficiently dense to get SNPs that are in linkage disequilibrium with most of the loci controlling important quantitative traits (OKEKE et al., 2017). Despite being a trivial

practice for some plant breeding programs, and especially for animal breeding programs, some plant breeding programs are still facing the transition's challenges from conventional plant breeding to a present scenario where conventional breeding is still essential but now it can also be supported by genome-assisted breeding evidence to drive decision making, by adding relevant information to speed up the breeding pipeline.

Selection decisions are based on genomic breeding values (GEBVs), which are calculated by first estimating the SNP effects in a training population (reference) with both phenotypic and genotypic information available; and then, the SNP effects are multiplied by the genotypes' marker matrix of a testing population (target) and summed to form the GEBVs (PRYCE et al., 2011). One of the essential factors to be considered is how the training sets are composed; they can be constituted of landraces and/or modern improved lines, clones from different breeding programs, across breeding stages and/or even from different countries (SOMO et al., 2020).

Joint evaluations can be desirable under certain situations, its practice is currently more common in animal breeding as crossbred predictions (POCRNIC et al., 2019; VANRADEN et al., 2020). However, for plants, recently Somo et al. (2020) attempted to examine diverse cassava germplasm assembled from two breeding programs in Tanzania at different breeding stages to predict traits. In their work, the use of clones in the same breeding stage to build TPs provided higher prediction accuracy than TPs with clones from multiple breeding stages together. They also included a Ugandan TP in either Tanzanian population and it did not prove successful to improve prediction accuracies.

This work aimed to: i) evaluate the feasibility of cross-country genomic predictions between germplasm's accessions from Embrapa (Cruz das Almas, Brazil) and IITA (Ibadan, Nigeria) for HCN (taking into consideration the hypothetical scenario of lack of phenotypic characterization of clones from a certain research institute/country and the need of predicting their GEBVs with SNPs effects predicted with a TP consisted of clones from other research institute/country, which may contemplate cases of interest, e.g. plant diseases); ii) identify the HCN variability in the cassava germplasm; iii) provide an assessment of population structure in the joint dataset (Embrapa+IITA); also iv) estimate the genetic parameters of GP based on a haplotype-approach and compare them to those based on single markers. This seems to be among the first attempts to evaluate the cross-country genomic prediction in cassava.

4.2. MATERIAL AND METHODS

4.2.1. Plant material

A total of 1,230 cassava clones from the Cassava Breeding Program (CBP) of Embrapa Mandioca e Fruticultura (Cruz das Almas, Brazil) were genotyped and phenotyped for HCN. The set of clones consisted of landraces and modern breeding lines, collected from different Brazilian growing regions. It consisted of a selected subset from a panel of 1,536 clones, which were revealed as a unique subset out of an original set of 3,354 clones by identity-by-state analysis carried out by Ogbonna et al. (2020b). The dataset from the International Institute of Tropical Agriculture (IITA) was obtained from the open-source cassava bredbase (https://cassavabase.org/), it included the genotypic and phenotypic information of 590 cassava clones evaluated in Ibadan, Nigeria.

4.2.2. Phenotypic characterization of HCN

Field trials were carried out in Cruz das Almas (State of Bahia, Brazil, 12º40'39"S, 39º06'23"W) from 2016 to 2019, with the number of replications varying from three replications (2017, 2018, 2019) to four replications (2016) in a randomized block design. The plots consisted of two rows with 16 plants per plot. Planting was performed from May to July (during the rainy season). Spacings between rows and plants were 0.90 and 0.80 meters, respectively. All trial management was performed, whenever necessary, in accordance with the technical recommendations and standard agricultural practices for cassava. For the IITA dataset, phenotypic information was collected from Cassavabase (https://cassavabase.org/), the final file consisted of 590 clones evaluated in Ibadan, Nigeria, from 1998 to 2012, the number of replications was mostly four, but it ranged from two to six for some clones. This dataset was originated from a previous curatorship in which Ogbonna et al (2020a) retrieved 228 trials out of 393 trials, and then we further filtered by location (Ibadan) and years (1998-2012). The total number of observations accounted was 8,355 and 5,158 for Embrapa and IITA datasets, respectively (Table 1).

Table 1. Experimental areas, cities, geographical coordinates, years, number of phenotypical observations and clones.

Institute	City	Geographical Coordinates	Years	Observations	Clones
Embrapa	Cruz das Almas	12°40'42.4"S 39°05'27.8"W	2016-2019 (4 years total)	8,355	1,230
IITA	Ibadan	7°29'44.5"N 3°53'51.4"E	1998-2012 (15 years total)	5,158	590

For both datasets, HCN was measured according to the picrate titration method described by Fukuda et al. (2010), that is a qualitative determination of HCN content in fresh roots basis based on a 1-9 color scale; with 1 and 9 representing the extremes of low and high HCN, respectively. For HCN Embrapa dataset, roots with uniform shape and size from different plants that represent the plots were analyzed for HCN. Plants were harvested 11-12 months after planting. Cross sectional 1 cm^3 cut was made at mid-root position, between the peel and the parenchyma center. The cube of root and five drops of toluene were added to a glass test tube and they were tightly sealed with a stopper. In order to determine the qualitative score of HCN content, a strip of filter paper was dipped into a freshly prepared alkaline picrate mixture until saturation, then, the saturated filter paper was placed above the root cube in the tube. Tubes were sealed for 10-12 hours before the color intensity evaluation. HCN represented the total cyanogenic glucosides (HCN/CN^-, linamarin and acetone cyanohydrin) in cassava root (BRADBURY et al., 1999).

4.2.3. Genotyping and data quality control

The DNA was extracted from young leaves of cassava accessions according to a protocol described by Albuquerque et al. (2018), Ogbonna et al. (2020a) and Ogbonna et al. (2020b), for Embrapa panel of clones, briefly it used cetyltrimethylammonium bromide according to the protocol of Doyle and Doyle (1990). Then, the DNA was diluted in TE buffer (10 mM Tris-HCl and 1 mM EDTA) and adjusted to a final concentration of 60 ng/µl. The DNA quality was checked by the digestion of 250 ng of genomic DNA from 10 random samples with the restriction enzyme *Eco*RI (New England Biolabs, Boston, MA) and, then, the samples were sent to the laboratory for library preparation, sequencing and bioinformatics analyses. Genotyping was performed by genotyping by sequencing (GBS) methodology, which is described in detail by Elshire et al. (2011). The DNA was digested with the *Ape*KI restriction enzyme and the Illumina sequencing read lengths of 150 bp.

Marker genotypes were called with the TASSEL GBS pipeline V5 (GLAUBITZ et al., 2014) and aligned to the cassava reference genome version 6.1. The joint dataset (Embrapa + IITA) was intersected using the GATK CombineVariants and Intersect functions (McKENNA et al., 2010).

Filtering steps and parameters as well as imputation were performed as described in Ogbonna et al. (2020a). Filtering steps and parameters included: mean depth values greater than 5, missing data up to 0.20 per loci and minor allele frequency of 0.01 per loci, also, individuals with more than 0.80 of missing data per chromosome were removed. Imputation was performed for each chromosome using beagle 4.0 (BROWNING; BROWNING, 2009). Imputed markers were subjected to further filtering using an allelic correlation greater or equal to 0.80. Final cassava clone set was obtained after filtering available phenotypic data on trial location and years basis, additional minor allele frequency filtering (0.05) step was carried out, achieving total sets with 1,230 clones and 14,323 SNPs for Embrapa, 590 clones and 13,524 SNPs for IITA, 1,820 clones and 14,924 SNPs for the joint dataset (Embrapa + IITA) (Table 2) out of an intersected file with 4,814 clones and 17,773 SNPs.

Table 2. Number of single nucleotide polymorphisms (SNP) in each data file after quality control, and number of shared SNPs between data files after quality control[/1].

Analyses	# SNPs
Embrapa	14,323
IITA	13,524
Embrapa+IITA	14,924
Common SNPs	11,883

[/1] SNPs with minor allele frequency > 0.05 were kept.

4.2.4. Population structure

The principal component analysis (PCA) and discriminant analysis of principal components (DAPC) analyses were performed with the joint dataset (Embrapa+IITA; 1820 clones; 14,924 SNPs), with the *prcomp* function and the package "adegenet" (JOMBART, 2008, JOMBART et al., 2010) in R software (R CORE TEAM, 2019), respectively. The PCA analysis and the PCs components reported were based on the genomic kinship coefficients between clones. However, for DAPC analysis the number of retained PCs (explained ~ 90%) were based on the PCA of the marker matrix. Bayesian Information Criterion (BIC) was used to identify the optimal number of clusters. Additional clustering step was carried out by Tocher's method (https://www.rdocumentation.org/packages/

biotools/versions/3.1/topics/tocher), grouping the DACP clusters based on their HCN means.

The fixation index (FST) (WRIGHT, 1965) between germplasm's sources (Embrapa and IITA) as well as between clusters identified in the joint dataset's DAPC analysis was estimated by using the method of Weir and Cockerham (1984) with the package "hierfstat" (GOUDET; JOMBART, 2015) in R software (R CORE TEAM, 2019).

4.2.5. Building the haplotype matrix

SNPs obtained after quality control were used to build the haplotype blocks. Haplotypes were identified by the Gabriel's method (GABRIEL et al., 2002), implemented on PLINK 1.90b5.3 (PURCELL et al., 2007). The method is based on a confidence interval of DPrime (D'). The pairs of SNPs were considered in strong linkage disequilibrium (LD) if the upper boundary confidence interval of D′ was higher than 0.98 and the lower boundary was higher than 0.80. The maximum length of the blocks was set to 200 Kb. We assumed very strong linkage between markers within the haplotype blocks and, then, the 'haplotype matrix' was numbered (0, 1) as it follows: for each haplotype block the class '0' was the one that contained exclusively 0 for all the haplotype blocks' markers, whereas the class '1' consisted of all remaining haplotypes.

4.2.6. Genomic prediction analyses

Clone's GEBVs for HCN were predicted based on SNP effects estimated by the following analyses: i) genomic prediction with training population of Embrapa genotypic and phenotypic data, ii) genomic prediction with training population of IITA genotypic and phenotypic data, and iii) genomic prediction with training population of the joint (Embrapa+IITA) datasets. The SNP effects were estimated in the training populations, and then, these three vectors of SNPs effects were multiplied by the genotype's incidence codes of the mutual testing populations and summed to form the GEBVs.

The models used for the genomic prediction analyses by single markers and haplotypes were:

$$\mathbf{y_1} = \mathbf{X_1 b_1} + \mathbf{Z_{11}} \boldsymbol{a_{G1}} + \mathbf{Z_{21}} \boldsymbol{a_{YR1}} + \boldsymbol{\varepsilon_1}$$
$$\mathbf{y_2} = \mathbf{X_2 b_2} + \mathbf{Z_{12}} \boldsymbol{a_{G2}} + \mathbf{Z_{22}} \boldsymbol{a_{YR2}} + \boldsymbol{\varepsilon_2}$$
$$\mathbf{y_3} = \mathbf{X_3 b_3} + \mathbf{Z_{13}} \boldsymbol{a_{G3}} + \mathbf{Z_{23}} \boldsymbol{a_{YR3}} + \boldsymbol{\varepsilon_3}$$

where $\mathbf{y_1}$, $\mathbf{y_2}$ and $\mathbf{y_3}$ are the vectors of phenotypic observations and $\boldsymbol{b_1}$ and $\boldsymbol{b_2}$ are the scalars of mean fixed effect, for analyses of datasets of Embrapa and IITA separately, respectively, and $\boldsymbol{b_3}$ is the vector of fixed effects of country added to the mean for the joint dataset's analysis. The vectors $\boldsymbol{a_{Gij}}$ refers to the random additive genomic clone effects associated with the i-th dataset ($i = 1, 2, 3$), $\boldsymbol{a_{Gij}} \sim$ N($\mathbf{0}$, $\sigma^2_{aGij}\mathbf{G_{ij}}$) where $\mathbf{G_{ij}}$ is the additive genomic relationship matrices as defined bellow, σ^2_{aGij} is the genetic variance associated with $\boldsymbol{a_{Gij}}$. The random effects vectors $\boldsymbol{a_{YRi}}$ refers to the combination of year and replication associated with the i-th dataset ($i = 1, 2, 3$), $\boldsymbol{a_{YRi}} \sim$ N($\mathbf{0}$, $\sigma^2_{\boldsymbol{aYRi}}\mathbf{I_{yri}}$), where $\mathbf{I_{yri}}$ is an identity matrix and $\sigma^2_{\boldsymbol{aYRi}}$ is the variance estimate due the combination of year and replication. $\mathbf{X_i}$, $\mathbf{Z_{1i}}$ and $\mathbf{Z_{2i}}$ are the incidence matrices for fixed effects and the random effects of genotypes and combination of year and replication, respectively, associated with the i-th dataset ($i = 1, 2, 3$). $\boldsymbol{\varepsilon_i}$ represents the residual vector associated with the i-th dataset ($i = 1, 2, 3$), $\boldsymbol{\varepsilon_i} \sim$ N($\mathbf{0}$, $\sigma^2_{ei}\mathbf{I_{ei}}$), where $\mathbf{I_{ei}}$ is an identity matrix and σ^2_{ei} is the residual variance. The genomic predictions were based on the GBLUP (genomic best linear unbiased prediction) method, with the REML (restricted maximum likelihood) estimation of variance components. The whole procedure uses the additive genomic relationship matrix $\mathbf{G_{ij}}$. Likelihood Ratio Tests (LRT) were undertaken to test the random effects using the chi-square distribution, with degrees of freedom equal the difference in the number of parameters for the two models. Genomic prediction analyses were performed using the package "sommer" (COVARRUBIAS-PAZARAN et al., 2016) in R software (R CORE TEAM, 2019).

Heritability was estimated as the ratio of the genetic variance to the sum of the genetic variance, the variance due to the combination of year and replication and the residual variance, for each analysis.

To capture genetic relatedness, the genomic relationship matrix ($\mathbf{G_{ij}}$; with the i-th dataset ($i = 1, 2, 3$) and the j-th genotyping method (j = SNP, hap) for the genomic prediction of single markers and haplotypes were obtained as it follows:

$$\mathbf{G_{i\_SNP}} = \frac{\mathbf{MM'}}{\sum_{i=1}^{n} 2p_i q_i}$$

$$\mathbf{G}_{i\_hap} = \frac{\mathbf{HH}'}{\sum_{i=1}^{n} p_i q_i}$$

where p_i and q_i are the allele frequencies for single markers' analyses and the 'allele' frequencies for haplotypes' analyses, **M** is the markers incidence matrix (numbered as 0, 1, 2) centered by the mean $2p_i$ and **H** is the haplotypes incidence matrix (numbered as 0, 1). The sizes of $\mathbf{G}_{ij}$ matrixes were 1230×1230, 590×590 and 1820×1820, for analyses (datasets) 1, 2 and 3, respectively. And the sizes of $\mathbf{I}_{yri}$ identity matrixes for the combination of year and replication were 26×26, 60×60 and 86×86, for analyses 1, 2 and 3, respectively.

The SNP and haplotype effects vectors ($\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{h}}$) were calculated as it follows:

$$\widehat{\boldsymbol{m}} = (\boldsymbol{MM}')^{-1}\boldsymbol{M}'\widehat{\boldsymbol{a}}_{\boldsymbol{Gij}}$$
$$\widehat{\boldsymbol{h}} = (\boldsymbol{HH}')^{-1}\boldsymbol{H}'\widehat{\boldsymbol{a}}_{\boldsymbol{Gij}}$$

The vector $\boldsymbol{a}_{\boldsymbol{Gij}}$ refers to the random additive genetic effects ($\boldsymbol{a}_{\boldsymbol{Gij}} \sim N(0, \sigma^2_{aGij}\mathbf{G}_{ij})$) associated with the i-th dataset ($i = 1,2,3$) and the j-th genotyping method (j = SNP, hap). The GEBVs' vectors were obtained as follows: $\widehat{\boldsymbol{GEBV\_SNP}} = \boldsymbol{M}\widehat{\boldsymbol{m}}$ or $\widehat{\boldsymbol{GEBV\_hap}} = \boldsymbol{H}\widehat{\boldsymbol{h}}$, for using SNPs or haplotypes, respectively.

For the cross-country analyses, simulating the hypothetical situation where we would not have the phenotypic information of a group of clones from a certain research institute/country, for the vector of GEBV prediction, we have used the estimated SNP effects with data from other research institute/country, after filtering the matrix M to encompass only common SNPs that were in both datasets.

Trait estimates of predictive ability, accuracy and bias were calculated from cross-validation with the training sets for each analysis. Those parameters were estimated for each fold, and the value presented in this study is the mean of the folds. The cross-validation method used was the k-fold, k=10, with the clones being randomly assigned to each fold. The training set, composed of 9 of the 10 subsets, was used to estimate marker effects and the remaining subset was the validation set. These marker effects estimates were used to predict the genomic breeding values of the validation set individuals. This process was repeated until all 10 subsets was used as the validation population once.

The predictive ability was given by the correlation coefficient between predicted genetic values (GEBVs) and the phenotypes in the validation population. The accuracy was calculated as the ratio between the predictive ability and the square root of the phenotypic trait heritability. The phenotypes (PP) were linearly regressed on the GEBVs, and the regression coefficient $\hat{b}$PP,GEBV was used to measure the degree of bias of the GEBV prediction. The bias relates to the size of the absolute differences between clones' predicted genetic values and their phenotypes. The estimated magnitude of these differences can be quantified by the $\hat{b}$PP,GEBV regression coefficient and can be overestimated ($\hat{b}$PP,GEBV < 1) or underestimated ($\hat{b}$PP,GEBV > 1). A regression coefficient equal to one indicates no bias. Then, here we will represent bias as one unit minus the regression coefficient $\hat{b}$PP,GEBV (Bias = 1 - $\hat{b}$PP,GEBV).

Statistical analyses were based on the HCN phenotypes originated by the picrate method 1-9 color scale; however, for data visualization clone's GEBV were further classified into three classes: sweet, intermedium and bitter, as clones ranged from 1.0 to 4.0, from 4.1 to 5.0 and from 5.1 to 9.0, respectively.

4.3. RESULTS

4.3.1. Population structure

Principal components analysis in our joint dataset revealed a clear pattern of grouping clones according to their origin (Embrapa/Brazil and IITA/Nigeria), with the first three PCs (based on the genomic kinship coefficients) accounting for 76.3% of the genetic variation (58.24, 12.21 and 5.85% for PC1, PC2 and PC3, respectively) (Fig. 1A).

For DAPC, 500 PCs (based on markers matrix) that explained ≈ 90% of the genetic variance were kept in our dataset. Based on the Bayesian Information Criterion (BIC), an optimal number of 12 genetic clusters was suggested to explain the total genetic variability (Supplementary Fig. 1), with cluster size ranging from 61 (cluster 10) to 299 individuals (cluster 2). Out of the 12 clusters, only four had clones from both Embrapa and IITA germplasm (clusters 1, 2, 6 and 7) with unbalanced proportions and a dominant origin within those clusters. The remaining eight clusters were exclusively composed by clones from Embrapa (clusters 5, 9, 10, 11 and 12) or IITA (clusters 3, 4 and 8). The HCN means ranged

from 3.75±0.68 (cluster 10) to 7.15±0.84 (cluster 5) (Table 3). Principal components plot according to the clustering from DAPC analysis is shown on Fig. 1B.

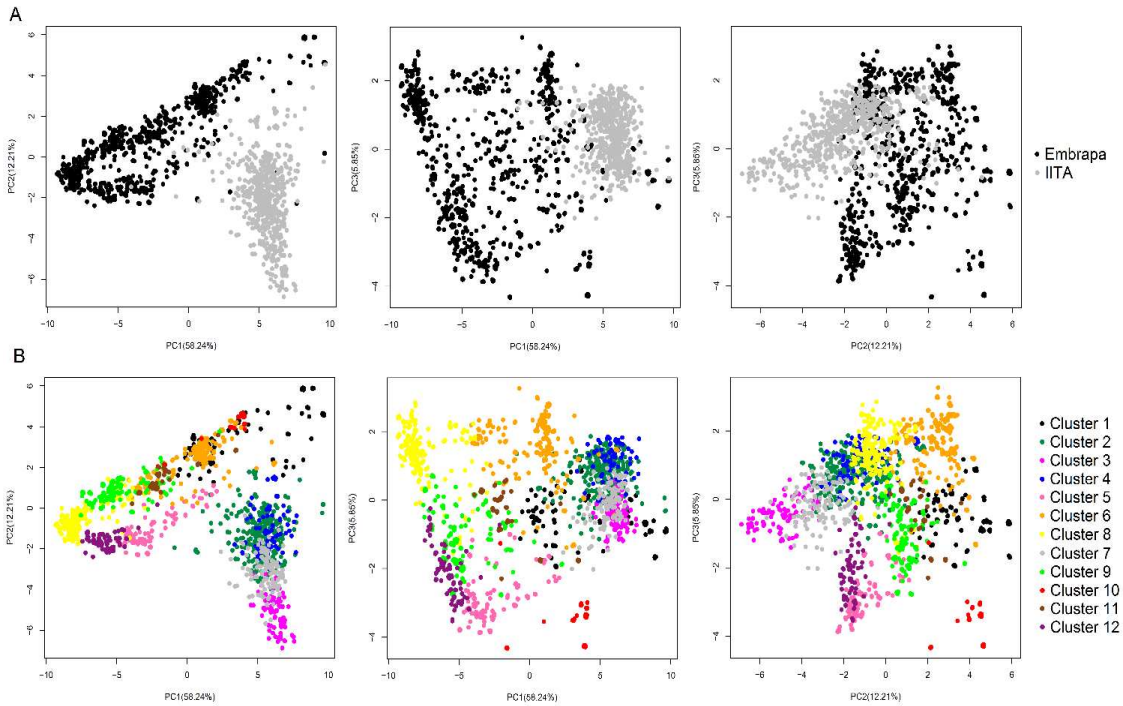


Fig 1. Principal components analysis (PCA) of the genomic kinship coefficients between cassava clones. (A) PCA from the genomic relationship matrix between all cassava clones (N=1,820; 14,924 SNPs) from Brazil (Embrapa) and Nigeria (IITA), showing the first three principal components and the variance explained by each component in parenthesis on the corresponding axis (58.24, 12.21 and 5.85% for PC1, PC2 and PC3, respectively). In black representing Embrapa clone's dispersion and in grey representing IITA's. (B) PC diagram highlighting the 12 clusters identified by the DAPC analysis.

Table 3. Discriminant analysis of principal components (DAPC) that accounted for most (> 95%) of the total genetic variability, and genetic clusters were inferred (k=12) based on Bayesian Information Criterion (BIC). Total number of clones (N), number of clones from Embrapa (n_1) and number of clones in IITA (n_2) in each cluster.

Cluster	N(Total)[/1]	n_1(Embrapa)	n_2(IITA)	Mean HCN[/2]
1	143	134	9	4.39±0.94 [a]
2	299	10	289	5.37±0.84 [b]
3	86	-	86	5.51±0.89 [b]
4	79	-	79	5.53±0.79 [b]
5	133	133	-	7.15±0.84 [c]
6	269	255	14	5.44±1.38 [b]
7	297	296	1	6.80±0.97 [c]
8	112	-	112	5.60±1.02 [b]
9	173	173	-	5.26±1.37 [b]
10	61	61	-	3.75±0.68 [d]
11	68	68	-	4.69±1.37 [a]
12	100	100	-	7.08±0.71 [c]
Total	1820	1230	590	5.57±1.38

[/1] Size of the dataset used for the DAPC: 1820 clones and 14,924 SNPs.
[/2] Clustering by the Tocher method (Inter-cluster distance limit = 0.64).
[a, b, c, d] Mean clustering of the clusters by Tocher's method.

Additional clustering step was carried out by Tocher's method, grouping the DACP clusters based on their HCN means. The intergroup distance limit was 0.64; DACP clusters 5 (7.15±0.84), 7 (6.80±0.97) and 12 (7.08±0.71) were allocated in the same Tocher group, high chances of the bitterest clones belonging to that group of clusters. They were exclusively composed by clones from Embrapa, except for cluster 7 which had 296 clones from Embrapa and only one from IITA. On the other hand, the 'sweetest' (low HCN) Tocher's group was composed by the cluster 10 (3.75±0.68), only, exclusively composed by clones from Embrapa.

The F_{ST} values between the proposed mutual training and validation sets, composed by Embrapa and IITA's genotypic datasets, were moderate (0.072). Estimated fixation indexes with markers between the twelve clusters identified within the joint dataset (Embrapa+IITA) varied from 0.002 (clusters 2 and 4, clusters 4 and 6) to 0.091 (clusters 3 and 9) (Table 4). Cluster 2 (mean HCN = 5.37±0.84) was composed mainly of clones from IITA, while cluster 4 (mean HCN = 5.53±0.79) was composed exclusively by clones from IITA, whereas cluster 6 (mean HCN = 5.44±1.38) had mainly of clones from Embrapa. Clusters 3 (mean HCN = 5.51±0.89) and 9 (mean HCN = 5.26±1.37) had clones exclusively from IITA and Embrapa, respectively. The above-mentioned clusters were all in the same Tocher's group based on HCN mean (Table 3).

Although the highest F_{ST} were expected to be among groups of clones from Embrapa and IITA, there were groups with clones exclusively from Embrapa (clusters 5 and 9) and exclusively from IITA (clusters 3 and 8) that presented moderate F_{ST} between them of 0.090, evidencing the genetic variability within the germplasms of each origin. The same way, there were clusters from different origins that presented low F_{ST}. The heatmap of the kinship matrix **G** also illustrated the population structure (Fig. 2), highlighting the high coefficients between clusters 8 and 9 (F_{ST} = 0.004; Table 4), composed of clones from IITA and Embrapa, respectively.

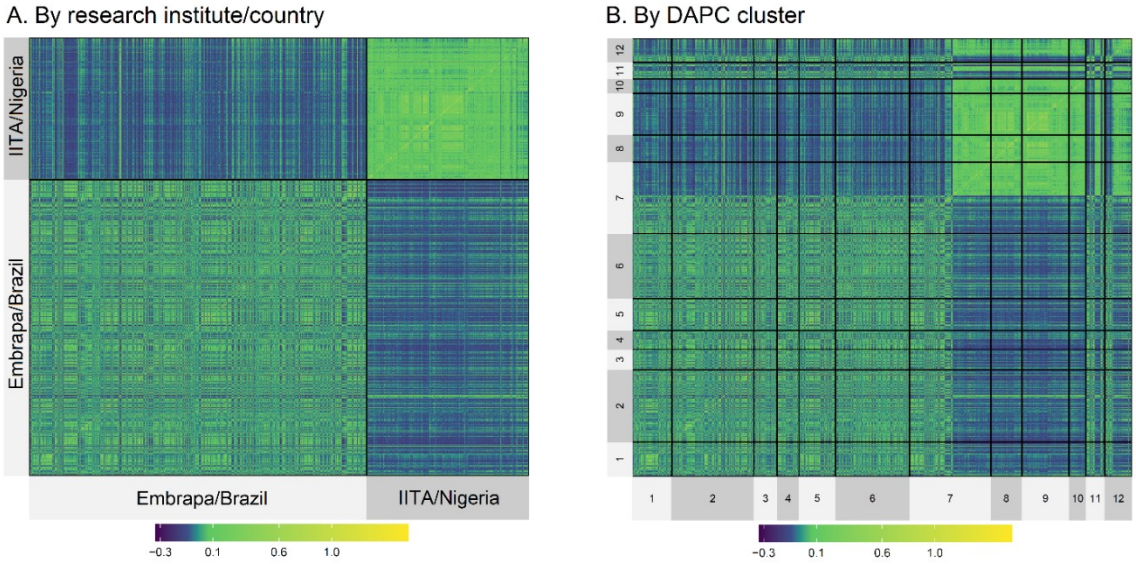


Fig 2. Heatmap of the kinship G matrix by Institute/Country (A) and (B) by DAPC Cluster (1-12).

Table 4. Fixation index (F_{ST}) estimated by single nucleotide polymorphisms (SNPs) between cassava clones from Embrapa and IITA and between clusters identified within the joint dataset (Embrapa+IITA).

F_{ST} between genotypic data from Embrapa and IITA = 0.072											
Clusters	**Clusters**										
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	0.004	0.013	0.007	0.005	0.009	0.018	0.084	0.084	0.074	0.023	0.030
2	0.000	0.004	0.002	0.009	0.003	0.019	0.086	0.086	0.073	0.019	0.026
3	0.004	0.000	0.003	0.020	0.005	0.023	0.090	0.091	0.076	0.020	0.024
4	0.002	0.003	0.000	0.011	0.002	0.018	0.087	0.086	0.072	0.020	0.023
5	0.009	0.020	0.011	0.000	0.013	0.023	0.090	0.090	0.082	0.030	0.037
6	0.003	0.005	0.002	0.013	0.000	0.018	0.084	0.083	0.070	0.023	0.024
7	0.019	0.023	0.018	0.023	0.018	0.000	0.032	0.030	0.023	0.007	0.009
8	0.086	0.090	0.087	0.090	0.084	0.032	0.000	0.004	0.023	0.043	0.047
9	0.086	0.091	0.086	0.090	0.083	0.030	0.004	0.000	0.020	0.040	0.043
10	0.073	0.076	0.072	0.082	0.070	0.023	0.023	0.020	0.000	0.027	0.027
11	0.019	0.020	0.020	0.030	0.023	0.007	0.043	0.040	0.027	0.000	0.008
12	0.026	0.024	0.023	0.037	0.024	0.009	0.047	0.043	0.027	0.008	0.000

4.3.2. Genomic analyses for Embrapa, IITA and joint datasets

All datasets had significant ($p < 0.01$) genetic and year-replication variability for HCN (Table 5). The phenotype-based heritabilities were 0.76, 0.18 and 0.56 for Embrapa, IITA and the joint Embrapa-IITA datasets, respectively; while the SNP-based heritabilities were 0.97, 0.20 and 0.65 for Embrapa, IITA and the joint Embrapa-IITA datasets, respectively. The Embrapa's dataset presented highest mean and lowest coefficient of residual variation for HCN when compared to IITA and joint datasets.

Table 5. Estimates of phenotype-based heritability (H^2), SNP-based heritability or haplotype-based heritability (h^2), predicted mean, coefficient of residual variation (CVe), genetic variance (σ_G^2), variance due the combination of year and replication (σ_{YR}^2), residual variance (σ_ε^2), predictive ability (PA), accuracy (Ac) and bias for individual markers and haplotypic block genomic analyses for hydrogen cyanide (HCN) content in cassava.

Single markers										
Datasets	**H^2**	**h^2**	**Mean**	**CVe**	**σ_G^2**	**σ_{YR}^2**	**σ_ε^2**	**PA**	**Ac**	**Bias[/2]**
Embrapa	0.76	0.97	7.11	0.13	31.90*[/1]	0.25*	0.79	0.59±0.05	0.67±0.06	0.41±0.06
IITA	0.18	0.20	4.95	0.28	0.64*	0.62*	1.92	0.26±0.12	0.61±0.29	0.39±0.37
Embrapa+IITA	0.56	0.65	5.57	0.21	3.68*	0.59*	1.39	0.64±0.05	0.85±0.07	0.23±0.06
Haplotype blocks										
Datasets	H^2	h^2	Mean	CVe	σ_G^2	σ_{YR}^2	σ_ε^2	PA	Ac	Bias
Embrapa	0.76	0.96	5.91	0.16	31.00*	0.27*	0.86	0.48±0.07	0.56±0.08	0.58±0.08
IITA	0.18	0.20	5.05	0.27	0.65*	0.62*	1.92	0.22±0.14	0.51±0.34	0.48±0.38
Embrapa+IITA	0.56	0.62	5.74	0.21	3.31*	0.59*	1.42	0.60±0.04	0.80±0.05	0.30±0.05

[/1] *Significant at 0.01 by the likelihood ratio test (LRT). LRT = 2*(complete model log-likelihood - reduced model log-likelihood); [/2] Bias = 1 - $\hat{b}_{PP,GEBV}$.

For the single markers' analyses, the highest predictive ability, accuracy and the lowest bias were achieved when using the joint dataset, being 0.64±0.05, 0.85±0.07 and 0.23±0.06, respectively. For Embrapa and IITA datasets alone, the SNP-based predictive abilities were 0.59±0.05 and 0.26±0.12, the accuracies were 0.67±0.06 and 0.61±0.29 and the biases were 0.41±0.06 and 0.39±0.37, respectively. For all genetic parameters, the highest standard deviations of 10-folds were found when using the IITA dataset.

From the practical point of view, for HCN there is still a reclassification, group-allocating according to the picrate method (1-9)'s records, which are: ranging from 1.0 to 4.0 - Sweet; from 4.1 to 5.0 - Intermedium; from 5.1 to 9.0 - Bitter. Fig. 3 illustrates boxplots contrasting the HCN estimated GEBVs (predicted by single markers) on own country-dataset, cross-country-dataset and considering both datasets jointly to characterize sweet, intermedium and bitter classes for Embrapa and IITA germplasm. For the GEBVs of clones

from IITA predicted with Embrapa dataset, the predicted values exceeded the sample space limits of the picrate method (1-9). It seems that cross-country genomic predictions, in the present work, underestimated the GEBVs for Embrapa clones predicted by markers effects from IITA's dataset GP and overestimated the GEBVs for IITA clones predicted by markers effects from Embrapa's dataset GP.

4.3.3. Genomic analyses with single markers vs. haplotypes

Haplotype blocks with strong LD between markers were identified by the Gabriel's method (Gabriel et al. 2002). The maximum length of the blocks was set to 200 Kb. For Embrapa, IITA and the joint datasets, 3,132, 2,678 and 3,337 haplotype blocks were found with mean lengths (Kb) of 20.400, 33.189 and 19.636 Kb and mean number of SNPs of 3.616, 4.086 and 3.569 SNPs, respectively (Supplementary Table 1).

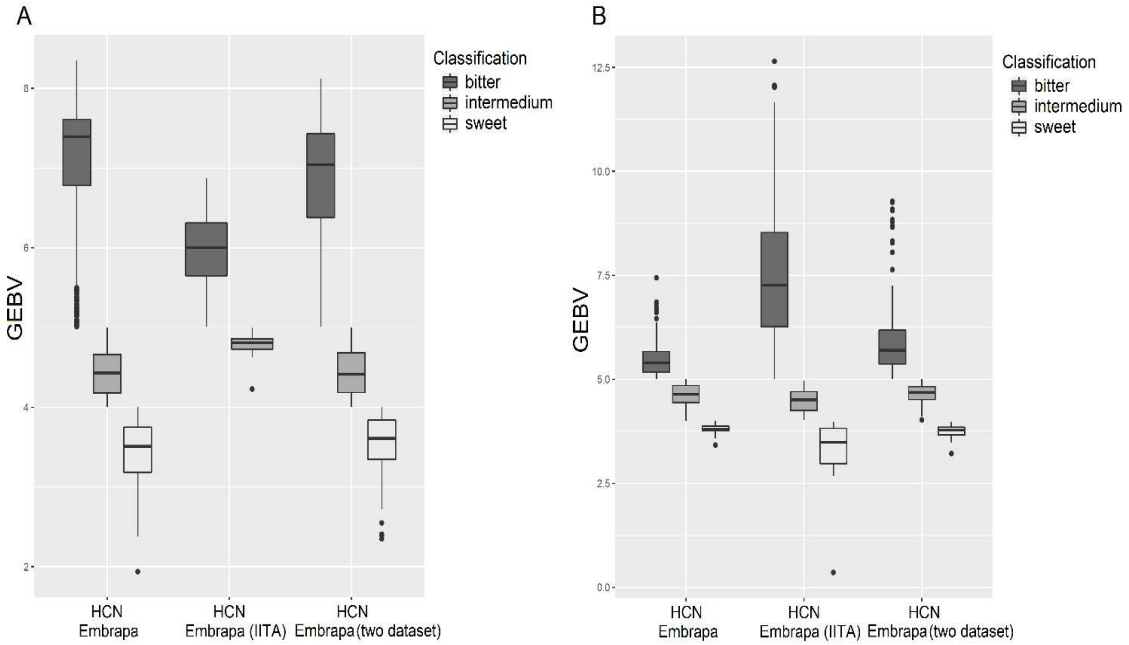


Fig 3. Boxplots contrasting the hydrogen cyanide predictions on country-dataset and cross-country-dataset to show sweet, intermedium and bitter classes. (A) Hydrogen cyanide for Embrapa's clones based on GEBV prediction on Embrapa's dataset, IITA's dataset and the two datasets together, respectively. (B) Hydrogen cyanide for IITA's clones based on GEBV prediction on IITA's dataset, Embrapa's datasets and the two datasets together, respectively.

Although we have assumed a simplified approach to number the 'haplotype matrix', for the haplotype genomic prediction analyses (Table 5), the accuracy recovering rates with haplotype blocks analyses compared to single markers were high: 84% (0.56/0.67), 84%

(0.51/0.61) and 94% (0.80/0.85), for the three datasets. Predictions from haplotype blocks had lower accuracies and higher biases compared to single markers; however, the correlation between the GEBVs estimated by the analysis of single markers and haplotype blocks were 0.99, 0.95 and 0.99 for Embrapa, IITA and the joint dataset, respectively. Model complexity reductions (haplotypes number/single markers number) from using haplotypes were down to 22% (3,132/14,323), 20% (2,678/13,524) and 22% (3,337/14,924) of the original complexity, for Embrapa, IITA and the joint dataset, respectively.

4.3.4. Comparisons between GEBVs rankings

For Embrapa clones' GEBVs estimated by markers effects in the three analyses (analysis 1: Embrapa phenotypic and genotypic data, analysis 2: IITA phenotypic and genotypic data and analysis 3: the two datasets together), high correspondences between GEBVs estimated by analyses 1 and 2, 1 and 3, and 2 and 3 (correlations of 0.55, 0.96 and 0.60, respectively; whereas for Kendall's correlations they were 0.38, 0.82 and 0.43, respectively) were identified. For IITA clones' GEBVs estimated by markers effects in the three analyses, weak but positive correspondences were obtained between GEBVs estimated by analyses 1 and 2, and 1 and 3 (correlations of 0.10 and 0.15, respectively; whereas for Kendall's correlations they were 0.07 and 0.10, respectively). For analyses 2 and 3 a good correspondence was observed (correlations of 0.92; and 0.78 for Kendall's correlation).

We accessed the correlation between the common SNP's effects vectors (total of 11,883 SNPs) originated from the three analyses. For SNP effects from analyses 1 and 2 the correlation was 0.01, from analyses 1 and 3, and 2 and 3, they were 0.69 and 0.42, respectively.

We also compared the top (lower GEBVs for HCN) 100 clones ranking contrasting the predicted and estimated GEBVs for both Embrapa and IITA datasets (analyses 1 and 2, respectively), considering the hypothetical situation that in the absence of phenotypic characterization for one research institute/country, we might want to use the marker effects obtained from the GP using data from the other research institute/country to predict the GEBVs' performance of their clones. For Embrapa, from the top 100-lists' comparison, we have found 12 coincident clones (Supplementary Table 2), while for IITA's lists we have found 24 coincident clones (Supplementary Table 3). Out of the coincident clones, those from Embrapa are in DAPC clusters 1 and 10 and those from IITA are in DAPC clusters 1, 2, 4, 6 and 8. From these clusters, 1, 2 and 6 are mixed, while 4, 8 and 10 are composed

exclusively by IITA or Embrapa clones. In Supplementary Fig. 2 is illustrated where those coincident clones between own- and cross-country genomic predictions fall in the PCA plot.

4.4. DISCUSSION

4.4.1. Population structure

This study was conceptualized as a proof-of-concept to assess the feasibility of using cross-country genomic predictions to support the selection of most suitable clones for germplasm exchange. Moreover, the germplasm characterization for preemptive breeding purposes, in case of quarantine diseases, i.e. when we might not have the phenotypes (resistance to certain disease because the screening in greenhouse or in the field is forbidden) is another important field that can benefit from the results of the present study. In such case, the estimation of the clones' GEBVs by using markers' effects estimated in different training population can be used to rank the most promising clones with resistance to the disease. However, it is well known the population structure has a strong effect on genomic prediction estimates, as observed in both animals (HAYES et al., 2009a) and plants (TECHNOW et al., 2013).

We performed PCA and DAPC for the joint dataset to understand the genomic structure among the Embrapa-Brazil and IITA-Nigeria cassava germplasms. Understanding how the cassava clones from these two breeding programs are related to each other is important in order to better plan the germplasm exchange dynamic between Brazil and Nigeria, or even to understand the outcomes from the joint datasets and the mutual cross-country predictions. By plotting the three first PCs, two big groups were clearly arrayed to separate clones by origin (Fig 1A), and then, by color coding according to the DAPCs clusters, not surprisingly the groups were composed by almost exclusively clones from one or the other research institute/country's origin (Fig 1B; Table 3). Regarding genetic information captured by SNPs, group differences might be due to group specific SNP allele effects (RIO et al., 2020).

The fixation index (F_{ST}) between the proposed mutual training and validation sets, composed by Embrapa and IITA's germplasms, were moderate (0.072), representing a weak population differentiation. When studying population structure of Brazilian germplasm from Embrapa separately, Ogbonna et al. (2020b) found almost the same value, an overall pairwise F_{ST} of 0.073 in 10 clusters obtained by DACP. The moderate F_{ST} is consistent with

the fact that the joint analysis showed greater accuracy and lower bias compared to the analyses of each dataset alone; if the divergence between datasets was greater, a sharp drop in accuracy would be expected. Scutari et al. (2016) found that the correlation between true and predicted values decays approximately linearly with respect to *fixation index* between the training and the target populations.

To provide comprehensive genetic architecture of HCN in cassava, Ogbonna et al. (2020a) performed a GWAS with Brazilian dataset and validated in an African population, and joint GWAS analysis between Africa and Brazil. These authors also showed evidences that the genetic architecture of HCN is conserved. This study revealed HCN is regulated in an oligogenic manner with two major loci explaining the variation across their datasets. Wolfe et al. (2019) identified regions in the genome of cultivated cassava (*Manihot esculenta*) in Africa that had introgressions of *Manihot glaziovii*, the legacy of crosses during the 1930s to improve varieties and mitigate the effects of emerging plant diseases. Although crosses with *Manihot glaziovii* were not frequent in Brazil as they were in Africa at that time, recently Ogbonna et al. (2020b) found introgressions of *M. glaziovii* in the Brazilian germplasm as well (on chromosomes 5 and 17, while those reported in African germplasm were on chromosomes 1 and 4).

4.4.2. Genomic analyses for Embrapa, IITA and joint datasets

Comparing the dataset's genetic parameters estimates, Embrapa's dataset yielded the largest genetic variance estimates for HCN compared to IITA, whereas the latter had the largest estimates for year-replication and residual variance estimates. Highest mean for HCN were observed for Embrapa dataset. Ogbonna et al. (2020a) explored the distribution of cyanide across sub-Saharan Africa datasets, by leveraging open-source data; their analysis indicated that Central and Southern Africa showed on average higher cyanide varieties compared to West Africa, and they also revealed a very slight trend of lowering cyanide on landraces compared to improved varieties.

Phenotype-based heritability ranged from 0.18 (IITA) and 0.76 (Embrapa) and SNP- and haplotype-based heritabilities ranged from 0.20 (IITA) and 0.97 (Embrapa) (Table 5). The observed differences between heritability estimates, could be attributed to several factors, e.g. environmental influence, genetically related plants, phenotypic measurement errors, imputation details, and could have been potentialized by germplasm's mislabeling.

Clone mislabeling has been reported to reduce the genetic gain by up to 40% in Africa (YABE et al., 2018).

In our study, pooling the two datasets together improved HCN predictive ability, accuracy as well as reduced the bias. Multi-population TP combining breeding populations can be employed to assess the potential for broadly focused GS training approaches (FAVILLE et al., 2018). Multi-group training sets has already shown good predictive ability for animals (PRYCE et al., 2011) and plants (TECHNOW et al., 2013). In pigs, Song et al. (2019) reported the GP using combined populations with different genetic backgrounds or from different breeds have not shown a clear advantage over using within-population or within-breed GPs. Wolfe et al. (2017) reported that cross-population accuracy was generally low, with mean of 0.18, but for prediction of cassava mosaic disease it has increased up to 0.57 in one Nigerian population when data from another related population were combined. In Somo et al. (2020) work, by adding the two cassava populations together to increase the TP size did not improve trait prediction accuracy. They evaluated cross-location predictions and compared the trait prediction accuracy estimates with the within-location prediction accuracy estimates. They reported that, from their nine traits evaluated, except for mean cassava green mite severity, the cross-location predictions were generally low, averaging between 0.10 and 0.14 when the Kibaha (Tanzania) population was used to predict the Ukiriguru (Tanzania) set and vice versa, respectively. The authors also included a Ugandan TP in either population, but that did not improve trait prediction accuracy neither. In animal breeding, for cross-breed prediction, this limitation has been reported to be due the non-persistent association between SNPs and QTL across breeds (HAYES et al., 2009b). Anyway, it is worthy mention that pooling together datasets are expected to increase precision and accuracy, due to increase in sample size; by its turn, cross-population/location predictions tend to decrease them, due to change in the target, which motivated the narrow down of multilocation datasets to dataset strictly focused on Ibadan in our case. Those referred increase and decrease are in comparison with within population prediction.

Boxplots of estimated GEBVs group-allocated (sweet, intermedium and bitter) on own country-dataset, cross-country-dataset and considering both datasets jointly (Fig. 3) showed that, in the present work, cross-country genomic predictions seemed to had underestimated the GEBVs for Embrapa clones predicted by markers effects from IITA's dataset GP and overestimated the GEBVs for IITA clones predicted by markers effects from Embrapa's dataset GP. This could possibly be due in part to the difference between the mean magnitude for the Embrapa and IITA datasets. Therefore, focusing on the distribution and range of

classes would also be important. The number of clones allocated to each class of HCN content, by research institute and their considered analysis, is presented in Table 6.

Table 6. Number of cassava clones allocated in each class of HCN content, for Embrapa and IITA, according to the GEBVs estimated with marker effects from the three referred analyses.

Institute	Embrapa			IITA		
Classification[/1]	Analysis1[/2]	Analysis2	Analysis3	Analysis1	Analysis2	Analysis3
Sweet	263	0	243	18	36	22
Intermedium	165	79	187	260	72	171
Bitter	802	1151	800	312	482	397
Total clones	1230	1230	1230	590	590	590

[/1] Classification according with the picrate method scale: ranging from 1.0 to 4.0 – sweet; from 4.1 to 5.0 – intermedium; from 5.1 to 9.0 – bitter.

[/2] Analysis 1, 2 and 3: marker effects' prediction based with training datasets of Embrapa, IITA and Embrapa+IITA, respectively.

4.4.3. Genomic analyses with single markers vs. haplotypes

The haplotype GP for HCN was an attempt to better capture the genomic similarity between clones due to the increased LD between haplotypes and causal genetic variants as well as to capture local allelic interactions. Although using haplotypes is often expected to improve the GP's accuracy over single SNPs, in the present study it was not observed. Lower predictive ability and accuracies were found by haplotype-based GP compared to single marker GP. Sallam et al. (2020) found, for *k*-fold cross validation in wheat yield, that using haplotypes of 5, 10, 15 and 20 adjacent markers increased in 6.3, 2.9, 5.3, and 2.2%, respectively in the predictive ability over single markers. The authors pointed a trade-off regarding the haplotype length: on one hand the increase of haplotype length is expected to capture LD between markers in blocks with QTL, thereby increasing the accuracy of prediction, on the other hand this may also increase the number of haplotype 'allelic' classes, which may reduce the accuracy of prediction due to smaller sample sizes representing these classes. Our simplified approach proposed to build the haplotype's matrix, could possibly explain why we observed slightly lower accuracies: our haplotypes were generated by Gabriel's method (GABRIEL et al., 2002) implemented on PLINK, and length set to maximum 200 Kb. Thus, the mean number of SNPs were of 3-4 SNPs spanning segment-length averaging of 20-33 Kb. However, the predicted GEBVs between the two approaches (haplotype vs. single markers) were highly correlated (> 0.95).

A study performed by Ogbonna et al. (2020a) revealed that HCN is regulated in an oligogenic manner with a few major genomic regions explaining a great proportion of the variation. Our efforts on using haplotype blocks is aligned with the previous findings; the search for a simplified structure that can explain the behavior of the trait and predictability of clone performance across countries. The accuracy recovering rates with haplotype blocks analyses compared to single markers we have got from using only around 3,000 haplotypes is very interesting and appealing. These haplotypes are likely to encompass polygenic blocks with coadapted genes controlling important quantitative traits. This capturing of the important causal genetic variants and of the LD between markers and QTL in blocks enables increasing the accuracy. These genes in blocks are relevant for both local variation (and adaptation) and for stability (predictability) when they are genetically conserved across locations.

4.4.4. Comparisons between GEBVs rankings

One location's data can be used to predict performance in another location, which can be helpful in accelerating the breeding process. However, according to Rio et al. (2020), when targeting a group-specific predicted set, training a model on a different group can eventually decrease accuracy as shown in several species. We compared the top (lower GEBVs for HCN) 100 ranking lists contrasting the predicted and estimated GEBVs for both Embrapa and IITA datasets (analysis 1 – TP with Embrapa's dataset – and analysis 2 – TP with IITA's dataset –, respectively), considering the hypothetical situation that in the absence of phenotypic characterization for one research institute/country, we might want to use the marker effects obtained from the GP using TP data from the other research institute/country to predict the performance of their clones.

For Embrapa clones, a correlation between GEBVs predicted with TP from Embrapa and IITA was 0.55, whereas for IITA clones, it was 0.10. For Embrapa, from the 100-lists comparison, we have found 12 coincident clones, while for IITA lists we have found 24 coincident clones, which lead us think that although the observed correspondence between GEBVs vectors for IITA using marker effects from analyses 1 and 2 was weak (correlation of 0.10 between the vectors of estimated GEBVs), they could be enough to at least giving some clue in clonal selection for germplasm exchange with much better probabilities than a blind-scenario, where it is not feasible to send all the clones you have, and you have plenty of good clones related to important agronomic traits.

It seems reasonable to attribute part of the success of these coincidences between the predictions of GEBVs obtained with the effects of markers from the own-country's analysis vs. from other country's analysis to the similarities of the regions (Cruz das Almas-Brazil and Ibadan-Nigeria) in terms of temperature and rainfall distribution. Both are in a low latitude range, belonging to tropical latitude zones; according to Köppen-Geiger Climate *Classification*, Cruz das Almas and Ibadan are lying into tropical rainforest climate (Af) and tropical savanna, wet (Aw), respectively. Other interesting variables to be mentioned: Cruz das Almas' altitude of 224 m above sea level, 1,136 mm annual precipitation and average temperature of 23ºC (within a range of 18-34ºC), whereas Ibadan's altitude of 199 m above sea level, 1,311 mm annual precipitation and average temperature of 26.5ºC (within a range of 20-34ºC) (en.climate-data.org). Similar environments are likely to display low genotype x environment interaction and then, higher coincidences between the predictions of GEBVs obtained with the effects of markers from the own-country's analysis vs. from other country's analysis are obtained.

The poor prediction result for cross-program prediction reported here, using Embrapa's TP to predict IITA clones' GEBV (0.10) makes a bit hard to conceive the cross-program predictions to be adopted in the daily routine of the referred breeding programs. However, in the lack of a better strategy, especially for the case of exchanging material in the context of quarantine diseases, which is a more delicate scenario with an existing restriction on clone movement, information at the molecular level could aid to the understanding of populations' structure, diversity and cross-country genomic predictions could corroborate to guide a germplasm selection to exchange by a reasonably accurate selection of clones since an unlimited shipping would be unfeasible.

Nevertheless, when using IITA's TP to predict Embrapa clones' GEBV we got a correlation of 0.55 which seems to be good enough for such kind of study. This discrepancy within the mutual cross-country GP may perhaps have been influenced by experimental technical details, since the same trait showed so disparate heritabilities and coefficients of variation when evaluated in different research institutes/countries. From the GWAS results for HCN obtained by Ogbonna et al. (2020a) on IITA's historical data, the error variances were higher compared to Embrapa's as well. IITA's dataset spans over more than a decade of experimental work which could certainly was influenced by changes in human resources, management, not to mention year variance which is larger. Thus, the potential impacts on prediction accuracy can change the ranking of top performing clones in the validation population. Although it seems to have worked reasonably well in the present scenario, to

make a conclusive statement regarding the use of cross-country predictions in cassava, further evidences are needed. This seems to be among the first attempts to evaluate the cross-country genomic selection in cassava and in plants. As such, a lot of useful new information was provided on the subject, which can guide new research on this very important and emerging field.

4.5. CONCLUSIONS

Twelve clusters were revealed by the assessment of population structure in South-American and African cassava. Fixation index among the clusters identified within the joint dataset ranged from 0.002 to 0.091. The joint dataset (Embrapa+IITA) provided an improved accuracy compared to the prediction accuracy of either germplasm' sources individually. When using the marker effects estimated with IITA dataset as the training to predict the Embrapa clones' GEBVs, we could say it was a successful case of cross-country prediction, taking into consideration the medium-to-high correlation of 0.55 between their vectors of estimated GEBVs; regardless, the reverse presented a low correlation of 0.10. Genomic predictions from haplotype blocks had slightly lower accuracies and higher biases compared to single markers. The correlation between the GEBVs estimated by the analyses of single markers and by haplotype blocks varied from 0.95 to 0.99, with high accuracy recovering rates of the haplotype blocks compared to single markers, ranging from 0.84 to 0.94. Also, with the haplotype blocks analyses the model sizes were reduced, with complexity reduction rates going down to 0.20-0.22. Cross-country genomic predictions proved to have potential use under the present study's scenario, i.e., investigated trait heritabilities higher than 0.18 and conserved genetic architecture across locations, with clones from regions that are not too much divergent in terms of meteorological conditions and with at least a minimal historic of germplasm exchange.

4.6. ACKNOWLEDGMENTS

The authors thank CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*), FAPEMIG (*Fundação de Amparo à Pesquisa do Estado de Minas Gerais*), FAPESB (*Fundação de Amparo à Pesquisa do Estado da Bahia*), and CAPES (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) for financial support. This work was also partially supported by the NEXTGEN Cassava project, through a grant to

Cornell University by the UK's Foreign, Commonwealth & Development Office (FCDO) and the Bill & Melinda Gates Foundation (Grant INV-007637 http://www.gatesfoundation.org). LGT was supported by scholarships from CNPq and CAPES.

4.7. REFERENCES

ALBUQUERQUE, H.; CARMO, C.; BRITO, A.; OLIVEIRA, E. Genetic diversity of *Manihot esculenta* Crantz germplasm based on single-nucleotide polymorphism markers. **Annals of Applied Biology**, 2018. https://doi.org/10.1111/aab.12460.

ANDRADE, L.; SOUSA, M.; OLIVEIRA, E.; RESENDE, M.; AZEVEDO, C. Cassava yield traits predicted by genomic selection methods. **PLoS ONE**, v. 14, n. 11, p. e0224920, 2019, https://doi.org/10.1371/journal.pone.0224920

BART, R.; TAYLOR, N. New opportunities and challenges to engineer disease resistance in cassava, a staple food of African small-holder farmers. PLoS Pathog, v. 13, n. 5, p. e1006287, 2017. https://doi.org/10.1371/journal.ppat.1006287

BRADBURY, M. G.; EGAN, S. V.; BRADBURY, J. H. Picrate paper gits fork determination of total cyanogens in Cassava roots and all forms of cyanogens in Cassava products. **Journal of the Science of Food and Agriculture**, v. 79, n. 4, p. 593-601, 1999.

BREDESON, J.; LYONS, J. B.; PROCHNIK, S. E.; WU, G. A.; HA, C. M. et al. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. **Nature Biotechnology**, v. 34, 562-570, 2016.

BRITO, A. C.; OLIVEIRA, S.; OLIVEIRA, E. J. Genome-wide association study for resistance to cassava root rot. **The Journal of Agricultural Science**, v. 155, n. 9, p. 1424-1441, 2017.

BROWNING, B.; BROWNING, S. A Unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. **American Journal of Human Genetics**, v. 84, n. 2, p. 210-23, 2009.

CEBALLOS, H.; ROJANARIDPICHED, C.; PHUMICHAI, C.; BECERRA, L.; KITTIPADAKUL, P.; IGLESIAS, C.; GRACEN, V. Excellence in Cassava breeding: perspectives for the future. **Crop Breed Genet. Genom.**, v. 2, n. 2, p. e200008, 2020. https://doi.org/10.20900/cbgg20200008.

CEREDA, M.; MATTOS, M. Linamarin: the toxic compound of cassava. **Journal of Venomous Animals and Toxins**, v. 2, p. 06-12, 1996.

Code of Practice for the Reduction of Hydrocyanic Acid (HCN) in Cassava and Cassava Products (CAC/RCP 73-2013).

COVARRUBIAS-PAZARAN, G. Genome assisted prediction of quantitative traits using the R package sommer. **PLoS ONE**, v. 11, n. 6, p. 1-15, 2016.

DE HAAS, Y.; CALUS, M.; VEERKAMP, R.; WALL, E.; COFFEY, M.; DAETWYLER, H.; HAYES, B.; PRYCE, J. Improved accuracy of genomic prediction for dry matter intake of dairy cattle from combined European and Australian data sets. **J. Dairy Sci.**, v. 95, p. 6103-6112, 2012.

DOYLE, J.; DOYLE, J. Isolation of plant DNA from fresh tissue. **Focus**, v. 12, p. 13-15, 1990.

ELIAS, A.; RABBI, I.; KULAKOW, P.; JANNICK, J. Improving genomic prediction in cassava field experiments using spatial analysis. **G3: Genes, Genomes, Genetics**, v. 8, n. 1, p. 53-62, 2018. https://doi.org/10.1534/g3.117.300323,

ELSHIRE, R.; GLAUBITZ, J.; SUN, Q.; POLAND, J.; KAWAMOTO, K.; BUCKLER, E.; MITCHELL, S. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS ONE**, v. 6, p. 1-10, 2011.

ESUMA, W.; HERSELMAN, L.; LABUSCHAGNE, M.; RAMU, P.; LU, F.; BAGUMA, Y. et al. Genome-wide association mapping of provitamin A carotenoid content in cassava. **Euphytica**, v. 212, p. 97-110, 2016.

FAVILLE, M.; GANESH, S.; CAO, M.; JAHUNER, M.; BILTON, T.; EASTON, H. et al. Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. Theor. Appl. Genet., v. 131, n. 3, p. 703-720, 2018.

FUKUDA, W. M. G.; GUEVARA, C. R.; KAWUKI, R.; FERGUSON, M. E. Selected morphological and agronomic descriptors for the characterization of cassava. **International Institute of Tropical Agriculture (IITA)**, Ibadan, Nigeria. 2010. 19 p.

GABRIEL, S.; SCHAFFNER, S.; NGUYEN, H.; MOORE, J.; ROY, J.; BLUMENSTIEL, et al. The structure of haplotype blocks in the human genome. **Science**, v. 296, p. 2225-2229, 2002.

GLAUBITZ, J.; CASSTEVENS, T.; LU, F.; HARRIMAN, J.; ELSHIRE, R.; SUN, Q. et al. TASSEL-GBS: A high-capacity genotyping by sequencing analysis pipeline. PloS One 9(2):e90346, 2014.

GOUDET, J.; JOMBART, T. **Hierfstat**: Estimation and tests of hierarchical F-statistics. R. Package version 004-22 10, 2015.

HAYES, B.; BOWMAN, P.; CHAMBERLAIN, A.; VERBYLA K.; GODDARD, M. Accuracy of genomic breeding values in multi-breed dairy cattle populations. **Genet. Sel. Evol.**, v. 41, p. 1-9, 2009a. doi: 10.1186/1297-9686-41-51.

HAYES, B.; BOWMAN, P.; CHAMBERLAIN, A.; GODDARD, M. Invited review: Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science**, v. 92, n. 2, p, 433-443, 2009b. https://doi.org/10.3168/jds.2008-1646.

HILLOCKS, R.; THRESH, J.; BELLOTTI, A. (Ed.). **Cassava:** Biology, production and utilization. Wallingford, CT: CABI, 2002.

IKEOGU, U.; AKDEMIR, D.; WOLFE, M.; OKEKE, U.; CHINEDOZI, A.; JANNICK J. et al. Genetic correlation, genome-wide association and genomic prediction of portable NIRS predicted carotenoids in cassava roots. **Front. Plant Sci.**, v. 10, p. 1570, 2019. https://doi: 10.3389/fpls.2019.01570

INTERNATIONAL CASSAVA GENETIC MAP CONSORTIUM (ICGMC). High resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* Crantz) from ten populations. **G3: Genes, Genomes, Genet.**, 2014. https://doi:10.1534/g3.114.015008.

JARVIS, A.; RAMIREZ-VILLEGAS, J.; HERRERA CAMPO, B.; NAVARRO-RACINES, C. Is cassava the answer to African climate change adaptation? **Trop. Plant Biol.**, v. 5, p. 9-29, 2012.

JOMBART, T. Adegenet: A R package for the multivariate analysis of genetic markers. **Bioinformatics**, v. 24, n. 11, p. 1403-1405, 2008. https://doi.org/10.1093/bioinformatics/btn129.

JOMBART, T.; DEVILLARD, S.; BALLOUX, F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. **BMC Genetics**, v. 11, p. 94, 2010.

KAYONDO, S.; CARPIO, D.; LOZANO, R.; OZIMATI, A.; WOLFE, M.; BAGUMA, Y. et al. Genome-wide association mapping and genomic prediction unravels CBSD resistance in a *Manihot esculenta* breeding population. **Scientific Reports**, v. 8, p. 1549, 2018.

McKENNA, A.; HANNA, M.; BANKS, E.; SIVACHENKO, A.; CIBULSKIS, K.; KERNYTSKY, A. et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. **Genome Research**, v. 20, n. 9, p. 1297-1303, 2010.

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v. 157, p. 1819, 2001.

OBI, C.; OKEZIE, O.; UKAEGBU, T. Fermentation reduces cyanide content during the production of cassava flours from sweet and bitter cassava tuber varieties. **Asian Food Science Journal**, v. 11, n. 1, p. 1-10, 2019.

OBUEH, H.; KOLAWOLE, S. Comparative study on the nutritional and anti-nutritional compositions of sweet and bitter cassava varieties for garri production. **J. Nutr. Health Sci.**, v. 3, n. 3, p. 302, 2016.

OGBONNA, A.; ANDRADE, L.; RABBI, I.; MUELLER, L.; OLIVEIRA, E.; BAUCHET, G. Genetic architecture and gene mapping of cyanide in cassava *(Manihot esculenta Crantz.*). 2020a. https://www.biorxiv.org/content/10.1101/2020.06.19.159160v1 (Posted June 20, 2020a).

OGBONNA, A.; ANDRADE, L.; RABBI, I.; MUELLER, L.; OLIVEIRA, E.; BAUCHET, G. Comprehensive genotyping of Brazilian cassava (*Manihot esculenta* Crantz) germplasm bank: Insights into diversification and domestication. https://www.biorxiv.org/content/10.1101/2020.07.13.200816v1 (2020b) (Posted July 14, 2020b)

OJO, R.; OBI, N.; AKINTAYO, C.; ADEBAYO-GEGE, G. Evaluation of cyanogen contents of cassava and cassava based food products in Karu, Nasarawa State, North-Central Nigeria. **IOSR J. Environ Sci. Toxicol. Food Technol.**, v. 6, n. 1, p. 47-50, 2013.

OKEKE, U.; AKDEMIR, D.; RABBI, I.; KULAKOW, P.; JANNICK, J. Accuracies of univariate and multivariate genomic prediction models in African cassava. **Genetics Selection Evolution**, v. 49-88, 2017.

OLIVEIRA, E.; RESENDE, M.; SANTOS, V.; FERREIRA, C.; OLIVEIRA, G.; SILVA, M. et al. Genome-wide selection in cassava. **Euphytica**, v. 187, p. 263, 2012.

POCRNIC, I.; LOURENÇO, D.; CHEN, C.; HERRING, W.; MISZTAL, I. Crossbred evaluations using single-step genomic BLUP and algorithm for proven and young with different sources of data. **J. Anim. Sci.**, v. 97, n. 4, p, 1513-1522, 2019.

POULTON, J. Cyanogenesis in plants. **Plant Physiol.**, v. 94, p. 401-405, 1990.

PROCHNIK, S.; MARRI, P.; DESANY, B.; RABINOWICZ, P.; KODIRA, C.; MOHIUDDIN, M. et al. The cassava genome: Current progress, future directions. **Tropical Plant Biology**, v. 5, p. 88-94, 2012.

PRYCE, J.; GREDLER, B.; BOLORMAA, S.; BOWMAN, J.; EGGER-DANNER, C.; FUERST, C. et al. Short communication: Genomic selection using a multi-breed, across-country reference population. **Journal of Dairy Science**, v. 94, n. 5, p. 2625-2630, 2011. https://doi.org/10.3168/jds.2010-3719, 2011.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A.; BENDER, D. et al. Plink: A tool set for whole-genome association and population-based linkage analyses. **The American Journal of Human Genetics**, v. 81, n. 3, p. 559-575, 2007.

R CORE TEAM. **R:** A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2019. Available in: http://www.R-project.org.

RABBI, I.; UDOH, L.; WOLFE, M.; PARKES, E.; GEDIL, M.; DIXON, A. et al. Genome-wide association mapping of correlated traits in cassava: Dry matter and total carotenoid content. **Plant Genome**, v. 10, n. 3, 2017.

RIO, S.; MOREAU, L.; CHARCOSSET, A.; MARY-HUARD, T. Accounting for group-specific allele effects and admixture in genomic predictions: Theory and experimental evaluation in maize. **Genetics**, v. 216, p. 27-41, 2020.

SALLAM, A.; CONLEY, E.; PRAKAPENKA, D.; DA, Y.; ANDERSON, J. Improving Prediction accuracy using multi-allelic haplotype prediction and training population optimization in wheat. **G3: Genes, Genomes, Genet.**, v. 10, n. 7, p. 2265-2273, 2020.

SCUTARI, M.; MACKAY, I.; BALDING, D. Using genetic distance to infer the accuracy of genomic prediction. **PLoS Genet.**, v. 12, p. e1006288, 2016.

SOMO, M.; KULEMBEKA, H.; MTUNDA, K.; MREMA, E.; SALUM, K.; WOLFE, M. Genomic prediction and quantitative trait locus discovery in a cassava training population

constructed from multiple breeding stages. **Crop Science**, v. 60, p. 896-913, 2020. doi:10. 1002/csc2.20003

SONG, H.; YE, S.; JIANG, Y.; ZHANG, Z.; ZHANG, Q.; DING, X. Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. **Genetics Selection Evolution**, v. 51, 2019.

TECHNOW, F.; BÜRGER, A.; MELCHINGER, A. Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. **G3 Genes, Genomes, Genet.**, v. 3, p. 197-203, 2013. doi:10.1534/g3.112.004630.

TORRES, L.; RESENDE, M.; AZEVEDO, C.; FONSECA, F.; OLIVEIRA, E. Genomic selection for productive traits in biparental cassava breeding populations. **PLoS ONE**, v. 14, n. 7, p. e0220245, 2019.

VANRADEN, P.; TOOKER, M.; CHUD, T.; NORMAN, H.; MEGONIGAL JR., J.; HAAGEN I. et al. Genomic predictions for crossbred dairy cattle. **J. Dairy Sci.**, v. 103, p. 1620-1631, 2020.

WHITE, W.; ARIAS-GARZON, D.; McMAHON J.; SAYRE, R. Cyanogenesis in cassava. **Plant Physiol.**, v. 116, p. 1219-1225, 1998.

WHO. **Hydrogen cyanide and cyanide:** Human health aspects. Geneva, 2004. 67 p.

WOLFE, M.; RABBI, I.; EGESI, C.; HAMBLIN, M.; KAWUKI, R.; KULAKOW, P. et al. Genome-wide association and prediction reveals the genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. **Plant Genome**, v. 9, p. 1-13, 2016.

WOLFE, M.; DEL CARPIO, D.; ALABI, O.; EZENWAKA, L.; IKEOGU, U.; KAYONDO, I. et al. Prospects for genomic selection in cassava breeding. **Plant Genome**, v. 10, 2017. https://doi.org/10.3835/plantgenome2017.03.0015.

WOLFE, M.; BAUCHET, G.; CHAN, A.; LOZANO, R.; RAMU, P.; EGESI, C. et al. Historical introgressions from a wild relative of modern cassava improved important traits and may be under balancing selection. **Genetics**, v. 213, n. 4, p. 1237-1253, 2019.

WEIR, B.; COCKERHAM, C. Estimating F-statistics for the analysis of population structure. **Evolution**, v. 38, p. 1358-1370, 1984.

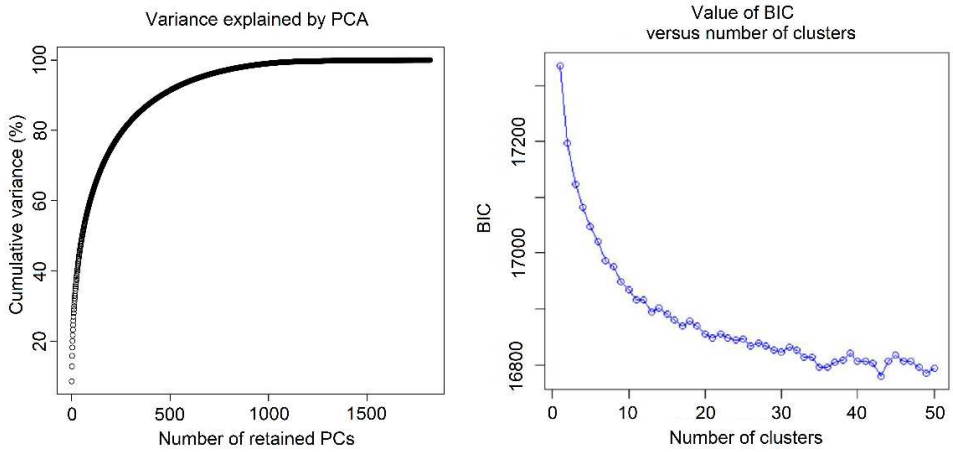
WERNER, C. R.; GAYNOR, R. C.; GORJANC, G.; HICKEY, J. M.; KOX, T.; ABBADI, A.; LECKBAND, G.; SNOWDON, R. J.; STAHL, A. How population structure impacts genomic selection accuracy in cross-validation: Implications for practical breeding. **Frontier in Plant Science**, v. 11, p. 1-14, 2020. https://doi.org/10.3389/fpls.2020. 592977.

WRIGHT, S. The interpretation of population structure by F-statistics with special regard to systems of mating. **Evolution**, v. 19, p. 395-420, 1965.

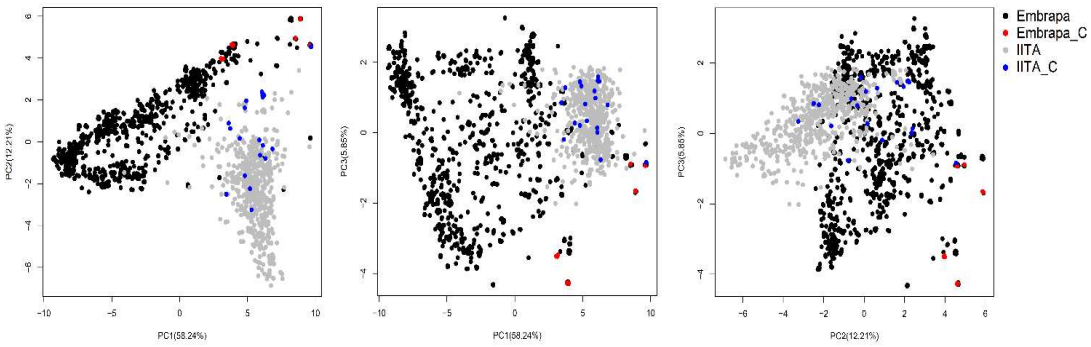
YABE, S.; IWATA, H.; JEAN-LUC, J. Impact of mislabeling on genomic selection in cassava breeding. **Crop Science**, v. 58, p. 1470-1480, 2018. https://doi.org/10.2135/cropsci 2017.07.0442

YONIS, B.; DEL CARPIO, D.; WOLFE, M.; JANNICK, J.; KULAKOW, P.; RABBI, I. Improving root characterisation for genomic prediction in cassava. **Scientific Reports**, v. 10, 2020.

4.8. SUPPLEMENTARY MATERIAL



Supplementary Fig 1. Discriminant analysis of principal components (DAPC); (A) Variance explained by the principal components (extracted from genomic marker matrix); (B) Value of Bayesian Information Criterion (BIC) versus number of clusters.



Supplementary Fig 2. Principal components analysis (PCA) based on the genomic kinship coefficients between clones: PCA derived from the genomic relationship matrix between all cassava clones (N=1,820; 14,924 SNPs) from Brazil (Embrapa) and Nigeria (IITA), showing the first three principal components and the variance explained by each component in parenthesis on the corresponding axis (58.24, 12.21 and 5.85% for PC1, PC2 and PC3, respectively); In black representing Embrapa clone's dispersion and in grey representing IITA's; Highlighting the coincident clones between own- and cross-country genomic predictions, in red representing Embrapa coincident clone's dispersion and in blue representing IITA's.

S1Table. Number, maximum, mean and minimum length of haplotype blocks and number of SNPs within each one. Results from three datasets.

Datasets	# Haplotype Blocks	Length (Kb)			# SNPs		
		Max.[/1]	Mean	Min.	Max.	Mean	Min.
Embrapa	3,132	199.541	20.400	0.002	26	3.616	2
IITA	2,678	199.636	33.189	0.002	22	4.086	2
Embrapa+IITA	3,337	199.541	19.636	0.002	20	3.569	2

[/1] Maximum length set to 200 Kb.

S2Table 2. Coincident twelve clones from the 'top 100 clone's list' for Embrapa germplasm.

Clone	GEBV1_1[/1]		GEBV1_2[/2]	
CNPMF1317.250372086	3.22	(sweet)	5.03	(interm.)
CNPMF1332.250372041	3.10	(sweet)	5.03	(interm.)
CNPMF1414.250372209	2.38	(sweet)	5.00	(interm.)
CNPMF1444.250372212	2.53	(sweet)	4.99	(interm.)
CNPMF1446.250372223	3.17	(sweet)	5.03	(interm.)
CNPMF1507.250372194	2.58	(sweet)	5.03	(interm.)
CNPMF1579.250407031	2.83	(sweet)	5.02	(interm.)
CNPMF187.250370458	3.12	(sweet)	4.88	(interm.)
CNPMF212.250370447	3.18	(sweet)	4.68	(interm.)
CNPMF213.250370448	3.28	(sweet)	4.70	(interm.)
CNPMF286.250370544	2.63	(sweet)	4.74	(interm.)
Rosinha365.250437688	3.14	(sweet)	5.02	(interm.)

/1 GEBV1 = GEBVs estimated for EMBRAPA clones by genomic prediction with EMBRAPA phenotypic and genotypic data; /2 GEBV1_2 = GEBVs estimated for EMBRAPA clones via SNPs' effect estimated by genomic prediction with IITA phenotypic and genotypic data.

S3Table. Coincident twenty-four clones from the 'top 100 clone's list' for IITA germplasm.

Clone	GEBV2_2[/1]		GEBV2_1[/2]	
I000378.250300406	4.03	(interm.)	4.54	(interm.)
I030060.250303093	4.06	(interm.)	4.08	(interm.)
I030061.250300215	4.45	(interm.)	4.83	(interm.)
I030141.250300223	4.04	(interm.)	4.82	(interm.)
I030256.250300226	4.13	(interm.)	4.45	(interm.)
I062116.250304713	4.20	(interm.)	4.51	(interm.)
I101014.250099332	3.87	(sweet)	0.36	(sweet)
I101573.250399860	4.17	(interm.)	4.08	(interm.)
I60142.250304629	4.24	(interm.)	4.08	(interm.)
I920325.250304478	4.00	(sweet)	4.24	(interm.)
I970299.250300471	4.18	(interm.)	4.25	(interm.)
TMEB117.250304093	4.16	(interm.)	3.45	(sweet)
TMEB199.250253760	3.80	(sweet)	3.61	(sweet)
TMEB225.250253783	3.77	(sweet)	4.22	(interm.)
TMEB239.250253789	4.16	(interm.)	4.78	(interm.)
TMEB270.250253800	4.34	(interm.)	4.51	(interm.)
TMEB403.250253858	4.32	(interm.)	3.51	(sweet)
TMEB435.250253875	3.75	(sweet)	4.07	(interm.)
TMEB455.250253885	3.76	(sweet)	4.09	(interm.)
TMEB459.250253886	4.21	(interm.)	2.94	(sweet)
TMEB469.250253889	4.36	(interm.)	3.75	(sweet)
TMEB477.250253892	3.77	(sweet)	4.07	(interm.)
TMEB480.250253894	3.79	(sweet)	4.25	(interm.)
TMEB785.250254011	4.51	(interm.)	3.76	(sweet)

[/1] GEBV2_2 = GEBVs estimated for IITA clones by genomic prediction with IITA phenotypic and genotypic data; [/2] GEBV2_1 = GEBVs estimated for IITA clones via SNPs' effect estimated by genomic prediction with EMBRAPA phenotypic and genotypic data.

5. CHAPTER III: WHAT ARE THE BENEFITS OF INCREASED MARKER COVERAGE AND DENSITY IN GENOME-WIDE ASSOCIATION ANALYSIS? – A CASE STUDY WITH PROTEIN AND OIL CONTENT IN SOYBEAN

Abstract - Improved SNP datasets have been made possible due to advancements in genotyping methods and the continuous decrease in cost per sample. Datasets derived from whole-genome sequencing (WGS) represent the culmination of such a process and offer a more exhaustive coverage of the genome as well as a greater marker density. How and to what extent do such enriched datasets improve the results of association studies compared to commonly used datasets composed of tens of thousands of markers? The objective of this study was to explore these questions by comparing the results of a genome-wide association study (GWAS) for seed oil and protein content performed with either 17K GBS-derived SNPs or 2.18M GBS-/WGS-derived SNPs on the same set of soybean accessions. New regions associated with protein and oil content were found by revisiting previous phenotypic dataset, as well as few overlapping and neighboring regions. In general, peak SNPs identified with denser marker set explained a higher percentage of the variation and the significant regions around them were more narrowly defined. All of these factors have impacted the identification of candidate genes, however, in both scenarios the identification of candidate genes with relevant annotation towards protein and oil content were made possible. Besides providing more likely chances of new marker-trait discovery, using a complete marker coverage seems to be advantageous to gain in resolution, delimiting targeting associated regions, thus, reducing mapping noise. Nevertheless, our results revealed that previous GWAS, even with very limited genotypic dataset, were successful identifying relevant associated regions. Our results provide new insights into applying WGS-imputation in soybean breeding, genetic architecture and candidate genes for protein and oil content in short-season soybean.

Keywords: Soybean. GWAS. Genotyping-by-sequencing. Whole-genome-sequencing. association mapping. Seed protein content. Seed oil content.

5.1. INTRODUCTION

Advances in genotyping technologies, and the decreasing cost per sample, followed by the development of new tools and statistical methods applied to plant genetics are making it possible to obtain vastly improved datasets for genetic analysis. This in-depth characterization of genetic variation generates exceptional data for genotype-phenotype association studies, such as in genome-wide association studies (GWAS) (TORKAMANEH et al., 2018a).

In most analyses conducted to date, SNP datasets have been generated using either GBS or SNP arrays and these will often capture only part of the underlying genetic variation that exists among the accessions used. It is quite typical in such studies to report between 10K and 50K SNPs (SONAH et al., 2014; BANDILLO et al., 2015; LEAMY et al., 2017; COPLEY et al., 2018; LEE et al., 2019; STEKETEE et al., 2020). This incomplete marker coverage can be problematic in at least three ways: 1) some regions of the genome harboring determinants of the trait of interest (a gene or QTL) may not be covered in the sense that no marker is present within some LD blocks; 2) limited marker coverage will make it difficult to precisely define haplotype blocks within which one can search for potential candidate genes; and 3) there is a very low probability of capturing a causal variant. In addition, being able to look back on past analyses and assess how well such partial marker coverage had succeeded in capturing the genetic architecture underlying variation in a trait of interest is also useful. For some researchers, e.g. breeders, a tightly associated marker may be the level of resolution needed for marker-assisted selection (MAS) and it may not be justified to perform the additional work needed to identify and validate a candidate gene and causal variant.

Soybean [*Glycine max* (L.) Merr.] is an important source of protein and oil for animal feed, human consumption, renewable biofuel and also provides industrial raw materials across the world. Increasing seed protein content of soybean cultivars has been one of the main goals of many soybean breeding programs for decades. The growing interest in increasing protein content is particularly driven by the use of the defatted soybean meal as the main protein source in poultry and livestock feed (LEE et al., 2019). However, seed oil content is also a valuable trait for the soybean industry and the strong negative correlation between protein and oil content has hampered progress in the breeding for these traits simultaneously (RINKER et al., 2014; PATIL et al., 2017; LEE et al., 2019). Another

important observed negative correlation is between yield and protein content. Selecting exclusively for higher yields can be detrimental to the total soybean protein content.

In the last decade, the Upper Midwest in the US and the prairie provinces of Canada have experienced a considerable growth in the soybean cropping area. However, protein levels have been found to be lower in these relatively new areas compared to more traditional cropping areas. As protein content is an important seed quality trait and also an important export determinant, impacts of environment, genetics and management are being targeted for studies in the hope of diminishing the impact of this trade-off.

In spite of the fact that protein and oil content can be improved via traditional breeding methods, MAS can be a more efficient means of achieving this, especially in the context of the negative correlation between traits (LI et al., 2019). This is where genome-wide association studies (GWAS) can be useful. Once one has the phenotypes and the SNP data for a group of accessions, GWAS enables the detection of genetic differences accounting for the phenotypic variation (COPLEY et al., 2018), thus identifying markers that are significantly associated with the trait of interest.

The study of the genetic variability among individuals through GWAS can contribute to a better understanding of the genetic control of traits, since it allows the identification of the regions in the genome that are related to the trait of interest as well as the specific allelic variants. Given sufficiently dense genome coverage, functionally relevant polymorphisms are expected to be in linkage disequilibrium (LD) with at least one marker. Numerous GWAS have been successfully conducted in several plant species.

In soybean, GWAS has been used to identify loci associated with agronomic traits (HWANG et al., 2014; SONAH et al., 2014; CONTRERAS-SOTO et al., 2017; LEAMY et al., 2017; COPLEY et al., 2018; LEE et al., 2019), abiotic stress (MAMIDI et al. 2014) and disease resistance (BASTIEN et al., 2014; IQUIRA et al., 2015; CHANG et al., 2016; WEI et al., 2017).

To achieve an extensive coverage that provides high QTL-detection power, Bastien et al. (2014) estimated that at least 68K well-distributed SNPs would be needed to cover the entire soybean genome in a set of Canadian elite accessions. Realistically, in the absence of a way to ensure such an exhaustive and uniform coverage of all LD blocks, it is certain that marker coverage in excess of 100K markers would be required to ensure that coverage is indeed exhaustive. Whole-genome sequencing (WGS) would be a good alternative to increase the amount of data generated, but WGS is still a costly

genotyping technology and genotyping a large number of individuals can prove too costly.

As reviewed in Torkamaneh et al. (2018a), the imputation of missing data allows for a combined approach in which data from an inexpensive genome scan (e.g. GBS or a SNP array) can be combined with extensive WGS data. In a first step, a GWAS panel can be genotyped at tens of thousands of SNPs (called “target” population) and, in a second step, WGS on a subset of these lines can be used as a relevant reference panel (called “reference” population) for the imputation of millions of SNPs on the entire set of lines. Up to date, the vast majority of case studies of possible uses of WGS-imputation approach as well as supporting initiatives are in human genetics (FRAZER et al., 2007; McVEAN et al., 2012) and animal breeding (DAETWYLER et al., 2014), in plant breeding they are still scarce and such research efforts are important not only to justify the resequencing cost of the reference panel but the extra work revisiting phenotypic datasets that were already studied in the light of incomplete marker coverage. A recent extensive WGS-based characterization of elite Canadian soybean germplasm has provided such a reference panel suitable for the imputation of missing loci (TORKAMANEH et al., 2018b). To date, few studies performed GWAS using imputed WGS data in plant breeding.

The objective of this study was to use the phenotypic data from a previous GWAS study on seed protein and oil content in soybean and to perform GWAS with a pruned set of 243K SNPs (derived from a set of 2.18M SNPs) obtained through a joint GBS and WGS approach. By comparing the results with a previous GWAS conducted with 17K GBS-derived SNPs (SONAH et al., 2014), we ask four main questions: 1) were new chromosomal regions associated with these traits discovered? 2) was the association or the amount of variation explained by the peak SNP increased? 3) were the haplotype blocks containing the peak SNP more narrowly defined? and 4) how did this impact the identification of candidate genes?

5.2. MATERIAL AND METHODS

5.2.1. Plant material and assessment of soybean protein and oil content

The GWAS panel used in this work was comprised of a core set of 137 short-season (mostly MG0) soybean lines representing the genetic diversity of soybeans cultivated in Canada. These lines were phenotyped for seed protein and oil content in six different environments: three sites (Ottawa, Ontario; Woodstock, Ontario; and St-Mathieu-de-Beloeil, Quebec) over two years (2012 and 2013). All lines were planted in single-row plots with two replications using a generalized lattice design at all the sites. Protein and oil content were measured using a DA 7200 Near Infrared Analyzer spectrometer (Perten Instruments®, Hägersten, Sweden). The phenotypic dataset (except for two lines that were removed) as well as the phenotypic data analysis were the same that were carried out by Sonah et al. (2014).

5.2.2. Genotyping and quality control

Genotyping-by-sequencing was used to provide genome-wide medium-density marker coverage on all lines. DNA samples were extracted using the Qiagen DNeasy 96 Plant kit. The *Ape*K1 restriction enzyme was used for library preparation as per Elshire et al. (2011). A total of ~203 million 100-bp Illumina HiSeq2000 single-end reads were obtained from sequencing a single 192-plex GBS library (which included unrelated samples). The original sequence reads (same as in SONAH et al., 2014) were newly processed using the recently developed SNP-calling pipeline Fast-GBS (TORKAMANEH et al., 2017) as well as a more recent annotation of the soybean genome Wm82.a2.v1 (SONG et al., 2016). Genotypes were called using a minimal read depth of two reads and loci with more than 80% missing data were removed.

Imputation of missing genotypes was first performed on a catalogue of 56K GBS-derived SNPs using BEAGLE v5 (BROWNING; BROWNING, 2007). A second round of imputation, aimed at imputed missing loci, was performed using a reference panel of 4.3M SNPs (derived from WGS of a panel of 102 Canadian elite soybean lines; Torkamaneh et al. 2018b). Among these 102 re-sequenced lines, 56 (41% overlap) were also part of the GWAS panel. After imputation of missing loci, SNPs with a minor allele frequency (MAF) <0.05 and heterozygosity >0.10 were removed using VCFtools (DANECEK et al., 2011), thus

generating a final catalogue of 2.18M SNPs. Finally, pruning was carried out with PLINK (PURCELL et al., 2007) to remove markers in high LD ($r^2 > 0.90$; sliding windows of 50 SNPs), as these could be considered redundant. A final set of 243,291 SNPs was used for the association analyses.

Briefly, for the genotype data of the GWAS study we compare our results to were GBS-sequenced, SNPs- called by the IGST-GBS pipeline and a total of 17,172 SNPs were selected out of 47,702 SNPs after excluding SNPs with more than 20% missing data and MAF < 0.05. More details could be found at Sonah et al. (2014).

5.2.3. Population structure and genetic relatedness

Population structure analysis was performed using fastSTRUCTURE with a set of 14K SNPs resulting from further pruning ($r^2 >0.50$) of the set of 243,291 SNP markers. The optimal number of sub-populations was determined with the chooseK tool as detailed by Malle et al. (2020). To capture genetic relatedness, a kinship matrix was produced using the method of VanRaden (VANRADEN, 2008) implemented in GAPIT (LIPKA et al., 2012).

5.2.4. Association mapping

In total, 243,291 SNPs were used to perform association analyses using a mixed linear model (MLM) in GAPIT (LIPKA et al., 2012). The model incorporated the kinship matrix, along with population structure covariates (K = 7). In order to use the same threshold as Sonah et al. (2014), the negative log(1/n), where n is the number of analyzed SNPs, was used in calling significant associations. The percentage of variation explained by the SNP-trait association (R^2) was estimated as the difference between R^2 of the model with and without the strongest associated SNP, as described by Garcia et al. (2019) Manhattan and quantile-quantile plots (QQ plots) were used to visualize the associations.

5.2.5. Comparison of the associated regions identified with complete vs. incomplete marker coverage

GWAS results obtained by a previous work with 17K GBS-derived set of SNPs (SONAH et al., 2014) were revisited. These results were converted to a more recent annotation of the soybean genome Wm82.a2.v1 to be comparable to the present work results.

The initial set of markers likely provided incomplete genome coverage, while the exhaustive set with 243,291 SNPs, obtained by a combination of GBS- and WGS-derived data followed by pruning, was considered a complete genome coverage here, due its representation as a better picture of the genetic variation among the short-season lines.

5.2.6. Definition of significant regions and identification of candidate genes

For comparison purposes and identification of candidate genes, the most significant SNPs (peak SNPs) were used to anchor QTLs. Since the significant regions reported by the previous work can prove too broad, ranging from one single SNP to 8.60 Mb (for oil content on Chr19; data not shown), we adopted a criterion to re-define the significant regions extension from the previous work as 100 Kb upstream and downstream from the identified peak SNP. For the present work, we assumed the significant region was delimited by the flaking markers that were in perfect LD (D'=1) with the peak SNP. To provide better definition of the genomic regions from the present work, in this step the full catalogue (2.18M SNPs) was used. When two significant SNPs at the same chromosome were found at distances > 800 Kb, with no other significant SNP in-between, they were considered anchors for different QTLs. Finally, the soybean reference genome was used to search for genes located within each significant region containing a peak SNP. Haploview (BARRET et al., 2004) was used for the visualization of genomic regions based on the r^2 (the squared allele frequency correlation between two loci) LD measurement. SNP effects were predicted with the SnpEff v4.3 (CINGOLANI et al., 2012), we built the database with reference genome sequences and gene annotation available at JGI Genome Portal (https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Gmax). Among the genes identified, genes with relevant annotations, regarding the referred traits, were considered as potential candidate genes for protein and oil content.

5.3. RESULTS AND DISCUSSION

5.3.1. Genotype data

To ensure an exhaustive genome-wide marker coverage, we adopted an integrated WGS-GBS genotyping approach in which all lines were subjected to GBS. Following imputation of the missing genotypes, the resulting GBS-derived panel (56K SNPs) was then

subjected to further round of imputation based on a reference panel of 102 re-sequenced lines to achieve complete genome-wide marker coverage (4.3M SNPs). The quality of the resulting data set has been thoroughly assessed by directly comparing the genotypes resulting from GBS + imputation to the actual genotype obtained via WGS and the accuracy of the resulting genotypes was quite high (96.4%) (TORKAMANEH et al., 2018b). The resulting catalogue of fully imputed data was filtered to retain SNPs with a MAF ≥ 0.05, heterozygosity ≤ 0.10, resulting in 2.18M SNPs, followed by pruning to remove markers in high LD ($r^2 > 0.90$). This resulted in a final catalogue of 243,291 SNPs that covered 948.8 Mb of the reference genome (97% of the total captured by the *Glycine max* Wm82.a2.v1 reference genome).

The largest numbers of SNPs were found on chromosomes 15 (27,983 SNPs) and 18 (25,160 SNPs), and the lowest numbers of SNPs were observed on chromosomes 12 (6,032 SNPs) and 11 (3,298 SNPs). The chromosome with the highest marker density was Chr15 (541 SNPs/Mb) and the chromosome with the lowest marker density was Chr11 (95 SNPs/Mb), while the average for the entire genome was 252 SNPs/Mb. Maldonado et al. (2019) reported their SNPs distribution as having a larger number of SNPs on largest chromosomes, with their maximum and minimum number of SNPs on chromosomes 18 and 11, respectively. The distribution of SNPs as well as the SNP density in terms of SNPs per Mb on each chromosome are summarized in Table 1.

Table 1. Distribution of GBS- and WGS-derived SNPs among chromosomes.

Chr	No. of SNPs	SNPs/Mb
1	8,944	157
2	9,603	198
3	18,217	398
4	15,526	296
5	7,306	173
6	11,729	228
7	10,309	231
8	9,342	195
9	13,022	260
10	13,241	257
11	3,298	95
12	6,032	150
13	14,894	325
14	10,108	206
15	27,983	541
16	13,406	354
17	7,502	180
18	25,160	434
19	10,763	212
20	6,906	144
Total/Mean	**243,291**	**252**

Recently, the use of WGS-imputed datasets either from GBS or SNP-chips have been reported in association studies (YAN; McHALE, 2017; BERG et al., 2019; FALKER-GIESKE et al., 2019; WU et al., 2019; BOUDHRIOUA et al., 2020) and also as an attempt to improve the accuracy of genomic selection analysis (SONG et al., 2019).

Wu et al. (2019) performed GWAS analyses with data before and after WGS-imputation from GBS for farrowing interval of different parities in two pig populations. They reported poor imputation accuracy, with average accuracies of whole genome of 0.42 for Landrace and 0.45 for Large White pigs. A possible reason for the relatively low accuracy compared to our study was the limited size of their reference population (20 and 40 Landrace and Large White pigs, respectively) and also the random selection of individuals to be WGS-sequenced. However, retaining only SNPs with imputation accuracies >0.3 and >0.8 resulted on average accuracies of 0.73-0.72 and 0.94-0.93, respectively, which were the files they used. As an example of the dataset in terms of marker enrichment, for Large White pigs, the data before imputation and after imputation/filtration varied in size from 326K (GBS) to 2.5M (>0.8) and 5.5M (>0.3) (WGS-imputed).

In Song et al. (2019), the imputation accuracy from 80K SNP-chip to WGS variants reached 0.92. As another example of applied WGS-imputation, in Falker-Gieske et al. (2019) GWAS work for meat and carcass traits in pigs, founders (24), F1 (91) and F2-designs (2,657 individuals) were genotyped by high coverage WGS, low coverage WGS and SNP chip, respectively. Their imputation scheme was as it follows: F1 were imputed to high WGS level and, posteriorly, founders and F1 animals served as the reference panel for the imputation of the F2 to WGS sequence level to subsequent GWAS.

Such genotyping approach and the use of WGS-imputed data is becoming more common over the second half of the decade. According to Falker-Gieske et al. (2019), it can be considered a reasonable strategy to increase the marker density to WGS level and provides affordable opportunity to improve the power and datasets usage potential.

5.3.2. Population structure

Population structure among accessions of this association panel was determined using fastSTRUCTURE and the optimum delta K ranged from 6 to 9 subpopulations, with the optimal number chosen to be seven (Figure 1). In the previous GWAS work (SONAH et al. 2014), the number of groups chosen was nine. The higher number of groups of the previous

work compared to the present work can be justified by the fact that for previous work the population structure analysis was done with a diverse set of 304 lines from which the association panel was later retrieved. Malle et al. (2020) reported that results from fastSTRUCTURE were consistent with results obtained by phylogenetic tree analysis as well as PCA-based population structure.

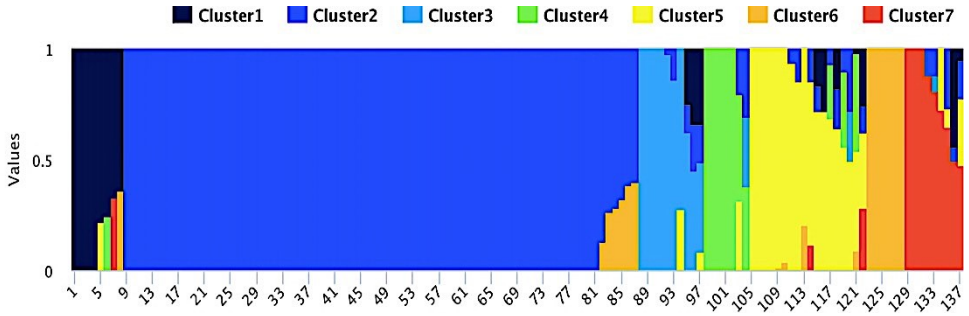


Figure 1. Population structure among a set of 137 Canadian soybean lines. The optimal number of clusters was found to be K=7. Each soybean line is represented by a single vertical line and each color represents one cluster.

5.3.3. Marker-trait associations with WGS-imputed dataset

Association analyses were performed to detect genomic regions controlling seed protein and oil content in soybean using 243K SNPs that were polymorphic within a set of 137 lines representative of the germplasm used in breeding short-season soybean varieties in Canada. An MLM model taking into consideration both kinship (K matrix) and population structure (Q matrix) was used. A negative log(1/n) threshold was used for declaring market-trait associations significance.

Seven and eight QTL regions were identified as associated with protein and oil content, respectively. As shown in the Manhattan plots (Figure 2), numerous highly significant marker-trait associations piled up neatly in QTL regions. The peak SNP of each QTL was taken to best capture each QTL, more details of their association are provided in Supplementary S1 and S2 Tables.

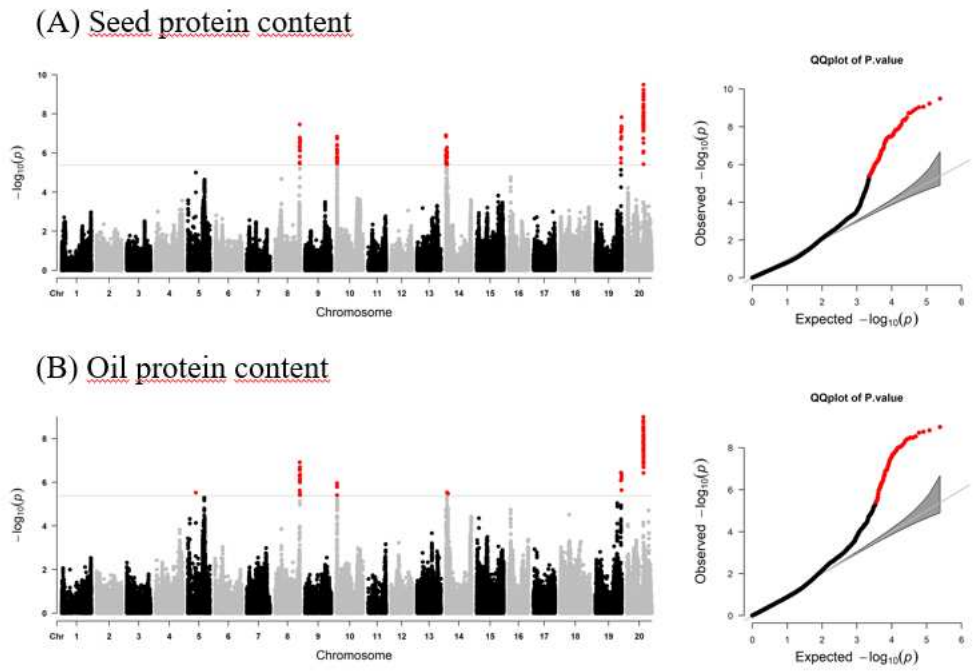


Figure 2. Manhattan plots for genome-wide association of protein and oil content. The data for the 137 Canadian soybean lines were analyzed using a mixed linear model for (A) seed protein content and (B) seed oil content. The -log10(p) values from the GWAS are plotted for each SNP position on each chromosome. The significance threshold (negative log(1/n)) is indicated by the horizontal line.

5.3.4. Comparing GWAS Results for complete and incomplete marker coverage

In order to investigate whether we could identify new QTLs associated with protein and oil content from the phenotypic dataset of 137 Canadian soybean lines by having a denser set of SNPs, we revisited results obtained by Sonah et al. (2014). The initial set of markers consisted of 17K SNPs likely provided an incomplete genome coverage, while the exhaustive set consisted of 243K SNPs obtained by a GBS-WGS approach was considered to be closer to a complete genome representation of the genetic variation.

In the previous work, with incomplete marker coverage, eight significant regions were found to be associated with protein content, whereas in the present work seven regions were identified (Table 2). Both GWAS results pointed significant associations for protein content on chromosomes 8, 10, 14, 19 and 20. Once the significant regions revealed by the previous work, with limited coverage, extended a genetic distance that translated into large genomic regions, that would be otherwise unfeasible to search for candidate genes within, here we limited them to 100 Kb upstream and downstream from the peak-SNP anchor. For the

present work, we delimited the significant region by flaking markers in perfect LD (D'=1) with the peak SNP.

Table 2. Characteristics of the peak SNPs defining each associated region with seed protein content.

Incomplete coverage

Chr	Peak SNP Position	QTL ID	*p*-value	R^2 [a]	Significant region (+/- 100 Kb from peak SNP) Start	End	Overlapping[b]
5	31,380,926	q_Pro_5-1	9.72E-07	0.13	31,280,926	31,480,926	X
8	46,415,576	q_Pro_8-1	6.53E-07	0.13	46,315,576	46,515,576	X
10	1,330,775	q_Pro_10-1	1.00E-07	0.16	1,230,775	1,430,775	X
14	126,916	q_Pro_14-1	3.36E-08	0.17	26,916	226,916	58.12-74.06 Kb
16	4,204,153	q_Pro_16-1	1.30E-06	0.13	4,104,153	4,304,153	X
19	50,539,611	q_Pro_19-1	2.48E-07	0.14	50,439,611	50,639,611	X
20	18,345,815	q_Pro_20-1	1.02E-05	0.10	18,245,815	18,445,815	X
20	31,756,869	q_Pro_20-2	3.55E-05	0.08	31,656,869	31,856,869	31.76-31.86 Mb

Complete coverage

Chr	Peak SNP Position	QTL ID	*p*-value	R^2	Significant region (flanking markers in high LD with peak SNP) Start	End	Overlapping
8	47,147,391	Q_Pro_8-1	3.48E-08	0.19	47,141,680	47,185,506	X
10	1,609,531	Q_Pro_10-1	1.48E-07	0.17	1,605,797	1,611,678	X
14	58,133	Q_Pro_14-1	1.23E-07	0.17	58,133	74,056	58.12-74.06Kb
14	1,771,702	Q_Pro_14-2	5.38E-07	0.15	1,771,271	1,829,944	X
19	49,037,060	Q_Pro_19-1	8.55E-08	0.18	48,991,643	49,076,176	X
19	49,914,498	Q_Pro_19-2	1.50E-08	0.20	49,912,458	49,915,487	X
20	31,874,531	Q_Pro_20-1	3.24E-10	0.26	31,764,481	31,874,531	31.76-31.86 Mb

[a] R^2 was calculated by the R^2 of the model with SNP minus R^2 of model without SNP

[b] Overlapping common significant region found in GWAS work with incomplete and complete genome coverage

Using the same method to calculate the thresholds, with incomplete marker coverage, 64 SNPs were identified as associated with protein content and with complete marker coverage the number of significant SNPs was 105. Out of the seven regions identified by the present work, two of them have some extent in common with regions reported by the previous work, on Chr14, Q_Pro_14-1 and q_Pro_14-1, peak SNPs at 58,133 (*p*-value = 1.23E-07) and 126,916 (*p*-value = 3.36E-08), and on Chr20, Q_Pro_20-1 and q_Pro_20-2, peak SNPs at 31,874,531 (*p*-value = 3.24E-10) and 31,756,869 (*p*-value = 3.55E-05) (Table 2). From the five regions left, Q_Pro_10-1 could be considered somewhat neighbor region (< 300 Kb apart) from previously found q_Pro_10-1, peak SNPs at 1,609,531 (*p*-value = 1.48E-07) and 1,330,775 (*p*-value = 1.00E-07), respectively; and Q_Pro_8-1, Q_Pro_14-2, Q_Pro_19-1 Q_Pro_19-2 could be considered new regions identified by the present work.

Although Q_Pro_8-1 would have some sharable extent with the original broad significant region (~1.4 Mb; data not shown) detected by the past work on Chr08 (before we redefined it to 100 Kb limiting around the peak SNP).

In general, the amount of variation explained (R^2) appeared to increase when using complete marker dataset, it ranged from 0.15 to 0.26, with incomplete marker it ranged from 0.08 to 0.17. Taking the significant regions that had some sharable extent as examples (Q_Pro_14-1 and q_Pro_14-1, Q_Pro_20-1 and q_Pro_20-2), the amount of variation explained by their peak SNPs were 0.17 and 0.26 (complete coverage) and 0.17 and 0.08 (incomplete coverage), for Chr14- and Chr20-regions, respectively (Table 2). In the Chr14-region, the percentage of variation explained by the peak SNP remained the same (0.17) using complete and incomplete coverage while for the Chr20-region it had an expressive increase (from 0.08 to 0.26). As it will be discussed later, this Chr-20 region is a well-known region reported in the literature as an important associated genome region with protein and oil content.

Another benefit of revisiting a study with a particular phenotype data with enriched genotype datasets, besides finding new regions and/or even validate previous identified regions with an increased amount of variation, is to narrow the significant regions into blocks in high linkage. In the manuscript of the previous work (SONAH et al., 2014), broad significant regions were reported, on average extending > 500 Kb for protein content associated regions, now that we limited the significant regions to 100 Kb up and downstream obviously this average was down to 200 Kb. For present work, limiting the haplotype blocks with a high LD, they were averaging 46 Kb for protein content associated regions.

For comparison purposes, the same method of calculating the significance threshold of the previous work was adopted for the present work, however, if we have adopted an FDR of 0.05 instead, as this threshold is also commonly used for GWAS studies, we would have found an additional overlapping significant region between GWAS work with incomplete and complete genome coverage on Chr16 (S3 Table).

For the above mentioned overlapping significant regions identified for protein content, on Chr 14 and 20 (Table 2), and using FDR<0.05, on Chr 16 (S3 Table), the linkage between their peak SNPs is presented on Table 3. They were quite high ($r^2 > 0.72$ and $D' > 0.87$), indicating a good correspondence between some of the previous and present work findings on the identification of associated regions of interest.

Table 3. Linkage between peak SNPs identified by GWAS work with incomplete and complete genome coverage for protein content.

Chr	Peak SNP Position Incomplete coverage	Peak SNP Position Complete coverage	r^2	D'
14	126,916	58,133	0.85	0.96
[a]16	4,204,153	4,099,404	0.79	0.94
20	31,756,869	31,874,531	0.72	0.87

[a] non-significant under the negative log (1/n) threshold (Table 2) but under FDR<0.05 it would be significant (S3 Table).

For oil content, eight significant regions were found in the previous and present work (Table 3). Both GWAS results pointed significant associations for oil content on chromosomes 5, 8, 10, 14, 19 and 20. The number of significant SNPs associated with oil content and with complete and incomplete marker coverage were 70 and 79, respectively.

None of the identified regions by the present work, with complete coverage, and the previous work, with incomplete coverage (after narrowing them 100 Kb around the peak SNP), have shared common extension, however, they were in some cases neighboring regions (< 300 Kb apart). For example, on Chr10, q_Oil_10-1 and Q_Oil_10-1, peak SNPs at 1,815,850 (*p*-value = 3.86E-07) and 1,609,531 (*p*-value = 1.11E-06), as well as on Chr20, q_Oil_20-2 and Q_Oil_20-1, peak SNPs at 31,866,045 (*p*-value = 1.18E-05) and 31,606,891 (*p*-value = 1.02E-09) (Table 4). The six regions left (identified by the present work), Q_Oil5-1, Q_Oil_8-1, Q_Oil_14-1, Q_Oil_14-2, Q_Oil_19-1 and Q_Oil_19-2 can be considered new regions when comparing the previous and current study, both carried out with the same phenotypic data. Although Q_Oil_8-1, Q_Oil_19-1 and Q_Oil_19-2 would have some sharable extent with the original broad significant region reported by the previous work on Chr 08 and 19 (1.4 and 8.6 Mb, respectively; data not shown) before we have narrowed it towards the peak SNP.

Table 4. Characteristics of the peak SNPs defining each associated region with seed oil content.

Incomplete coverage

Chr	Peak SNP Position	QTL ID	*p*-value	R^2 [a]	Significant region (+/- 100 Kb from peak SNP) Start	End	Overlapping[b]
5	31,380,926	q_Oil_5-1	1.61E-07	0.19	31,280,926	31,480,926	X
8	46,415,576	q_Oil_8-1	6.09E-06	0.14	46,315,576	46,515,576	X
10	1,815,850	q_Oil_10-1	3.86E-07	0.18	1,715,850	1,915,850	X
14	65,161	q_Oil_14-1	1.95E-06	0.15	1	165,161	X
16	4,204,153	q_Oil_16-1	1.62E-06	0.16	4,104,153	4,304,153	X
19	50,539,611	q_Oil_19-1	1.09E-06	0.16	50,439,611	50,639,611	X
20	18,345,815	q_Oil_20-1	1.02E-06	0.16	18,245,815	18,445,815	X
20	31,866,045	q_Oil_20-2	1.18E-05	0.13	31,766,045	31,966,045	X

Complete coverage

Chr	Peak SNP Position	QTL ID	*p*-value	R^2	Significant region (flanking markers in high LD with peak SNP) Start	End	Overlapping
5	15,338,484	Q_Oil_5-1	3.02E-06	0.15	15,316,474	15,410,120	X
8	47,343,134	Q_Oil_8-1	1.22E-07	0.20	47,316,318	47,365,208	X
10	1,609,531	Q_Oil_10-1	1.11E-06	0.16	1,605,797	1,611,678	X
14	1,755,977	Q_Oil_14-1	2.93E-06	0.15	1,754,439	1,779,146	X
14	3,795,066	Q_Oil_14-2	3.27E-06	0.15	3,794,998	3,796,247	X
19	49,037,060	Q_Oil_19-1	3.63E-07	0.18	48,991,643	49,076,176	X
19	49,914,498	Q_Oil_19-2	4.01E-07	0.18	49,912,458	49,915,487	X
20	31,606,891	Q_Oil_20-1	1.02E-09	0.27	31,605,400	31,623,926	X

[a] R^2 was calculated by the R^2 of the model with SNP minus R^2 of model without SNP

[b] Overlapping common significant region found in GWAS work with incomplete and complete genome coverage

For oil content, as it was for protein content, the amount of variation explained (R^2) appeared to increase when using complete marker dataset, it ranged from 0.15 to 0.27, comparatively, with incomplete marker that ranged from 0.13 to 0.19.

Broad significant regions were reported by the previous work, before we have delimited them to 100 Kb limits around the peak SNPs, sizes reached up to 8.6 Mb. For present work, with complete coverage, for oil content those regions were on average 35 Kb length.

If we have adopted a more permissive threshold (FDR < 0.05), we would have found three overlapping significant regions between GWAS work with incomplete and complete genome coverage for oil content, on Chr 5, 14 and 16 (S4 Table), respectively with q_Oil_5-1, q_Oil_14-1 and q_Oil_16-1 (Table 4). The same way it occurred for protein content, for oil content the linkage between peak SNPs identified by the previous GWAS work with

incomplete genome coverage and the present GWAS work with complete genome coverage (with FDR < 0.05; S4 Table) were high (r^2 > 0.81 and D' > 0.92; S5 Table).

Although the preliminarily expectation is to find more significant regions by having a more complete coverage, it seems each case should be investigated separately, once the outcome result can vary. In our work, using a denser marker dataset the number of significant associations for protein increased (from 64 to 105 SNPs) while it decreased for oil (from 79 to 70 SNPs), but in both cases it culminated in similar number of QTL regions compared to the previous work. Van der Berg et al. (2019) reported the detected QTLs increased with increasing marker density. In Wu et al. (2019), the number of significant SNPs decreased using WGS-imputed data on GWAS. Nevertheless, in the latter case the authors concluded association analyses using imputed WGS data rather than GBS data appeared to reduce the mapping noise and highlight the important peaks for their trait of interest.

For the present work, from the seven and eight regions described for protein and oil content, respectively, three peak SNPs (on Chr10 – position 1,609,531 – and Chr19 – positions 49,037,060 and 49,914,498 –) were identical for protein and oil content. Sonah et al. (2014) reported that the regions they identified for protein content were practically identical of those they found for oil content. In their work, from the eight regions, five peak SNPs were identical for protein and oil content. Both studies found that all favorable alleles for one trait were unfavorable to the other trait.

5.3.5. Significant region on Chr 20

Significant regions on 31.61-31.97Mb-segment of Chr20 were identified for protein and oil in both studies (using incomplete and complete marker coverage) drew attention at a first sight, because it seems those regions can be assumed as part of a major QTL region which is often described in the association studies' literature for protein and oil content in soybean.

For the present work, the QTL/peak SNP showing the strongest association with protein content was located on Chr20 at 31,874,531 (*p*-value = 3.24E-10) and accounted for 0.26 of the phenotypic variation for this trait. For oil content, the most significant QTL/peak SNP was also on Chr20 at 31,606,891 (*p*-value = 1.02E-09) and explained 0.27 of the variation. Significant regions surrounding the peak SNPs Gm20:31,874,531 and Gm20:31,606,891 spanned 110.05 and 18.53 Kb, respectively (Tables 2 and 4). In spite of the proximity of the two QTL regions (31.61-31.62 Mb and 31.76-31.87 Mb) and the

moderate linkage between their peak SNPs (r^2=0.54 and D'=0.79), they fall within two distinct significant regions, as illustrated in supplementary S1 Figure.

The peak SNPs identified on Chr20 by our study were Gm20:31,874,531 and Gm20:31,606,891 for protein and oil content, respectively; by the previous work they were Gm20:31,866,045 and Gm20:31,756,869, respectively. Important to note that the peak SNP for oil content identified by the previous work is inside that high linkage significant region identified by the present work for protein content (Gm20:31,766,045-31,966,045), whereas the peak SNP for protein content identified by the previous work is only 9 Kb downstream apart from this referred region.

Van and McHale (2017) carried out meta-analyses from over 80 studies for seed contents of protein and oil, for Protein+Oil they identified an interval from 27.99 to 34.05Mb on Chr20 which overlaps with the regions identified by the present and previous work (SONAH et al., 2014), with complete and incomplete coverage, respectively, for protein and oil content analyzed separately. For oil content, Van and McHale (2017) identified the region 28.54 – 35.08Mb on Chr20, which also overlaps the genomic region that we have found.

Several other studies reported an associated additive effect for seed protein and/or oil content detected on Chr20 (BOLON et al., 2010; HWANG et al., 2014; VAUGHN et al., 2014; BANDILLO et al., 2015; VAN; McHALE, 2017; LEE et al., 2019). Pioneer work on the identification of QTL for high protein and oil at Chr20 was done by Diers et al. (1992) with RFLP analysis. Nichols et al. (2006) fine mapped the QTL region and narrowed it as a 3 cM interval using BC5F5-derived near-isogenic lines (NILs) contrasting in seed protein and oil. Bolon et al. (2010) using a large number of simple sequence repeat (SSR) markers established an 8.4Mb region (24.54 to 32.92Mb) on Chr20 as a potential region to harbor candidate genes responsible for pleiotropic soybean seed protein and oil QTL. Further studies supported narrowing this region (HWANG et al., 2014; VAUGHN et al., 2014). Vaughn et al. (2014) reported an association between protein content at the 1.0Mb genomic region between 30.93 and 31.97 Mb on Chr20, with the greatest effect associated with the SNP marker at position 31,972,955, for the population MS-2000, with mostly MGV accessions; for oil content, the markers with the lowest significant *p*-values resided in the 31.15-31.64 Mb block (physical distances described in their manuscript were based on Gmax1.01 assembly). Thus, the present study as well as Sonah et al. (2014) further confirmed the importance of this QTL for soybean protein and oil content.

This all lead us hypothesize that the relationship between traits in this Chr20-genome region may be due to either pleiotropic gene effects or very tight linkage. Chung et al. (2013),

comparing a *Glycine soja* population with a *Glycine max* population, identified the 30.5-32.3 Mb region on Chr20 as being under selection related to domestication. Bandillo et al. (2015) reported that haplotype alleles on Chr20 displayed a negative relationship between the protein and oil content. Given their mapping resolution and the presence of few likely candidate genes in the detected genomic regions, they hypothesize pleiotropic gene effects underlie the observed negative correlation between oil and protein at the identified loci. They also have not excluded the possibility of multiple genes in very tight repulsion-phase linkage since the sizes of the haplotype blocks found by them in Chr20 region were > 500 Kb. If is the latter, it brings an issue/sort of questions like: how large would a segregating population have to be to obtain varieties with favorable alleles for protein and oil content in this region? Would that strategy be feasible when designing field trials? That is one of the reasons why is important to identify other important major QTLs for these traits and do not rely only on this genomic region for MAS when one wants to improve protein and oil content simultaneously.

5.3.6. Candidate genes

Haplotype blocks around the peak SNPs identified by the GWAS using complete coverage data were determined based on LD, while the significant regions from the previous work, with incomplete coverage were narrowed to 100 Kb up and downstream from the peak SNPs; those were used to guide the candidate genes search.

For the significant regions associated with protein content (Table 2), 146 and 34 genes were found on Soybase database, the soybean reference genome, for incomplete and complete coverage, respectively. For oil content's significant regions (Table 4), 140 and 21 genes for incomplete and complete coverage, respectively, were identified. Among the genes found within significant regions by using complete (WGS imputed) or incomplete coverage (GBS data), seven genes in common were found for protein content, while for oil content no common genes were identified. The reduced number of genes using the complete coverage datasets could be explained as the significant regions were narrowly defined in the present work, or understood as they were inserted in gene-poor regions, the first case seems to be the explanation.

Based on the currently available annotation, some of those genes cannot be readily be associated with protein and oil content. A list with candidate genes that seem to be related to the traits of interest as well as their annotations are provided in Supplementary S6 Table.

Out of the seven genes identified in common between the two studies, with complete and incomplete coverage, namely Glyma.14G000400, Glyma.14G000500, Glyma.20G085100, Glyma.20G085200, Glyma.20G085300, Glyma.20G085400 and Glyma.20G085500, none of them had functional annotations that led us to believe that they were directly associated in soybean protein biosynthesis or seed composition traits. However, recent studies indicate that two (Glyma.20G085100 and Glyma.20G085200) out of these seven may be important in determining the protein content in soybeans. Ward (2017), aiming to reduce candidate gene region of cqProt-003, designation of a QTL on Chr 20 given by Nichols et al. (2006) and one of the most studied QTL in soybean breeding (FLIEGE, 2019), narrowed cqSeed protein-003 to a 77.8 kb region on Chr 20 (31,744,150 – 31,821,947; Gmax2.0 map assembly) and performed the QTL fine mapping in a population of near isogenic lines (NILs) that were segregating for cqProt-003 (FLIEGE, 2019). Within the 77.8 kb interval, Ward (2017) identified three candidate genes: Glyma.20g085000, Glyma.20g085100, and Glyma.20g085200. As their functional annotation directly not seem to be involved with protein content, Fliege (2019) futher analyzed the sequence of these genes for potential variants. Fliege (2019) reported structural variant at Glyma.20G085100, and according to the author the insertion at the gene in *G. max* (Williams 82) caused a low protein phenotype in Williams 82, and at some frequency the sequence reverts to the ancestral, PI468916 (*G. soja*) form, producing seed with high protein content.

Among the candidate genes, with annotation that seem to be related to protein biosynthesis, within the significant regions identified by the present work (complete coverage) there is Glyma.14G024700, it is 9 Kb upstream Q_Pro_14-2 peak SNP, it has molecular function of protein serine/threonine phosphatase activity and is described as being involved with the biological process of proline transport. The gene Glyma.19G262900, identified within the significant region q_Pro_19-1 (by the previous work) and 14 Kb upstream its peak SNP, also were assign to be involved with proline transport. Proline is an important amino acid that have several regulatory roles besides participates in general protein synthesis and seed development (SZABADOS; SAVOURÉ 2010; KISHOR et al. 2015). Wang et al. (2014) reported the map-based cloning and functional characterization of proline responding1 gene (pro1) in maize. They also unraveled the mechanism related to this mutation. Lack of proline affects general protein synthesis/accumulation in pro1 mutants.

Glyma.19G242300, located 26 Kb downstream the peak SNP of Q_Pro_19-1, has cysteine synthase activity as one of its molecular functions and cysteine biosynthetic process among its molecular processes. Cysteine is one of the few sulfur-containing amino-acids.

Although cysteine plays a very important role in stabilization of protein structure at higher level, conferring rigidity and proteolytic resistance, soybean proteins are relatively poor in the sulfur-containing amino acids like cysteine (PANTHEE et al., 2006). Selecting cysteine-rich protein cultivars and identifying QTL associated with cysteine could assist in the enhancement of the nutritional quality of the soybean meal (PANTHEE et al., 2006; KRISHNAN et al., 2015).

Several of the identified candidate genes for protein content (with complete marker coverage) could be listed here, for example, genes with molecular functions of protein serine/threonine kinase and protein tyrosine kinase activities (Glyma.14G025200 and Glyma.19G242900), structural constituent of ribosome (Glyma.19G242700 and Glyma.20G085400), endopeptidase, peptidase and threonine-type endopeptidase activities (Glyma.19G242200), involved with biological processes such post-translational protein modification (Glyma.19G242400).

Relevant candidate genes for protein content were also found within the significant regions identified by the previous work (SONAH et al., 2014) associated with important biological process and/or molecular functions like methionine biosynthetic process (Glyma.08G350600), protein serine/threonine phosphatase activity (Glyma08.351000), acetyl-CoA metabolic process (Glyma.08G351200), endopeptidade, peptidase, ribonuclease and threonine-type activities (Glyma.10G014300), asparagine catabolic process (Glyma.10G015400), serine/threonine kinase activity (Glyma.14G000800), cysteine, methionine and threonine biosynthetic process (Glyma.14G001000), protein serine/threonine and tyrosine kinase activities (Glyma.14G001300 and Glyma.19G261700), regulation of translational initiation (Glyma.19G262500) and serine-type endopeptidase activity (Glyma.19262600).

The relationship between asparagine metabolism and protein concentration in soybean seed was investigated by Pandurangan et al. (2012), their results highlighted differences between high and low protein soybean genotypes related to the metabolism of asparagine in developing seed. Wang et al. (2019) discussed the possibility of protein and oil content being associated with soluble sugar utilization traits; they found that high levels of free asparagine, aspartic acid and glutamic acid in high protein genotypes may provide favorable conditions for storage protein synthesis, thus consuming an excess of C-skeletons and resulting in insufficient C-skeletons for oil synthesis, leading to lower seed oil content.

According to Hernández-Sebastià et al. (2005), the negative relationship between seed protein and oil content may be related to the regulation of carbon (C) flux between the

competing synthetic pathways, that require carbon C-skeletons derived from imported sucrose during seed development. The flux through pyruvate and to acetyl-CoA is conceivably an important regulator that can partially account for the negative correlations between protein and oil content (WILCOX; SHIBLES, 2001).

For oil content, within the significant regions identified by the present work (complete coverage), Glyma.19G242200 was the candidate gene identified. It is involved in the biologic process of fatty acid beta-oxidation. Although the high percentage of polyunsaturated fatty acids in soybean oil is desirable to meet human nutrition requirements, polyunsaturated fatty acids are susceptible to oxidative reactions.

A high number of candidate genes for oil content were found within the significant regions identified by the previous work (SONAH et al., 2014), associated with biological process like fatty acid biosynthetic process (Glyma.08G349200), acetyl-CoA metabolic process (Glyma.08G351200), lipid metabolic process (Glyma.19G261600, Glyma.19G262800, Glyma.19G262900 and Glyma.20G085600), lipid oxidation (Glyma.19G263300), lipid storage (Glyma.19G263400) and fatty acid biosynthetic and metabolic processes (Glyma.20G060100). Pathways for oil accumulation have been studied and characterized in plants, several key genes are involved in the process of fatty acid biosynthesis (ZHANG et al., 2016).

New regions were found revisiting previous phenotypic data, as well as regions identified previously with a less dense panel were not observed when exploring a higher genome coverage. Few regions identified by both GWAS works overlapped, and some of them were somewhat neighbors. Although the coincident significant regions represent good reproducibility, we have used a distinct genotypic dataset. Then, differences in the QTLs identification that encompasses the distinct input datasets were already expected. The higher marker coverage achieved by the imputation with WGS-markers allowed us to detect novel regions, and by exploring the linkage within markers narrowing the significant regions.

Yan et al. (2017) discussed the GWAS based on a whole-genome imputation approach can be a powerful strategy to analyze economically important complex traits in livestock. They rediscovered a missing QTL for lumbar number in Sutai pigs by carrying out a GWAS based on a whole-genome imputation strategy. In a soybean GWAS recent study, Boudhrioua et al. (2020) adopted an approach like that and discovered a novel marker-trait association with resistance to *Sclerotinia* stem rot on Chr 01. They have used almost the same phenotypic data of a previous work (BASTIEN et al., 2014), and that incremented panel with a higher number of SNPs made possible to find a new marker-trait discovery.

Falker-Gieske et al. (2019) by using WGS-imputed data reported that they not only confirmed previously reported but also discovered new QTLs. Wu et al. (2019) found association analyses using imputed WGS data rather than GBS data appeared to reduce the mapping noise and highlight the important peaks for their trait of interest. In our work, however, it was observed that candidate genes related to the traits of interest were found in significant regions identified using both approaches, imputed WGS and GBS data.

These results support the idea that by having a less expensive genome scan (e.g. GBS), with a uniform physical distribution of SNPs, can prove successful to detect trait-related QTLs, the incomplete coverage dataset was enough to identify major QTLs (as it was the case of Chr20 region for protein and oil content). It also provides insights into solutions that combine second-generation sequence data with GWAS as well as on genome important locations for key traits in soybean breeding, providing breeders with markers for selection aiming to improve protein and oil and content.

5.4. CONCLUSIONS

Achieving a dense catalogue of polymorphisms due integration of genotyping approaches allowed us to better exploit LD to delimit significant genomic regions, as well as to find associations with an increased amount of variation explained by their peak SNPs. Additionally, new marker-trait associated regions compared to previous work were identified. Regardless, the latter, even with a very limited genotypic dataset was able to successfully identify associated regions to the traits of interest. These results support the idea that on one hand by having a less expensive genome scan (e.g. GBS), with fairly uniform SNPs' distribution, can prove successful to major QTL identification, whereas on the other hand by having a denser marker-coverage in addition to validate previous results (if any) with more prominent peaks, it is possible to identify new marker-trait associations explaining an increased amount of variation, to further investigate the landscape around the peak SNP and to have haplotype blocks around peak SNPs more narrowly defined. According to our findings, the choice of using a GBS or a WGS-imputed data, or even WGS data, on GWAS, seems to rely mostly on the final-user objective, assuming the pros and cons that each strategy brings.

5.5. ACKNOWLEDGMENTS

Lívia Gomes Torres was supported by scholarships from *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) and *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq).

5.6. REFERENCES

BANDILLO, N.; JARQUIN, D.; SONG, Q.; NELSON, R.; CREGAN, P.; SPECHT, J.; LORENZ, A. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. **The Plant Genome**, v. 8, p. 1-13, 2015.

BARRETT, J. C.; FRY, B.; MALLER, J.; DALY, M. J. Haploview: Analysis and visualization of LD and haplotype maps. **Bioinformatics**, v. 21, p. 263-265, 2004.

BASTIEN, M.; SONAH, H.; BELZILE, F. Genome wide association mapping of *Sclerotinia sclerotiorum* resistance in soybean with a genotyping-by-sequencing approach. **The Plant Genome**, v. 7, p. 1-13, 2014.

BOLON, Y. T.; JOSEPH, B.; CANNON, S. B.; GRAHAM, M. A.; DIERS, B. W.; FARMER, A. D.; MAY, G. D.; MUEHLBAUER, G. J.; SPECHT, J. E.; TU, Z. J.; WEEKS, N.; XU, W. W.; SHOEMAKER, R. C.; VANCE, C. P. Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. **BMC Plant Biology**, 10:41, 2010.

BOUDHRIOUA, C.; BASTIEN, M.; TORKAMANEH, D.; BELZILE, F. Genome-wide association mapping of *Sclerotinia sclarotiorum* resistance in soybean using whole-genome resequencing data. **BMC Plant Biology**, v. 20, p. 195, 2020. https://doi.org/10.1186/s12870-020-02401-8,

BROWNING, S. R.; BROWNING, B. L. Rapid and accurate haplotype phasing and missing data inferente for whole genome association studies by use of localized haplotype clustering. **Am. J. Hum. Genet.**, v. 81, p. 1084-1097, 2007.

CINGOLANI, P.; PLATTS, A.; LE WANG, L.; COON, M.; NGUYEN, T.; WANG, L.; LAND, S. J.; LU, X.; RUDEN, D. M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. **Fly**, v. 6, n. 2, p. 80-92, 2012.

CHANG, H. X.; BROWN, P. J.; LIPKA, A. E.; DOMIER, L. L.; HARTMAN, G. L. Genome-wide association and genomic prediction identifies associated loci and predicts the sensitivity of Tobacco ringspot virus in soybean plant introductions. **BMC Genomics**, v. 17, p. 153, 2016.

CHUNG, W. H.; JEONG, N.; KIM, J.; LEE, W. K.; LEE, Y. G.; LEE, S. H.; YOON, W.; KIM, J. H.; CHOI, I. Y.; MOON, J. K.; KIM, N.; JEONG, S. C. Population structure and

domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. **DNA Res.**, v. 21, p. 153-167, 2013.

CONTRERAS-SOTO, R. I.; MORA, F.; OLIVEIRA, M. A. R.; HIGASHI, W.; SCAPIM, C. A.; SCHUSTER, I. A Genome-Wide Association Study for Agronomic Traits in Soybean Using SNP Markers and SNP-Based Haplotype Analysis. **Plos One**, v. 12, p. e0171105, 2017.

COPLEY, T. R.; DUCEPPE, M.; O'DONOUGHUE, L. S. Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines. **BMC Genomics**, v. 19, p. 167, 2018.

DAETWYLER, H.; CAPITAN, A.; PAUSCH, H. et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. **Nature Genetics**, v. 46, p. 858-865, 2014. https://doi.org/10.1038/ng.3034

DANECEK, P.; AUTON, A.; ABECASIS, G.; ALBERS, C. A.; BANKS, E.; DEPRISTO, M. A.; HANDSAKER, R. E.; LUNTER, G.; MARTH, G. T.; SHERRY, S. T.; McVEAN, G.; DURBIN, R. The variant call format and VCFtools. **Bioinformatics**, v. 27, p. 2156:2158, 2011.

DIERS, B. W.; KEIM, P.; FEHR, W. R.; SHOEMAKER, R. C. RFLP analysis of soybean seed protein and oil content. **Theoretical and Applied Genetics**, v. 83, p. 608-612, 1992.

ELSHIRE, R. J.; GLAUBITZ, J. C.; SUN, Q.; POLAND, J. A.; KAWAMOTO, K.; BUCKLER, E. S.; MITCHELL, S. E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS ONE**, v. 6, p. e19379, 2011.

FALKER-GIESKE, C.; BLAJ, I.; PREUß, S.; BENNEWITS, J.; THALLER, G.; TETENS, J. GWAS for meat and carcass traits using imputed sequence level genotypes in pooled F2-designs in pigs. **G3: Genes, Genomes, Genetics**, v. 9, p. 2823–2834, 2019. https://doi.org/10.1534/g3.119.400452

FLIEGE. Genomic changes underlying disease resistance and high protein QTL (Unpublished doctoral dissertation) University of Illinois at Urbana-Champaign. Urbana, IL, 2019. https://www.ideals.illinois.edu/bitstream/handle/2142/106255/FLIEGE-DISSERTATION-2019.pdf?sequence=1. Accessed in: 07 Aug 2020.

FRAZER, K.; BALLINGER, D.; COX, D. et al. A second generation human haplotype map of over 3.1 million SNPs. **Nature**, v. 449, p. 851-861, 2007. https://doi.org/10.1038/nature06258.

GARCIA, M.; ECKERMANN, P.; HAEFELE, S.; SATIJA, S.; SZNAJDER, B.; TIMMINS, A.; BAUMANN, U.; WOLTERS, P.; MATHER, D. E.; FLEURY, D. Genome-wide association mapping of grain yield in a diverse collection of spring wheat (*Triticum aestivum* L.) evaluated in Southern Austraria. **Plos ONE**, v. 14, n. 2, p. e0211730, 2019.

HERNÁNDEZ-SEBASTIÀ, C.; MARSOLAIS, F.; SARAVITZ, C.; ISRAEL, D.; DEWEY, R. E.; HUBER, S. C. Free amino acid profiles suggest a possible role for asparagine in the

control of storage-product accumulation in developing seeds of low- and high-protein soybean lines. **Journal of Experimental Botany**, v. 56, p. 1951-1963, 2005.

HWANG, E.-Y.; SONG, Q.; JIA, G.; SPECHT, J. E.; HYTEN, D.; COSTA, J.; CREGAN, P. B. A genome-wide association study of seed protein and oil content in soybean. **BMC Genomics**, v. 15, n. 1, p. 1-12, 2014.

IQUIRA, E.; SONAH, H.; BELZILE, F. Association mapping of QTLs for sclerotinia stem rot resistance in a collection of soybean plant introductions using a genotyping by sequencing (GBS) approach. **BMC Plant Biology**, p. 1-12, 2015. doi 10.1186/s12870-014-0408-y.

KISHOR, P. B. K.; KUMARI, P. H.; SUNITA, M. S. L.; SREENIVASULU, N. Role of proline in cell wall synthesis and plant development and its implications in plant ontogeny. **Frontiers in Plant Science**, v. 6, p. 544, 2015. doi: 10.3389/fpls.2015.00544

KRISHNAN, H. B.; KIM, W. S.; OEHRLE, N. W.; ALASWAD, A. A.; BAXTER, I.; WIEBOLD, W. J.; NELSON, R. L. Introgression of leginsulin, a cysteine-rich protein, and high-protein trait from an Asian soybean plant introduction genotype into a North American experimental soybean line. **J Agric**., v. 62, p. 2862-2869, 2015.

LEAMY, L. J.; ZHANG, H.; LI, C.; CHEN, C. Y.; SONG, B. A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). **BMC Genomics**, v. 18, 2017.

LEE, S.; VAN, K.; SUNG, M.; NELSON, R.; LAMANTIA, J.; MCHALE, L. K. Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. **Theoretical and Applied Genetics**, v. 132, p. 1639-1659, 2019.

LI, D.; ZHAO, X.; HAN, Y.; LI, W.; XIE, F. Genome-wide association mapping for seed protein and oil contents using a large panel of soybean accessions. **Genomics**, v. 111, p. 90-95, 2019.

LIPKA, A. E.; TIAN, F.; WANG, Q.; PEIFER, J.; LI, M.; BRADBURY, P. J.; GORE, M. A.; BUCKLER, E. S.; ZHANG, Z. GAPIT: genome association and production integrated tool. **Bioinformatics**, v. 28, p. 2397-2399, 2012.

MALLE, S.; MORRISON, M.: BELZILE, F. Identification of loci controlling mineral element concentration in soybean seeds. **BMC Plant Biology**, v. 20, n. 1, 1-14, 2020. https://doi.org/10.1186/s12870-020-02631-w.

MAMIDI, S.; LEE, R. K.; GOOS, J. R.; McCLEAN, P. E. Genome-wide association studies identifies seven major regions responsible for iron deficiency chlorosis in soybean (*Glycine max*). **Plos One**, 2014. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107469

McVEAN, G.; ALTSHULER, D.; DURBIN, R. et al. An integrated map of genetic variation from 1,092 human genomes. **Nature**, v. 491, p. 56–65, 2012. https://doi.org/10.1038/nature11632.

NICHOLS, D. M.; GLOVER, K. D.; CARLSON, S. R.; SPECHT, J. E.; DIERS, B. W. Fine Mapping of a seed protein QTL on soybean linkage group I and its correlated effects on agronomic traits. **Crop Science**, v. 46, p. 834–839, 2006. https://doi.org/10.2135/cropsci 2005.05-0168.

PANDURANGAN, S.; PAJAK, A.; MOLNAR, S. J.; COBER, E. R.; DHAUBHADEL, S.; HERNÁNDEZ-SEBASTIÀ, C.; KAISER, W. M.; NELSON, R. L.; HUBER, S. C.; MARSOLAIS, F. Relationship between asparagine metabolism and protein concentration in soybean seed. **Journal of Experimental Botany**, v. 63, n. 8, p. 3173-3184, 2012. https:// pubmed.ncbi.nlm.nih.gov/22357599/.

PANTHEE, D. R.; PANTALONE, V. R.; SAMS, C. E.; SAXTON, A. M.; WEST, D. R.; ORF, J. H.; KILLAM, A. S. Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. **Theoretical and Applied Genetics**, v. 112, p. 546-553, 2006.

PATIL, G.; MIAN, R.; VUONG, T.; PANTALONE, V.; SONG, Q.; CHEN, P.; SHANNON, G. J.; CARTER, T. C.; NGUYEN, H. T. Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. **Theoretical and Applied Genetics**, v. 130, p. 1975-1997, 2017.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P. I.; DALY, M. J.; SHAM, P. C. PLINK: a tool set for whole-genome association and population-based linkage analyses. **The American Journal of Human Genetics**, v. 81, p. 559575, 2007.

RINKER, K.; NELSON, R.; SPECHT, J.; SLEPER, D.; CARY, T.; CIANZIO, S. R.; CASTEEL, S.; CONLEY, S.; CHEN, P.; DAVIS, V.; FOX, C.; GRAEF, G.; GODSEY, C.; HOLSHOUSER, D.; JIANG, G.-L.; KANTARTZI, S. K.; KENWORTHY, W.; LEE, C.; MIAN, R.; MCHALE, L.; NAEVE, S.; ORF, J.; POYSA, V.; SCHAPAUGH, W.; SHANNON, G.; UNIATOWSKI, R.; WANG, D.; DIERS, B. Genetic improvement of U.S. soybean in maturity groups II, III, and IV. **Crop Science**, v. 54, p. 1419-1432, 2014.

SONAH, H.; O'DONOUGHUE, L.; COBER, E.; RAJCAN, I.; BELZILE, F. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. **Plant Biotechnology Journal**, v. 13, p. 211-221, 2014.

SONG, Q.; JENKINS, J.; JIA, G.; HYTEN, D.; PANTALONE, V.; JACKSON, S.; SCHMUTZ, J.; CREGAN, P. Construction of high-resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. **BMC Genomics**, v. 17, v. 33, 2016.

SONG, H.; YE, S.; JIANG, Y. et al. Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. **Genet Sel. Evol.**, v. 51, v. 58, 2019. https://doi.org/10.1186/s12711-019-0500-8.

STEKETEE, C. J.; SCHAPAUGH, W. T.; CARTER, T. E.; LI, Z. Genome-wide Association analyses reveal genomic regions controlling canopy wilting in soybean. **G3: Genes, Genomes, Genetics**, v. 10, n. 4, p. 1413-1425, 2020.

SZABADOS, L.; SAVOURÉ, A. Proline: a multifunctional amino acid. Trends **Plant Sci.**, v. 15, p. 89-97, 2010.

TORKAMANEH, D.; LAROCHE, J.; BASTIEN, M.; ABED, A.; BELZILE, F. Fast-GBS: a new pipeline for the efficient and highly accurate calling of SNPs from genotyping-by-sequencing data. **BMC Bioinformatics**, 2017.

TORKAMANEH, D.; BOYLE, B.; BELZILE, F. Efficient genome-wide genotyping strategies and data integration in crop plants. **Theoretical and Applied Genetics**, v. 131, p. 499-511, 2018a.

TORKAMANEH, D.; LAROCHE, J.; TARDIVEL, A.; O'DONOUGHUE, L.; COBER, E.; RAJCAN, I.; BELZILE, F. Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean. **Plant Biotechnol. J.**, v. 16, n. 3, p. 749-759, 2018b.

TORKAMANEH, D.; CHALIFOUR, F. P.; BEAUCHAMP, C. J.; AGRAMA, H.; BOAHEN, S.; MAAROUFI, H.; RAJCAN, I.; BELZILE, F. Genome-wide association analyses reveal the genetic basis of biomass accumulation under symbiotic nitrogen fixation in African soybean. **Theoretical and Applied Genetics**, v. 133, n. 2; p. 665-676, 2020.

VAN, K.; McHALE, L. K. Meta-analyses of QTLs associated with protein and oil contents and compositions in soybean [*Glycine max* (L.) Merr.) seed. **Int. J. Mol. Sci.**, v. 18, p. 1180, 2017.

van der BERG, S.; VANDENOLAS, J.; van EUEUWIJK, F. A.; BOUWMAN, A. C.; LOPES, M. S.; VEERKAMP, R. F. Imputation to whole-genome sequence using multiple pig populations and its use in genome-wide association studies. **Genet. Sel. Evol.**, v. 51, n. 2, 2019.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, p. 4414, 2008.

VAUGHN, J. N.; NELSON, R. L.; SONG, Q.; CREGAN, P. B.; LI, Z. The genetic architecture of seed composition in soybean is refined by genome wide association scans across multiple populations. **G3: Genes, Genomes, Genetics**, v. 4, p. 2283-2294, 2014.

WANG, G.; ZHANG, J.; WANG, G.; FAN, X.; SUN, X.; QIN, H. et al. Proline responding1 plays a critical role in regulating general protein synthesis and the cell cycle in maize. **Plant Cell**, v. 26, p. 2582-2600. doi: 10.1105/tpc.114.125559, 2014.

WANG, J.; ZHOU, P.; SHI, X.; YANG, N.; YAN, L.; ZHAO, Q.; YANG, C.; GUAN, Y. Primary metabolite contents are correlated with seed protein and oil traits in near-isogenic lines of soybean. **The Crop Journal**, v. 7, p. 651-659, 2019.

WARD, R. A. Genetic improvement of aphid resistance, protein, and elemental composition in soybean. (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign. Urbana, IL., 2017. https://www.ideals.illinois.edu/bitstream/handle/2142/98215/WARD-DISSERTATION-2017.pdf?sequence=1 Accessed in 07 Aug 2020.

WEI, W.; MESQUITA, A. C. O.; FIGUEIRÓ, A. A.; WU, X.; MANJUNATHA, S.; WICKLAND, D. P.; HUDSON, M. E.; JULIATTI, F. C.; CLOUGH, S. J. Genome-wide association mapping of resistance to a Brazilian isolate of *Sclerotinia sclerotiorum* in soybean genotypes mostly from Brazil. **BMC Genomics**, v. 18, n. 849, 2017.

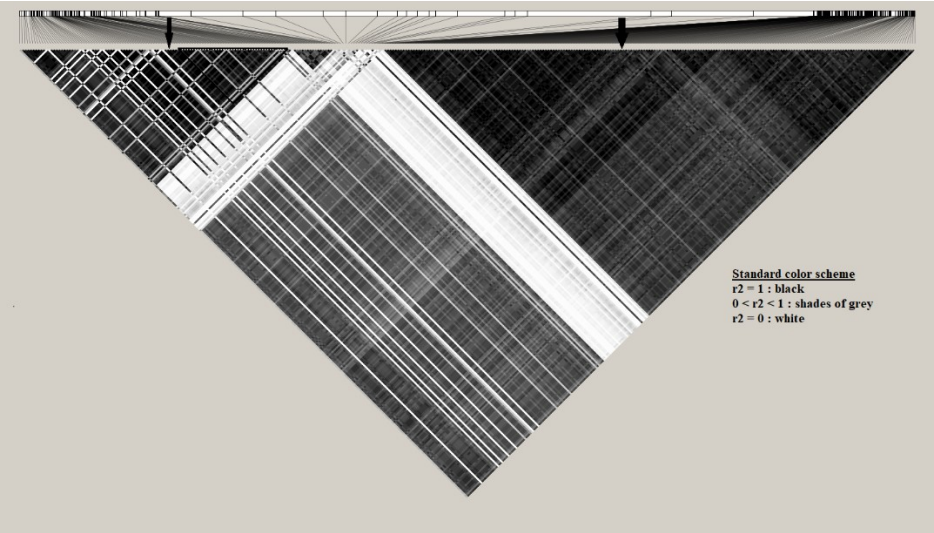
WILCOX, J. R; SHIBLES, R. M. Interrelationships among seed quality attributes in soybean. **Crop Sci.**, v. 41, p. 11-14, 2001.

WU, P.; WANG, K.; ZHOU, J.; CHEN, D.; YANG, Q.; YANG, X.; LIU, Y.; FENG, B.; JIANG, A.; SHEN, L.; XIAO, W.; JIANG, Y.; ZHU, L.; ZENG, Y.; XU, X.; LI, X.; TANG, G. GWAS on imputed whole-genome resequencing from genotyping-by-sequencing data for farrowing interval of different parities in pigs. **Front. Genet.**, v. 10, p. 1-13, 2019.

YAN, G.; QIAO, R.; ZHANG, F. et al. Imputation-based whole-genome sequence association study rediscovered the missing QTL for lumbar number in Sutai pigs. **Scientific Reports**, v. 7, n. 615, 2017. https://doi.org/10.1038/s41598-017-00729-0.

ZHANG, Y.; LU, X.; ZHAO, F. et al. Soybean GmDREBL Increases lipid content in seeds of transgenic arabidopsis. **Sci. Rep.**, v. 6, p. 34307, 2016. https://doi.org/10.1038/srep34307

5.7. SUPPLEMENTARY MATERIAL



S1Figure. The candidate regions of the major QTLs for seed protein content and seed oil content on Chr20. Plots depicting a ~ 305 Kb region (31.59 to 31.90 Mb) of Chr20 that is suspected of harboring candidate genes controlling soybean seed protein and oil, as reference: from 31.61 to 31.62 Mb corresponding a QTL region for oil content and from 31.76 to 31.87 Mb corresponding a QTL region for protein content. The figure depicts the extent of LD in this region based on r^2 (the squared allele frequency correlation between two loci). Black arrows show the position of the peak SNPs associated with oil (Gm20:31,606,891) and protein content (Gm20:31,874,531). In order to have a better picture of the linkage blocks, the marker positions were extracted from the previously mentioned catalogue with 2.18M SNPs (before pruning) instead of the pruned one (243K SNPs).

S1Table. Characteristics of the peak SNPs defining each QTL associated with seed protein content.

Chr	Peak SNP Position	QTL ID	*p*-value	FDR [a]	R^2 [b]	Effect [c]	Ref. Allele	Alt. Allele
8	47,147,391	Q_Pro_8-1	3.48E-08	2.82E-04	0.19	-2.07	T	C
10	1,609,531	Q_Pro_10-1	1.48E-07	8.27E-04	0.17	-2.28	T	C
14	58,133	Q_Pro_14-1	1.23E-07	7.28E-04	0.17	-1.37	G	T
	1,771,702	Q_Pro_14-2	5.38E-07	2.22E-03	0.15	-2.13	A	T
19	49,037,060	Q_Pro_19-1	8.55E-08	5.20E-04	0.18	-1.93	A	G
	49,914,498	Q_Pro_19-2	1.50E-08	1.92E-04	0.20	-2.16	G	C
20	31,874,531	Q_Pro_20-1	3.24E-10	5.59E-05	0.26	-2.49	G	A

[a] False Discovery Rate;
[b] R^2 was calculated by the R^2 of the model with SNP minus R^2 of model without SNP;
[c] Average change due allelic substitution.

S2Table. Characteristics of the peak SNPs defining each QTL associated with seed oil content.

Chr	Peak SNP Position	QTL ID	*p*-value	FDR [a]	R^2 [b]	Effect [c]	Ref. Allele	Alt. Allele
5	15,338,484	Q_Oil_5-1	3.02E-06	1.15E-02	0.15	1.05	G	A
8	47,343,134	Q_Oil_8-1	1.22E-07	8.24E-04	0.20	0.99	G	A
10	1,609,531	Q_Oil_10-1	1.11E-06	4.64E-03	0.16	1.03	T	C
14	1,755,977	Q_Oil_14-1	2.93E-06	1.13E-02	0.15	0.95	C	T
14	3,795,066	Q_Oil_14-2	3.27E-06	1.22E-02	0.15	0.73	A	G
19	49,037,060	Q_Oil_19-1	3.63E-07	2.05E-03	0.18	0.90	A	G
	49,914,498	Q_Oil_19-2	4.01E-07	2.17E-03	0.18	0.93	G	C
20	31,606,891	Q_Oil_20-1	1.02E-09	1.08E-04	0.27	1.44	C	T

[a] False Discovery Rate.
[b] R^2 was calculated by the R^2 of the model with SNP minus R^2 of model without SNP.
[c] Average change due allelic substitution.

S3Table. Characteristics of the peak SNPs defining each significant region associated with seed protein content, adopting the complete genotypic dataset and FDR<0.05.

Chr	Peak SNP Position	*p*-value	R^2 [a]	Significant region (flanking markers in high LD with peak SNP)		
				Start	End	Overlapping[b]
5	15,338,484	1.0E-05	0.12	15,316,474	15,410,120	X
5	32,053,693	2.6E-05	0.11	31,886,116	32,053,693	X
8	12,150,272	2.1E-05	0.11	12,150,272	12,233,480	X
8	47,147,391	3.5E-08	0.19	47,141,680	47,185,506	X
10	1,609,531	1.5E-07	0.17	1,605,797	1,611,678	X
14	58,133	1.2E-07	0.17	58,133	74,056	58.13-74.06 Kb
14	1,771,702	5.4E-07	0.15	1,771,271	1,829,944	X
16	4,099,404	1.8E-05	0.11	4,084,010	4,110,383	4.10-4.11 Mb
19	49,037,060	8.6E-08	0.18	48,991,643	49,076,176	X
19	49,914,498	1.5E-08	0.20	49,912,458	49,915,487	X
20	31,874,531	3.24E-10	0.26	31,764,481	31,874,521	31.76-31.86 Mb

[a] R^2 was calculated by the R^2 of the model with SNP minus R^2 of model without SNP.
[b] Overlapping common significant region found in GWAS work with incomplete and complete genome coverage.

S4Table. Characteristics of the peak SNPs defining each significant region associated with seed oil content, adopting the complete genotypic dataset and FDR<0.05.

Chr	Peak SNP Position	*p*-value	R^2 [a]	Significant region (flanking markers in high LD with peak SNP)		
				Start	End	Overlapping[b]
5	15,338,484	3.02E-06	0.15	15,316,474	15,410,120	X
5	31,365,336	5.06E-06	0.14	31,365,266	31,365,575	31.37-31.37 Mb
8	47,343,134	1.22E-07	0.20	47,316,318	47,365,208	X
10	1,609,531	1.11E-06	0.16	1,605,797	1,611,678	X
14	42,498	1.76E-05	0.12	42,498	74,056	58.13-74.06 Kb
14	1,755,977	2.93E-06	0.15	1,754,439	1,779,146	X
14	3,795,066	3.27E-06	0.15	3,794,998	3,796,247	X
16	4,108,747	1.85E-05	0.12	4,050,778	4,110,383	4.10-4.11 Mb
19	41,949,572	9.26E-06	0.13	41,948,686	41,955,153	X
19	49,037,060	3.63E-07	0.18	48,991,643	49,076,176	X
19	49,914,498	4.01E-07	0.18	49,912,458	49,915,487	X
20	31,606,891	1.02E-09	0.27	31,605,400	31,623,926	X
20	41,557,736	1.98E-05	0.12	41,557,736	41,583,627	X

[a] R^2 was calculated by the R^2 of the model with SNP minus R^2 of model without SNP

[b] Overlapping common significant region found in GWAS work with incomplete and complete genome coverage

S5Table. Linkage between peak SNPs identified by GWAS work with incomplete and complete genome coverage for oil content.

Chr	Peak SNP Position Incomplete coverage	Peak SNP Position Complete coverage	r^2	D'
[a]5	31,380,926	31,365,336	0.81	1.00
[a]14	65,161	42,498	0.85	0.92
[a]16	4,204,153	4,108,747	0.89	0.95

[a] Non-significant under the negative log (1/n) threshold (Table 4) but under FDR<0.05 it would be significant (S4 Table).

S6 Table. Candidate genes within significant regions for seed protein and oil content identified by GWAS with complete and incomplete coverage: location and gene ontology molecular function.

Trait	Coverage [a]	Gene	Biological Process	Molecular Function
Protein	Complete	Glyma.14G024700	proline transport	acid phosphatase activity; hydrolase activity; metal ion binding; protein serine/threonine phosphatase activity
Protein	Complete	Glyma.14G025200	protein phosphorylation	ATP binding; kinase activity; protein kinase activity; protein serine/threonine kinase activity; protein tyrosine kinase activity; transferase activity, transferring phosphorus-containing groups
Protein	Complete	Glyma.19G242200	fatty acid beta-oxidation; gluconeogenesis; glycolysis; proteasomal ubiquitin-dependent protein catabolic process; proteasome assembly; proteasome core complex assembly; proteolysis involved in cellular protein catabolic process; response to cadmium ion; response to misfolded protein; response to salt stress; toxin catabolic process; ubiquitin-dependent protein catabolic process	endopeptidase activity; peptidase activity; threonine-type endopeptidase activity
Protein	Complete	Glyma.19G242300	Golgi organization; aging; calcium ion transport; coumarin biosynthetic process; cysteine biosynthetic process; cysteine biosynthetic process from serine; gluconeogenesis; glucosinolate biosynthetic process; glycolysis; hyperosmotic response; indoleacetic acid biosynthetic process; metabolic process; photorespiration; response to cadmium ion; response to salt stress; response to temperature stimulus; response to wounding; water transport	catalytic activity; cysteine synthase activity; protein binding; pyridoxal phosphate binding
Protein	Complete	Glyma.19G242400	DNA methylation; actin nucleation; cell-cell signaling; cell adhesion; cell growth; chromatin organization; chromatin silencing; deadenylation-independent decapping of nuclear-transcribed mRNA; developmental growth; ethylene mediated signaling pathway; glucuronoxylan metabolic process; gravitropism; meristem initiation; miRNA	5'-3' exonuclease activity; 5'-3' exoribonuclease activity; exonuclease activity; nucleic acid binding; zinc ion binding

			catabolic process; nuclear-transcribed mRNA catabolic process; nuclear-transcribed mRNA catabolic process, exonucleolytic; nucleobase-containing compound metabolic process; positive regulation of cell proliferation; positive regulation of organelle organization; positive regulation of transcription, DNA-dependent; post-translational protein modification; production of miRNAs involved in gene silencing by miRNA; production of ta-siRNAs involved in RNA interference; protein N-linked glycosylation; protein glycosylation; reciprocal meiotic recombination; regulation of cell differentiation; regulation of chromosome organization; regulation of gene expression, epigenetic; regulation of transcription, DNA-dependent; reproduction; response to 1-Aminocyclopropane-1-carboxylic Acid; response to ethylene stimulus; trichome morphogenesis; unidimensional cell growth; virus induced gene silencing; xylan biosynthetic process	
Protein	Complete	Glyma.19G242700	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA); ribosome biogenesis; translation; translational elongation	nucleotide binding; structural constituent of ribosome
Protein	Complete	Glyma.19G242900	circadian rhythm; protein phosphorylation	ATP binding; kinase activity; protein kinase activity; protein serine/threonine kinase activity; protein tyrosine kinase activity; transferase activity, transferring phosphorus-containing groups
Protein	Complete	Glyma.20G085400	embryo development ending in seed dormancy; translation	structural constituent of ribosome
Protein	Incomplete	Glyma.08G350600	RNA splicing, via endonucleolytic cleavage and ligation; cytoskeleton organization; gluconeogenesis; methionine biosynthetic process; proteasomal protein catabolic process; protein sumoylation; response to heat	protein binding; protein tag
Protein	Incomplete	Glyma.08G351000	pollen germination; seed germination	acid phosphatase activity; hydrolase activity; metal ion binding; protein serine/threonine phosphatase activity

Protein	Incomplete	Glyma.08G351200	Golgi vesicle transport; acetyl-CoA metabolic process; brassinosteroid biosynthetic process; cellulose biosynthetic process; defense response signaling pathway, resistance gene-independent; protein retention in ER lumen; protein transport; sterol biosynthetic process	ER retention sequence binding; receptor activity
Protein	Incomplete	Glyma.10G014300	fatty acid beta-oxidation; photorespiration; proteolysis involved in cellular protein catabolic process; ubiquitin-dependent protein catabolic process	endopeptidase activity; peptidase activity; ribonuclease activity; threonine-type endopeptidase activity
Protein	Incomplete	Glyma.10G015400	1-aminocyclopropane-1-carboxylate biosynthetic process; asparagine catabolic process; aspartate transamidation; biosynthetic process; glutamate catabolic process to oxaloacetate	1-aminocyclopropane-1-carboxylate synthase activity; catalytic activity; pyridoxal phosphate binding; transaminase activity; transferase activity
Protein	Incomplete	Glyma.14G000800	ER to Golgi vesicle-mediated transport; MAPK cascade; N-terminal protein myristoylation; abscisic acid mediated signaling pathway; amino acid import; ammonium transport; anion transport; basic amino acid transport; cell communication; cellular membrane fusion; defense response to fungus; endoplasmic reticulum unfolded protein response; hyperosmotic salinity response; intracellular signal transduction; jasmonic acid mediated signaling pathway; negative regulation of defense response; negative regulation of programmed cell death; nucleotide transport; protein autophosphorylation; protein phosphorylation; protein targeting to membrane; regulation of anion channel activity; regulation of ion transport; regulation of plant-type hypersensitive response; regulation of stomatal movement; response to abscisic acid stimulus; response to auxin stimulus; response to cold; response to ethylene stimulus; response to jasmonic acid stimulus; response to water deprivation;	ATP binding; calcium ion binding; calmodulin-dependent protein kinase activity; kinase activity; protein binding; protein kinase activity; protein serine/threonine kinase activity; transferase activity, transferring phosphorus-containing groups

			response to wounding; signal transduction; systemic acquired resistance, salicylic acid mediated signaling pathway	
Protein	Incomplete	Glyma.14G001000	RNA splicing, via endonucleolytic cleavage and ligation; cysteine biosynthetic process; methionine biosynthetic process; response to bacterium; response to fungus; threonine biosynthetic process; threonine metabolic process	ATP binding; homoserine kinase activity
Protein	Incomplete	Glyma.14G001300	cellular response to nitrogen starvation; defense response to fungus; innate immune response; negative regulation of defense response; protein phosphorylation; regulation of defense response; response to chitin; salicylic acid biosynthetic process; salicylic acid mediated signaling pathway; systemic acquired resistance	ATP binding; kinase activity; protein kinase activity; protein serine/threonine kinase activity; protein tyrosine kinase activity; transferase activity, transferring phosphorus-containing groups
Protein	Incomplete	Glyma.19G261700	protein phosphorylation; transmembrane receptor protein tyrosine kinase signaling pathway	ATP binding; protein kinase activity; protein serine/threonine kinase activity; protein tyrosine kinase activity; transferase activity, transferring phosphorus-containing groups
Protein	Incomplete	Glyma.19G261800	protein catabolic process; ubiquitin-dependent protein catabolic process	ATP binding; ATPase activity; hydrolase activity; nucleoside-triphosphatase activity; nucleotide binding
Protein	Incomplete	Glyma.19G262500	cellular metabolic process; regulation of catalytic activity; regulation of translational initiation; translational initiation	GTP binding; guanyl-nucleotide exchange factor activity; translation initiation factor activity
Protein	Incomplete	Glyma.19G262600	metabolic process; negative regulation of catalytic activity; proteolysis	identical protein binding; serine-type endopeptidase activity
Protein	Incomplete	Glyma.19G262900	amino acid transport; lipid metabolic process; proline transport	carboxylesterase activity; hydrolase activity, acting on ester bonds; lipase activity
Oil	Complete	Glyma.19G242200	fatty acid beta-oxidation; gluconeogenesis; glycolysis; proteasomal ubiquitin-dependent protein catabolic process; proteasome assembly; proteasome core complex assembly; proteolysis involved in cellular protein catabolic process; response to cadmium ion; response to misfolded protein; response to salt stress; toxin catabolic process; ubiquitin-dependent protein catabolic process	endopeptidase activity; peptidase activity; threonine-type endopeptidase activity

Oil	Incomplete	Glyma.08G349200	fatty acid biosynthetic process	ACP phosphopantetheine attachment site binding involved in fatty acid biosynthetic process; acyl-[acyl-carrier-protein] hydrolase activity; thiolester hydrolase activity
Oil	Incomplete	Glyma.08G351200	Golgi vesicle transport; acetyl-CoA metabolic process; brassinosteroid biosynthetic process; cellulose biosynthetic process; defense response signaling pathway, resistance gene-independent; protein retention in ER lumen; protein transport; sterol biosynthetic process	ER retention sequence binding; receptor activity
Oil	Incomplete	Glyma.19G261600	lipid metabolic process	oxidoreductase activity, acting on the CH-CH group of donors
Oil	Incomplete	Glyma.19G262800	defense response to bacterium; defense response to fungus; defense response to fungus, incompatible interaction; induced systemic resistance, ethylene mediated signaling pathway; jasmonic acid and ethylene-dependent systemic resistance, ethylene mediated signaling pathway; lipid metabolic process; response to fungus; response to salicylic acid stimulus; systemic acquired resistance	carboxylesterase activity; hydrolase activity, acting on ester bonds; lipase activity
Oil	Incomplete	Glyma.19G262900	amino acid transport; lipid metabolic process; proline transport	carboxylesterase activity; hydrolase activity, acting on ester bonds; lipase activity
Oil	Incomplete	Glyma.19G263300	anther dehiscence; anther development; defense response; defense response by callose deposition; ethylene mediated signaling pathway; growth; jasmonic acid biosynthetic process; lipid oxidation; pollen development; response to bacterium; response to chitin; response to fungus; response to jasmonic acid stimulus; response to ozone; response to wounding; stamen filament development	lipoxygenase activity
Oil	Incomplete	Glyma.19G263400	N-terminal protein myristoylation; embryo development ending in seed dormancy; flower development; lipid storage; mRNA export from nucleus; meristem initiation; meristem structural organization; photomorphogenesis; protein import into nucleus; protein ubiquitination; regulation of	nucleobase-containing compound transmembrane transporter activity; tRNA binding

			flower development; response to freezing; seed dormancy process; seed germination; sugar mediated signaling pathway; tRNA export from nucleus; vegetative to reproductive phase transition of meristem	
Oil	Incomplete	Glyma.20G060100	fatty acid biosynthetic process; fatty acid metabolic process; long-chain fatty acid metabolic process	catalytic activity; long-chain fatty acid-CoA ligase activity
Oil	Incomplete	Glyma.20G085600	cellular response to phosphate starvation; lipid metabolic process	phosphatidate phosphatase activity

[a] Coverage: the candidate gene was identified by the GWAS using the genotype data with complete (243K SNPs) or incomplete (17K SNPs) marker coverage.

6. FINAL REMARKS

In the first chapter, the results indicated excellent potential of genomic predictions usage for early selection (at first stage of clonal evaluation – clonal evaluation trials – CET) in biparental cassava populations, when aiming at increasing starch production for commercial use. The correlations between the GEBVs of one-stage genomic analysis and the EBVs of the four-stage pedigree analysis were high for all traits. Moreover, the GEBV-based and EBV-based rankings of the top 10 best clones overlapped by 5 to 9 clones for individual traits, and the rankings of the 10% best clones exhibited 72% overlap for the selection index.

In the second chapter, cross-country genomic predictions proved to have potential use under the study's scenario. The joint dataset, with data from Embrapa (Brazil) and IITA (Nigeria) provided an improved accuracy compared to the prediction accuracy of either germplasm' sources individually. An assessment of population structure for the joint dataset was provided by PCA and DAPC analyses. Fixation index (F_{ST}) among the genetic groups identified within the joint dataset by DAPC analysis ranged from 0.002 to 0.091, indicating small to moderate genetic differentiation between groups. Regarding the cross-country genomic predictions, when using the marker effects estimated with IITA dataset as the training to predict the Embrapa clones' GEBVs, we could say it was a successful case of cross-country prediction, taking into consideration the medium-to-high correlation of 0.55 between their vectors of estimated GEBVs; regardless, the reverse presented a low correlation of 0.10. With the haplotype blocks analyses the model sizes were reduced, with complexity reduction rates going down to 0.20-0.22 and the correlation between the GEBVs estimated by the analyses of single markers and by haplotype blocks varied from 0.95 to 0.99.

In the third chapter, our take-home message was that the choice of using a GBS or WGS-imputed data, on GWAS, seems to rely mostly on the final-user objective, the level of resolution expected, as well as the trait of interest's genetic landscape and control, taking into consideration the pros and cons that each strategy brings. Achieving a dense catalogue of polymorphisms due integration of genotyping approaches can allow us to better exploit LD to delimit significant genomic regions, additionally to the identification of new marker-trait associated regions. However, with a very limited genotypic dataset it is still possible to successfully identify relevant regions associated to the trait of interest.