UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS, AMBIENTAIS E BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS AGRÁRIAS
CURSO DE MESTRADO

MAPEAMENTO ASSOCIATIVO E SELEÇÃO DE VARIÁVEIS RELACIONADAS AO DÉFICIT HÍDRICO EM MANDIOCA

PRISCILA PATRÍCIA DOS SANTOS SILVA

CRUZ DAS ALMAS - BAHIA
AGOSTO - 2018

MAPEAMENTO ASSOCIATIVO E SELEÇÃO DE VARIÁVEIS RELACIONADAS AO DÉFICIT HÍDRICO EM MANDIOCA

PRISCILA PATRÍCIA DOS SANTOS SILVA

Bióloga

Universidade Federal do Recôncavo da Bahia, 2016

Dissertação apresentada ao Colegiado do Programa de Pós-Graduação em Ciências Agrárias da Universidade Federal do Recôncavo da Bahia, como requisito parcial para a obtenção do Título de Mestre em Ciências Agrárias (Área de Concentração: Fitotecnia).

Orientador: Dr. Eder Jorge de Oliveira

Coorientadora: Drª. Massaine Bandeira e Sousa

CRUZ DAS ALMAS - BAHIA

AGOSTO – 2018

FICHA CATALOGRÁFICA

S586m Silva, Priscila Patrícia dos Santos.
Mapeamento associativo e seleção de variáveis relacionadas ao déficit hídrico em mandioca / Priscila Patrícia dos Santos Silva._ Cruz das Almas, BA, 2018.
120f.; il.

Orientador: Eder Jorge de Oliveira.
Coorientadora: Massaine Bandeira Cruz.

Dissertação (Mestrado) – Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas.

1.Mandioca – Melhoramento genético. 2.Mandioca – Variabilidade genética. 3.Deficiência hídrica – Análise. I.Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas. II.Título.

CDD: 633.682

Ficha elaborada pela Biblioteca Universitária de Cruz das Almas - UFRB. Responsável pela Elaboração – Antonio Marcos Sarmento das Chagas (Bibliotecário - CRB5 / 1615). Os dados para catalogação foram enviados pela usuária via formulário eletrônico.

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS, AMBIENTAIS E BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS AGRÁRIAS
CURSO DE MESTRADO

MAPEAMENTO ASSOCIATIVO E SELEÇÃO DE VARIÁVEIS RELACIONADAS AO DÉFICIT HÍDRICO EM MANDIOCA

COMISSÃO EXAMINADORA DA DEFESA DE DISSERTAÇÃO DE PRISCILA PATRÍCIA DOS SANTOS SILVA

Realizada em 09 de Agosto de 2018

Prof. Dr. Eder Jorge de Oliveira
Embrapa Mandioca e Fruticultura
Examinador Interno (Orientador)

Profa. Dra. Edna Lôbo Machado
Universidade Federal do Recôncavo da Bahia - UFRB
Examinadora Interna

Dr. Diego Fernando Marmolejo Cortes
Embrapa Mandioca e Fruticultura
Examinador Externo

DEDICATÓRIA

Dedico à minha amada família: Meu Pai Antônio (in memoriam), minha Mãe Ivanildes e a minha Avó Vanda.

AGRADECIMENTOS

Agradeço primeiramente a Deus, que em sua infinita bondade sempre me guiou e me deu forças para prosseguir, com o discernimento necessário.

À minha amada família por todo o incentivo, apoio e preocupação. Em especial a minha mãe Ivanildes, por ter me feito persistir mesmo quando achei que não fosse capaz, sem você eu não iria tão longe, essa conquista é por você e pra você!

À Fabíola, por todo incentivo, por todos os conselhos e até mesmo pelos sermões e puxões de orelha (risos), e principalmente por ser meu suporte muitas vezes. A Daniel, por toda disponibilidade, auxilio e dedicação quando precisei. E à Daiana, pelos momentos de descontração e fuga do cotidiano.

A Universidade Federal do Recôncavo da Bahia e ao Programa de Pós-graduação em Ciências Agrárias, pela oportunidade de realizar o curso de mestrado.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela concessão da bolsa.

A Empresa brasileira de pesquisa agropecuária (EMBRAPA), centro Mandioca e Fruticultura, pela concessão de uso dos laboratórios e campos experimentais para o desenvolvimento do projeto de pesquisa.

A meu orientador Eder Jorge de Oliveira, pela exigência e pelos conhecimentos compartilhados ao longo do mestrado.

À minha coorientadora Massaine Bandeira e Sousa, pelo auxílio na minha caminhada acadêmica, por toda paciência e dedicação.

Aos meus colegas do Laboratório de Biologia Molecular – Embrapa, em especial à: Ana Claudia, Jocilene, Luana, Hilçana, Andresa e o Vandeson, com vocês o trabalho se tornou mais leve e divertido, agradeço o companheirismo e por todo o auxílio e dedicação!

E a todos aqueles, mesmo que brevemente, fizeram parte desta etapa da minha vida...

...meu muito obrigado!

Faça um plano e tenha um objetivo. Trabalhe para alcança-lo, mas de vez em quando, olhe ao seu redor e aproveite, porque é isso...
Tudo pode acabar amanhã!

-Meredith Grey

Pra começar
Cada coisa em seu lugar
E nada como um dia após o outro

Pra que apressar?
Se nem sabe onde chegar
Correr em vão se o caminho é longo

E se tropeçar
Do chão não vai passar
Quem sete vezes cai, levanta oito

Quem julga saber
E esquece de aprender
Coitado de quem se interessa pouco

E quando chorar
Tristeza pra lavar
Num ombro cai metade do sufoco

O tempo dirá
O tempo é que dirá
E nada como um dia após o outro

-Tiago Iorc

Eis o meu segredo: só se vê bem com o coração. O essencial é invisível aos olhos. Os homens esqueceram essa verdade, mas tu não a deves esquecer.

- O Pequeno Príncipe

SUMÁRIO

Página

MAPEAMENTO ASSOCIATIVO E SELEÇÃO DE VARIÁVEIS RELACIONADAS AO DÉFICIT HÍDRICO EM MANDIOCA

Autora: Priscila Patrícia dos Santos Silva
Orientador: Dr. Eder Jorge de Oliveira
Coorientadora: Drª. Massaine Bandeira e Sousa

RESUMO: O presente estudo teve como objetivo a seleção de variáveis fisiológicas e agronômicas, visando estabelecer um modelo de predição para produtividade total de raízes em genótipos de mandioca sob déficit hídrico; bem como identificação de regiões genômicas associadas ao déficit hídrico via mapeamento associativo (GWAS). Foram avaliados 49 genótipos de mandioca em duas condições hídricas (irrigação – IN e déficit hídrico – DH) em dois anos agrícolas (2012/2013 e 2013/2014), utilizando seis modelos preditivos: *Classification and Regression Trees* (CART), *Artificial Neural Network* (ANN), *Support Vector Machines* (SVM), *Extreme Learning Machine* (ELM), *Generalized Linear Model with Stepwise Feature Selection* (GLMSS) e *Partial Least Squares* (PLS). Os modelos preditivos GLMSS, ELM e PLS apresentaram maior capacidade de predição ($r^2 > 0{,}75$) com *RMSE* variando entre 0,49 e 0,51. As características mais importantes para a produtividade total de raízes foram: número de raízes por planta, índice de área foliar, número de folhas mensurado no oitavo mês e produtividade da parte aérea. Em relação à GWAS, foram utilizados 25.597*single nucleotide polymorphism* (SNP), obtidos após o controle de qualidade utilizando *call rate* ≥ 0,80 e MAF<0,05. Os dados genômicos foram imputados pelo software Beagle. Foram obtidos 54 SNPs associados à produtividade total de raízes, produtividade da parte aérea, produtividade de amido, teor de matéria seca nas raízes, índice de tolerância a seca (DTI) e índice de estabilidade da tolerância a seca (DTSI). Os SNPs identificados estão próximos a 120 transcritos, previamente identificados, dos quais 24 possuem anotação funcional conhecida relacionados a algumas proteínas conhecidas, à exemplo do domínio Apetala 2 (AP2) e da proteína zíper de leucina; que estão envolvidas na tolerância ao déficit hídrico em outras espécies. Informações importantes para seleção de genótipos tolerantes ao déficit hídrico são apresentadas e discutidas para uso nos programas de melhoramento de mandioca.

Palavras-chave: tolerância à seca, modelos de predição, marcadores SNP.

GENOME-WIDE ASSOCIATION STUDY AND SELECTION OF VARIABLES RELATED TO WATER DEFICIT IN CASSAVA

Author: Priscila Patrícia dos Santos Silva
Adviser: Dr. Eder Jorge de Oliveira
Co-Adviser: Drª. Massaine Bandeira e Sousa

ABSTRACT: The present study aimed at the selection of physiological and morphological variables, aiming to establish a prediction model for root productivity in cassava genotypes under water deficit and the identification of genomic regions associated with water deficit via genome-wide association study (GWAS). 49 cassava genotypes were evaluated in two water conditions (normal irrigation - NI and water deficit - WD) in two years (2012/2013 e 2013/2014), using six predictive models: Classification and Regression Trees (CART), Artificial Neural Network (ANN), Support Vector Machines (SVM), Extreme Learning Machine (ELM), Generalized Linear Model with Stepwise Feature Selection (GLMSS) and Partial Least Squares (PLS). The predictive models GLMSS, ELM, and PLS presented higher reliability of prediction according to the values of $r^2 > 0{,}75$ with RMSE ranging from 0.49 to 0.51. The most important traits for the prediction of fresh root yield (RoY), were: number of roots (NR); area under progress curve of leaf expansion based on leaf area index (AUPC.LAI); Number of leaves measured in the eighth month (NS.8) and shoot yield (ShY). To GWAS, 25.597 single nucleotide polymorphism (SNP), with call rate ≥ 0.80 and MAF<0.05, were imputed by the software Beagle. We identified 54 marker-phenotype associated to the traits: RoY, ShY, starch yield (StY), dry matter content in the roots (DMC), drought tolerance index (DTI), and drought tolerance stability index (DTSI). The SNPs identified are close to 120 previously identified transcripts, of which 24 have known functional annotation related to some known proteins, to the example of Apetala 2 domain (AP2), Photosystem II oxygen-evolving enhancer protein; leucine zipper and bZIP transcription factor, which are involved in tolerance to water deficit in other species. This study describes important information for cassava breeding programs aimed at greater tolerance to water deficit.

Keywords: drought tolerance, prediction models, SNP markers.

REFERENCIAL TEÓRICO

Aspectos gerais da cultura e importância econômica

A mandioca (*Manihot esculenta* Crantz) pertencente à família Euphorbiaceae é originária da América tropical, sendo a região sul da bacia Amazônica do Brasil seu provável centro de origem (OLSEN & SCHAAL, 2001). Apresenta relevância mundial por ter se estabelecido como a terceira mais importante fonte de alimento, perdendo apenas para o arroz e o milho, constituindo a base alimentar de cerca de 800 milhões de pessoas na América Latina, Ásia e África (CEBALLOS et al., 2010; LIU et al., 2011; CIAT, 2017). Com a produção de 20 milhões de toneladas, o Brasil é o quarto maior produtor de mandioca, com destaque para o estado do Pará, maior produtor, seguido pelos estados do Paraná, Bahia e Maranhão (FAO, 2016; IBGE, 2017).

A cultura da mandioca possui uma vasta aplicabilidade, com aproveitamento de todas as partes da planta, podendo ser destinada para alimentação humana, animal ou utilizada como matéria-prima para uma ampla gama de produtos industriais (CEBALLOS et al., 2012). As folhas são utilizadas na alimentação humana e animal, e neste último caso, além das folhas a parte verde da haste superior é destinada a ração animal e as hastes lenhosas após serem moídas são usadas como substrato para o cultivo de cogumelos (FAO, 2013). As raízes são o principal produto e podem ser destinadas ao consumo *in natura* para uso na alimentação, seja cozida ou frita, podendo também ser raladas, fermentadas e posteriormente torradas com a finalidade de produzir uma farinha granulada, além de ser utilizada no preparo de bolos, biscoitos, pães e etc. O processamento das raízes origina o amido que possui variadas aplicações, tais como na indústria têxtil, plástica, siderúrgica, papeleira, farmacêutica, alimentícia e de biocombustíveis (CEBALLOS et al., 2012; FAO, 2013; CIAT, 2017).

A mandioca é uma cultura que apresenta adaptabilidade à escassez de água, solos degradados e alta temperatura tendo, portanto, alto potencial para ser cultivada em condições de seca, sendo produzida principalmente por pequenos agricultores com recursos limitados, tornando-se uma espécie de segurança alimentar em ambientes de cultivo marginais (LIU et al., 2011;

MORANTE et al., 2010; OKOGBENIN et al., 2013). Além disso, os pequenos agricultores realizam o cultivo da mandioca em consorcio com outras culturas à exemplo do milho (*Zea mays*), amendoim (*Arachis hypogaea*), feijão-guandu (*Cajanus cajan*), feijão comum (*Phaseolus vulgaris*) e soja (*Glycine max*), visando aumentar a produtividade, o lucro e a diversidade de produtos (PYPERS et al., 2011; FERREIRA et al., 2014).

Por outro lado, o cultivo da mandioca apresenta algumas limitações importantes tais como: a ocorrência da deterioração fisiológica pós colheita, que geralmente tem início de 24 à 48 horas após a colheita, sendo que as condições de manejo e de estocagem das raízes influenciam na sua intensidade; a incidência de pragas e doenças que são responsáveis por reduzir substancialmente os rendimentos resultando em perdas na produção; e a baixa produtividade total de raízes em consequência da tendência dos agricultores em implantar a mandioca nas piores área agrícolas, geralmente em solos degradados com baixo teor de matéria orgânica, nitrogênio, fósforo, zinco e ferro (CAMPO et al., 2011; ISLAMI et al., 2011; MORANTE et al., 2010). Apesar dessas limitações, a mandioca apresenta boa capacidade de recuperação após ataque de algumas doenças e pragas que afetam o seu cultivo, além de uma grande adaptabilidade as diversas condições edafoclimáticas. Isto é atribuído a sua ampla diversidade genética resultante da seleção natural no decorrer da evolução da cultura, na pré e pós-domesticação (HERSHEY, 1988; FUKUDA et al., 2005).

Déficit hídrico

Estresses abióticos como o hídrico (seca ou alagamento), salino, térmico, toxicidade química e estresse oxidativo são ameaças graves que afligem à agricultura, resultando na deterioração do ambiente. Sendo o principal motivo de perda de cultivos em todo o mundo (WANG et al., 2003). O déficit hídrico é um fator que tem influência direta sobre o cenário agrícola de diversos países produtores de alimentos. Além disso, por ser considerado o estresse abiótico mais significativo na agricultura, e por interferir diretamente no crescimento e desenvolvimento das plantas, ocasionando impactos negativos na produção; têm-se realizado esforços com o intuito de melhorar o

desempenho produtivo das culturas sob condições de estresse hídrico (CATTIVELLI et al., 2008).

A cultura da mandioca desenvolveu mecanismos fisiológicos complexos para suportar a escassez de água, de forma a evitar ou tolerar o estresse hídrico pelo desenvolvimento de um sistema radicular profundo, redução do crescimento, com a senescência das folhas e fechamento estomático. Contudo, o êxito desses mecanismos está sujeito à severidade e duração do período de seca (LOPES et al., 2011; TARDIEU, 2012; CLAEYS & INZE, 2013). O estresse provocado pelo déficit hídrico estimula a biossíntese, distribuição e acumulo do ácido abscísico (ABA) nos órgãos e nos tecidos das plantas, em especial nas folhas, ramos e raízes. Esse acumulo de ABA desempenha um papel importante, pois desencadeia os mecanismos fisiológicos já citados, para auxiliar a planta no processo de tolerância ao estresse hídrico (OKOGBENIN et al., 2013). Outro mecanismo de adaptação à seca desenvolvido pelas plantas é a redução do potencial hídrico devido ao acúmulo de solutos osmoticamente ativos, que propiciam a expansão celular e o seu crescimento, mesmo em situações de baixa disponibilidade hídrica, pois mantém parcialmente a abertura dos estômatos, permitindo a assimilação de CO_2 (PUGNAIRE et al., 1999). O aumento desses solutos, a exemplo dos açúcares solúveis, prolina e outros aminoácidos propicia o ajuste osmótico, conservando a integridade de enzimas, proteínas e membranas, retendo o equilíbrio de água na célula mesmo em condições ambientais desfavoráveis (BARTELS & SUNKAR, 2005).

Apesar da considerável tolerância ao déficit hídrico, a mandioca tem o seu crescimento e rendimento reduzidos por períodos de seca prolongados. A proporção da redução no rendimento das raízes depende da duração do estresse e do estágio de desenvolvimento em que a planta se encontra. O período crítico para o efeito mais severo do déficit hídrico na cultura estende-se do primeiro ao quinto mês após o plantio, que é o período correspondente à formação dos tubérculos e do rápido crescimento foliar (OKOGBENIN et al., 2013).

A mandioca possui mecanismos fisiológicos associados com a adaptação à escassez de água, para tolerar e evitar a desidratação, permitindo que a cultura suporte períodos de seca prolongados (EL-SHARKAWY, 2007).

Seus principais mecanismos de adaptação são a condutância estomática, redução do crescimento dos ramos, área foliar reduzida por meio da senescência das folhas mais velhas, surgimento de folhas novas para realização da fotossíntese e elongação da raiz principal para captação de água em áreas mais profundas do solo. Todos esses mecanismos estão relacionados principalmente com o uso limitado de carboidratos e a conservação da água, de modo a conservar esses recursos escassos (DUQUE & SETTER, 2013). Independente da intensidade do estresse hídrico há acúmulo de ABA nas folhas das plantas submetidas ao déficit hídrico, coincidindo com a manutenção do alto potencial de água nos tecidos e fechamento estomático (ALVES & SETTER, 2004; DUQUE & SETTER, 2013). Em relação ao ajustamento osmótico sob déficit hídrico, a concentração de açúcares totais no interior da célula sofre uma diminuição, expondo uma correlação negativa com o ajustamento osmótico; há um pequeno aumento na concentração de prolina, porém não é o suficiente para ocasionar alteração no ajuste osmótico, que é controlado por sais de potássio (ALVES & SETTER, 2004).

Sob restrição hídrica, o mecanismo natural de tolerância ao estresse que predomina nesta cultura é a manutenção do alto potencial hídrico nos tecidos, decorrente do fechamento estomático (ALVES & SETTER, 2004; EL-SHARKAWY, 2006). Apesar da produção de raízes ser reduzida sob déficit hídrico, a mandioca pode recuperar-se e atingir elevada capacidade após a reidratação, em virtude da rápida formação de folhas novas, com taxas fotossintéticas mais elevadas quando comparado com plantas que não sofreram o estresse, resultando em uma produção semelhante à de cultivos irrigados (EL-SHARKAWY, 2007).

Identificação de fontes de tolerância e melhoramento para mitigação do déficit hídrico em mandioca

A tolerância a seca é uma característica quantitativa complexa, decorrente da ação de vários genes, e isso implica em dificuldades no processo de seleção e melhoramento genético. E por ser um dos estresses abióticos que mais ocasiona impactos negativos na produção, o

desenvolvimento de cultivares tolerantes ao déficit hídrico é considerado uma abordagem promissora em ambientes onde a água é um fator limitante (OKOGBENIN et al., 2013).

Inicialmente a seleção de genótipos de mandioca com tolerância ao déficit hídrico baseava-se na fenotipagem dos genótipos em condições irrigadas e de sequeiro. O desenvolvimento de cultivares melhoradas com bom rendimento tanto em ambientes favoráveis quanto em ambientes sob estresse, utilizando técnicas clássicas de fenotipagem de características fisiológicas e agronômicas, como: senescência foliar tardia, elongação da raiz principal; elevada taxa fotossintética, extenso sistema de raízes finas; produção e alongamento das raízes adventícias vem sendo aplicado em diversas espécies (EL-SHARKAWY, 2007; SUBERE et al., 2009). Por outro lado, o melhoramento genético para mitigação do déficit hídrico vem sendo reforçado com o auxílio de ferramentas moleculares que podem ser utilizadas para identificar regiões genômicas associadas à tolerância à seca, permitindo a compreensão da base molecular do controle genético desta característica (PROCHNIK et al., 2012).

Atualmente, a combinação de ferramentas moleculares com bons protocolos de fenotipagem, possibilita a produção de variedades adaptadas a seca com um bom rendimento na produção tanto em ambientes favoráveis quanto em ambientes desfavoráveis (OKOGBENIN et al., 2013). De acordo com Turyagyenda et al. (2013a), a caracterização fisiológica e molecular de resposta à seca e identificação de genes de tolerância candidatos em mandioca pode ser realizada com sucesso. Assim sendo, faz-se necessário a compreensão básica dos mecanismos agronômicos e fisiológicos que estão diretamente relacionados aos altos níveis de produção mesmo que submetidos ao estresse hídrico, viabilizando desta forma, uma seleção eficiente das características fenotípicas mais estritamente relacionadas à tolerância do déficit hídrico (TURYAGYENDA et al., 2013b).

Seleção de variáveis para avaliação da tolerância ao déficit hídrico e modelos de predição

A seleção das variáveis mais importantes para explicar determinado fenótipo é utilizada com a finalidade de aprimorar o desempenho do modelo de

predição, resultando em melhores acurácias (ANDERSEN & BRO, 2010; MEHMOOD et al., 2012). Modelos de calibração multivariada são comumente aplicados para que se possa realizar a predição de um ou vários parâmetros (ANDERSEN & BRO, 2010), à exemplo do modelo *Classification and Regression Trees* (CART), *Artificial Neural Network* (ANN), *Support Vector Machines* (SVM), *Extreme Learning Machine* (ELM), *Generalized Linear Model with Stepwise Feature Selection* (GLMSS) e *Partial Least Squares* (PLS).

O CART é um modelo de classificação e regressão que desenvolve uma árvore de decisão, de modo a reproduzir graficamente a associação existente entre as variáveis preditoras e a variável a ser predita. Essa árvore de decisão é desenvolvida por intermédio de divisões binárias continuas do conjunto inicial de dados, até que os critérios predeterminados para o crescimento da árvore sejam alcançados (BREIMAN et al., 1984).

O modelo ANN utiliza algoritmos fundamentados na estrutura de neurônios biológicos, composta por uma rede com parâmetros de entrada (variáveis preditoras) e parâmetros de saída (variáveis a serem preditas) interconectadas, onde as conexões possuem diferentes pesos. Os parâmetros do modelo e os pesos das conexões são calibrados durante o treinamento do modelo, com a finalidade de capacitar a rede a produzir uma saída que melhor explique a variável predita (HAN et al., 2011).

O SVM consiste em construir um plano para separar grupos distintos, otimizando a distância de separação. Não sendo possível realizar uma separação linear, um hiperplano é criado sendo capaz de mapear os dados e separa-los linearmente. A regressão por vetores suporte é designada pelo vetor que estabelece a melhor relação dentro de uma margem ε pré-estabelecida, sendo que quanto menor for o valor de ε, maior será a complexidade do modelo gerado e consequentemente, maior será o número de vetores suporte produzidos para explicar a função (SCHÖLKOPF et al., 2000).

O ELM foi proposto inicialmente para o treinamento de redes neurais *feedforward* de camada oculta (SLFNs). No ELM os parâmetros de camada oculta são selecionados aleatoriamente e os pesos ou conexões entre a camada oculta e a camada de saída são analiticamente determinados. Desta forma, as variáveis preditoras são selecionadas de acordo com a conexão ou

peso que exercem em relação a camada oculta, até que o modelo esteja completo e possibilite a predição (HUANG et al., 2015).

Já o método GLMSS considera uma variável resposta a ser predita e variáveis explicativas, consideradas como preditoras, onde a média da variável resposta é modelada por meio de uma estrutura de regressão que envolve amostras aleatórias com n observações, que são analisadas afim de identificar quais variáveis possuem maior contribuição na predição utilizando uma função de ligação entre as variáveis preditoras e a variável a ser predita (CANTERLE & BAYER, 2015). O método *stepwise* é utilizado com o propósito de reduzir o número elevado de variáveis estudas, selecionando as variáveis que mais contribuem para o modelo de predição, reduzindo com isso o número de variáveis a compor a equação de regressão. Este método é realizado de modo interativo, adicionando (*forward selection*) ou removendo variáveis (*backward selection*), a partir do critério de informação de Akaike (AIC) (AKAIKE, 1974), que é uma medida baseada na verossimilhança de ajuste do modelo, que explica o número de parâmetros estimados, de modo que o modelo com o menor AIC possui o melhor ajuste relativo, conforme o número de parâmetros incluídos (AKAIKE, 1974; ALVES et al., 2013).

O método PLS relaciona a matriz X (variáveis dependentes) à matriz Y (composta por variáveis independentes), permitindo analisar dados com desequilíbrio entre o número de variáveis e observações; com forte correlação e com elevados níveis de ruído. Esta técnica de regressão dá origem a um conjunto de parâmetros que propiciam informações sobre a estrutura e comportamento de X e Y. A seleção do número de variáveis a serem mantidas no modelo é realizada com base na avaliação de significância em termos preditivos de cada variável, a inclusão de variáveis no modelo é interrompida quando essas variáveis deixam de ser significativas (WOLD et al., 2001; ANDERSEN & BRO, 2010).

Ainda que tenham a mesma finalidade, os modelos apresentados realizam a análise de predição de formas distintas, estando sujeitos ao conjunto de dados utilizado (PARK et al., 2005; RUß, 2009). Tendo isto em vista, para obter uma predição acurada, é necessário testar vários modelos e selecionar aqueles que melhor se ajustam ao conjunto de dados em análise.

Ou seja, ao invés de selecionar um modelo especifico para predição, geralmente é mais eficaz comparar modelos baseados em diferentes pressuposições estatísticas e à partir da análise dos parâmetros de qualidade da predição definir aquele com maior acurácia de predição (RUß, 2009).

Esses modelos têm sido utilizados na predição da produtividade agrícola (PARK et al., 2005; MUCHERINO et al., 2009; RUß, 2009; WEBER et al., 2012; MEHMOOD et al., 2012), podendo ser úteis na identificação das características mais relevantes para explicar o rendimento mesmo com baixa disponibilidade hídrica. As variáveis são selecionadas de modo que a relação entre as variáveis estudadas e a resposta tenha um alto grau de correspondência com elevada acurácia de predição (DAN et al., 2016). Como foi observado por Ferraro et al., (2009) ao realizaram um estudo com o método CART para analisar as variáveis que mais influenciam no rendimento da produtividade da cana-de-açúcar. RUß (2009) utilizou além do método CART o ANN e SVM para realizar a predição da produtividade de trigo, enquanto Bari et al., (2014) utilizaram os métodos ANN e SVM para realizar a predição de variedades de trigo resistentes à ferrugem (*Puccinia striiformis*). A maioria destes estudos apresentaram elevada acurácia de predição. Outro exemplo, foi a utilização da técnica de PLS em experimentos de campo com diferentes regimes hídricos, em que Weber et al. (2012) relataram que os modelos de predição explicaram uma maior proporção de variação genética para produtividade de grãos de milho submetidos ao estresse hídrico em comparação com condições bem irrigadas.

Marcadores moleculares aplicados no melhoramento da mandioca

No melhoramento clássico da mandioca, a seleção dos genitores para cruzamento é realizada de forma criteriosa, tendo como base características agronômicas com genótipos contrastantes. As principais técnicas utilizadas no melhoramento convencional da mandioca são a introdução e a seleção de variedades, indução de poliploides e hibridações intra e interespecíficas, que podem ser controladas ou por policruzamentos (LARA et al., 2008). Por ser uma cultura altamente heterozigótica, apresenta uma ampla segregação na primeira geração após a hibridação, podendo resultar em híbridos superiores,

que ao serem identificados, são fixados por meio da propagação vegetativa (FARIAS et al., 2006).

Além da utilização de características fenotípicas para seleção de genótipos superiores, o uso de ferramentas moleculares foi incorporado nos programas de melhoramento genético à partir da década de 90 para auxiliar no processo de seleção, onde marcadores de polimorfismos amplificados aleatoriamente (*Random Amplified Polymorphism DNA* - RAPDs) e polimorfismo de comprimento de fragmentos de restrição (*Restriction Fragment Length Polymorphism* - RFLPs) foram utilizados pela primeira vez na cultura com a finalidade de estudar a diversidade genética no gênero *Manihot* (MARMEY et al., 1994). Em seguida, os marcadores microssatélites (*Simple Sequence Repeats* - SSR) foram utilizados para a identificação de duplicatas em coleções de germoplasma (CHAVARRIAGA-AGUIRRE et al., 1999), e marcadores associados ao polimorfismo de comprimento de fragmentos amplificados (*Amplified Fragment Length Polymorphism* - AFLP) foram utilizados na compreensão da diferenciação genética da mandioca (ROA et al., 1997; ELIAS et al., 2000; FREGENE et al., 2000). Além disso, marcadores AFLP, isoenzimas, e RAPD foram utilizados para construção do primeiro mapa de ligação genética da cultura (FREGENE et al., 1997). Posteriormente diversos estudos com marcadores moleculares na cultura foram utilizados para diferentes finalidades, como análise da diversidade genética (ASANTE & OFFEI, 2003); detecção e quantificação de doenças virais (MONGER et al., 2001); diagnóstico de doenças como couro de sapo (ALVAREZ et al., 2009); dentre outras.

Os SNPs (*single nucleotide polymorphism*) são marcadores moleculares que têm como base as alterações da molécula de DNA, ou seja, mutações em bases únicas da cadeia de bases nitrogenadas. São marcadores binários e codominantes, capazes de discriminar eficientemente os alelos homozigóticos e heterozigóticos (BROOKES, 1999). Esses marcadores fornecem uma elevada cobertura do genoma, e quando encontrados em sequências codificadoras podem resultar em alterações no fenótipo, e assim, expor associação a características especificas (JEHAN & LAKHANPAUL, 2006). Na mandioca, diversos estudos já fizeram uso dos marcadores SNPs em estudos

de diversidade; para construção de mapa de ligação genético; como meio de rastrear as origens evolutivas e geográficas da mandioca; na identificação de genótipos com amido ceroso relacionados à expressão do gene GBSSI, dentre outros (OLSEN, 2004; RABBI et al., 2012; AIEMNAKA et al., 2012).

Os marcadores SNPs têm se destacado em virtude das recentes tecnologias desenvolvidas para genotipagem, que envolvem o processamento de um grande número de amostras e marcadores simultaneamente, reduzindo assim os custos e proporcionado maior rendimento e reprodutibilidade em comparação outras estratégias de genotipagem (JEHAN & LAKHANPAUL, 2006; MAMMADOV et al., 2012). Com o desenvolvimento do sequenciamento genômico completo da mandioca associado ao uso de ferramentas de bioinformática, tem sido possível desenvolver novas ferramentas moleculares com ampla cobertura genômica, para facilitar e consolidar as pesquisas para o melhoramento genético da cultura (CEBALLOS et al., 2012; PROCHNIK et al., 2012). O advento do sequenciamento de DNA teve início com o método Sanger (SANGER & COULSON, 1975). Posteriormente com a automatização do método, está técnica se tornou a principal ferramenta de sequenciamento do genoma humano e de outros organismos (LANDER et al., 2001; GARCIA-HERNANDEZ et al., 2002).

Em 2005 surgiu um novo método de sequenciamento de DNA, denominado sequenciamento de nova geração, ou NGS (*Next Generation Sequencing*). Inicialmente foi lançado o sequenciador 454 da Roche, que realiza o sequenciamento por síntese, onde a leitura das bases ocorre à medida que é adicionada ao fragmento de DNA recém-formado, diferindo assim do método Sanger, onde a leitura da base era realizada por marcação fluorescente e pelo peso molecular da sequência parcial do DNA contido na molécula, por meio da análise eletroforética (MARGULIES et al., 2005; SCHUSTER, 2008). Subsequentemente outras plataformas de sequenciamento foram criadas fazendo uso de diferentes técnicas de sequenciamento, apesar de divergirem em alguns aspectos metodológicos, todos os sequenciadores de nova geração (NGS) se baseiam no processamento massivo de fragmentos de DNA.

A genotipagem por sequenciamento (*Genotyping-by-sequencing* - GBS) tem como base o sequenciamento de nova geração (NGS), com o diferencial de conduzir simultaneamente a genotipagem e a detecção de polimorfismos (ELSHIRE et al., 2011). Atualmente a técnica de GBS tem sido bastante utilizada por propiciar a geração de um grande volume de dados de SNPs de forma rápida e com um custo relativamente baixo. A GBS utiliza a enzima de restrição *Ape*KI para efetuar a digestão do DNA genômico alvo; esta enzima é uma endonuclease de restrição do tipo II que reconhece uma sequência degenerada de cinco bases (GCWGC, onde W pode ser A ou T), originando fragmentos com tamanhos de 200 à 400pb (pares de base), com uma extremidade 5' (3pb). Essa extremidade, é reconhecida pelo adaptador, propiciando a ligação deste com o DNA genômico cortado pela enzima, reduzindo assim a complexidade do genoma de modo rápido e fácil, promovendo o sequenciamento das extremidades de todos os fragmentos decorrentes da digestão. Em virtude disto, GBS é um método relativamente simples de genotipagem, apropriada para caracterização de germoplasma, estudos populacionais e de melhoramento genético em diversos organismos (ELSHIRE et al., 2011). Na cultura da mandioca, vários estudos foram realizados utilizando marcadores SNPs obtidos por GBS, como por exemplo análise de estrutura populacional e identificação de variedades de mandioca (RABBI et al., 2015); construção de mapas genéticos com a localização genética e física de genes relacionados à imunidade à patógenos (SOTO et al., 2015); e mapeamento genético visando a identificação de regiões relacionadas ao aumento do teor de carotenoides nas raízes de armazenamento; a resistência ao vírus do mosaico africano da mandioca (CMD) e rendimento em áreas com alta pressão de CMD (RABBI et al., 2014).

Os marcadores moleculares têm sido utilizados amplamente com a finalidade de realizar o mapeamento de QTL (*Quantitative trait loci*), em muitas culturas. Os QTLs são regiões do genoma responsáveis pela expressão de caracteres fenotípicos. Na mandioca já foram identificados QTLs que controlam a acumulação de glucosídeos cianogênicos e teor de matéria seca (KIZITO et al., 2007); teores de caroteno (MARÍN et al., 2009) e tolerância à deterioração pós-colheita (CORTÉS et al., 2002). SNPs também já foram utilizados para o

mapeamento de QTL que controlam as propriedades de viscosidade do amido e resistência a pragas e doenças, tais como: à bacteriose; vírus do mosaico africano; vírus do castanho listrado da mandioca (CBSD) e o ácaro da mandioca (RABBI et al., 2014; NZUKI et al., 2017). Porém, em alguns casos de mapeamento de QTL na mandioca, houve uma baixa resolução de marcadores ligados a QTL, decorrente do uso de um número limitado de marcadores e de um reduzido número de amostras. Para análise mais abrangente da arquitetura genética se faz necessário o uso de populações múltiplas que representem uma amostra maior da variação genética disponível nas espécies. Uma alternativa viável é o uso de uma nova abordagem, utilizando mapeamento associativo via mapeamento de QTL ou associação genômica ampla (GWAS) (HOLLAND, 2007; FERGUSON, et al., 2012).

Associação genômica ampla (GWAS)

A GWAS tem sido utilizada para determinar a base genética de características complexas (ABDURAKHMONOV et al., 2008), tendo como base o desequilíbrio de ligação (LD), que é a associação não aleatória de alelos em loci diferentes e é afetada por padrões de recombinação do genoma, pela estrutura da população e pelo sistema de reprodução da espécie. Desta forma, o mapeamento de QTLs (*Quantitative Trait Loci*) via GWAS, pode estimar a localização e o número de genes que possuem controle sobre a variação fenotípica de uma caraterística, por meio da associação do genótipo com o fenótipo (FLINT-GARCIA et al., 2003; GUIMARÃES et al., 2009; MYLES et al., 2009). Uma das vantagens é que na procura por uma associação entre um marcador e uma determinada característica de interesse, o efeito do ambiente é reduzido (ABDURAKHMONOV et al., 2008). Na mandioca, a GWAS já foi realizada com êxito para mapeamento de características agronômicas como: conteúdo de carotenoides pró-vitamínicos A (ESUMA et al., 2016); resistência à podridão radicular (BRITO et al., 2017); e para severidade dos sintomas foliares e nas raízes causados pelo CBSD (KAYONDO et al., 2018).

A GWAS tem se destacado por possuir vantagens como a alto poder de resolução, eficiência de mapeamento e baixo custo em relação a outras estraetgias de mapeamento genético. Além de ser uma poderosa ferramenta

para a identificação de variações genéticas que compõem a base de fenótipos de elevada importância, tais como a resistência a doenças e a tolerância à seca (IQUIRA et al., 2015). Informações sobre QTLs que regulam a resposta ao déficit hídrico e genes responsáveis por efeitos de QTL podem ser utilizados para elucidar a base fisiológica da tolerância à seca e na seleção de genótipos com maior produtividade em condições limitadas de água (TUBEROSA & SALVI, 2006). Em grandes culturas como a cevada, o milho e a soja, o método de mapeamento associativo (GWAS) tem sido utilizado com sucesso na identificação de loci controladores de caracteres quantitativos (QTL), por meio da análise de ligação entre SNPs e regiões cromossômicas associadas com a tolerância ao estresse hídrico (DESHMUKH et al., 2014; WEHNER et al., 2015; WANG et al., 2016). Entretanto, o mapeamento de QTL via GWAS para tolerância ao déficit hídrico tem sido pouco explorado na cultura da mandioca.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDURAKHMONOV, I.Y.; ADDUKARIMOV, A. Application of association mapping to understanding the genetic diversity of plant germplasm resources. **International Journal of Plant Genomics**, v.2008, n.1, p.1-18, 2008.

AIEMNAKA, P.; WONGKAEW, A.; CHANTHAWORN, J.; NAGASHIMA, S.K.; BOONMA, S.; et al. Molecular characterization of a spontaneous *waxy* starch mutation in cassava. **Crop Science**, v. 52, p.2121-2130, 2012.

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, v.19, n.6, p.716-723. 1974.

ALVAREZ, E.; MEJÍA, J.F.; LLANO, G.A.; LOKE, J.B.; CALARI, A.; et al. Characterization of a phytoplasma associated with frogskin disease in cassava. **Plant Disease**, v.93, n.11, p.1139-1145, 2009.

ALVES, A.A.C.; SETTER, T.L. Abscisic acid accumulation and osmotic adjustment in cassava under water deficit. **Environmental and Experimental Botany**, v.51, n.3, p.259-271, 2004.

ALVES, M.F.; LOTUFO, A.D.P.; LOPES, M.L.M. Seleção de variáveis *stepwise* aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v.1, n.1, p. 010144.1-010144.6. 2013.

ANDERSEN, C.M.; BRO, R. Variable selection in regression - A Tutorial. **Journal of Chemometrics**, v.24, n.11-12, p.728-737, 2010.

ASANTE, I.K.; OFFEI, S.K. RAPD-based genetic diversity study of fifty cassava (*Manihot esculenta* Crantz) genotypes. **Euphytica**, v.131, n.1, p.113-119, 2003.

BARI, A.; AMRI, A.; STREET, K.; MACKAY, M.; DE PAUW, E.; et al. Predicting resistance to stripe (yellow) rust (*Puccinia Striiformis*) in wheat genetic resources using focused identification of germplasm strategy. **The Journal of Agricultural Science**, v.152, n.6, p.906-916, 2014.

BARTELS, D.; SUNKAR, R. Drought and salt tolerance in plants. **Critical Reviews in Plant Sciences**, v.24, n.1, p.23-58, 2005.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. Classification and regression trees. Calif: **Wadsworth International,** v.543, p.577, 1984.

BRITO, A.C.; OLIVEIRA, S.A.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.

BROOKES, A.J. The essence of SNPs. **Gene**, v.234, n.2, p.177-186, 1999.

CAMPO, B.V.H.; HYMAN, G.; BELLOTTI, A. Threats to cassava production: known and potential geographic distribution of four key biotic constraints. **Food Security**, v.3, n.3, p.346, 2011.

CANTERLE, D.R.; BAYER, F.M. Testes de especificação para a função de ligação em modelos lineares generalizados para dados binários. **Ciência e Natura**, v.37, n.1, p.1-11, 2015.

CATTIVELLI, L.; RIZZA, F.; BADECK, F.W.; MAZZUCOTELLI, E.; MASTRANGELO, A.M.; FRANCIA, E.; MARÈ, C.; TONDELLI, A.; STANCA, A. M. Drought tolerance improvement in crop plants: an integrated view from breeding to genomics. **Field Crops Research**, v.105, n.1, p.1-14, 2008.

CEBALLOS, H.; OKOGBENIN, E.; PÉREZ, J.C.; LÓPEZ-VALLE, L.A.B.; DEBOUCK, D. Cassava. In: Root and tuber crops. **Springer**, v.1, p.53-96, 2010.

CEBALLOS, H.; KULAKOW, P.; HERSHEY, C.; Cassava breeding: current status, bottlenecks and the potential of biotechnology tools. **Tropical Plant Biology**, v.5, p.73-87, 2012.

CHAVARRIAGA-AGUIRRE, P.; MAYA, M.M.; TOHME, J.; DUQUE, M.C.; IGLESIAS, C.; et al. Using Microsatellites, isozymes and AFLP to evaluate genetic diversity and redundancy in the cassava core collection and to assess the usefulness of DNA-based markers to maintain germplasm collections. **Molecular Breeding**, v.5, p.263-273, 1999.

CLAEYS, H.; INZÉ, Dirk. The agony of choice: how plants balance growth and survival under water-limiting conditions. **Plant Physiology**, v.162, n.4, p.1768-1779. 2013.

CIAT. **International Center for Tropical Agriculture**. 2017. Disponível em: <http://ciat.cgiar.org/what-we-do/breeding-better-crops/rooting-for-cassava/>. Acesso em: dez, 2017.

CORTÉS, D.; REILLY K.; OKOGBENIN J.; BEECHING J.R.; IGLESIAS C.; TOHME J. Mapping wound-response genes involved in post-harvest physiological deterioration (PPD) of cassava (*Manhiot esculenta* Crantz). **Euphytica**, v.128, n.1, p.47–53, 2002.

DAN, Z.; HU, J.; ZHOU, W.; YAO, G.; ZHU, R.; ZHU, Y.; HUANG, W. Metabolic prediction of important agronomic traits in hybrid rice (*Oryza sativa* L.). **Scientific Reports**, v.6, p.1-9, 2016.

DESHMUKH, R.; SONAH, H.; PATIL, G.; CHEN, W.; PRINCE, S.; et al. Integrating omic approaches for abiotic stress tolerance in soybean**. Frontiers in Plant Science**, v.5, p.244-253, 2014.

DUQUE, L.O.; SETTER, T.L. Cassava response to water deficit in deep pots: root and shoot growth, ABA, and carbohydrate reserves in stems, leaves and storage roots. **Tropical Plant Biology**, v.6, n.4, p.199-209, 2013.

ELIAS, M.; PANAUD, O.; ROBERT, T. Assessment of genetic variability in a traditional cassava (*Manihot esculenta* Crantz) farming system, using AFLP markers. **Heredity**, v.85, n.3, p.219-230, 2000.

EL-SHARKAWY, M.A. International research on cassava photosynthesis, productivity, eco-physiology, and responses to environmental stresses in the Tropics**. Photosynthetica**, v.44, n.4, p.481-512, 2006.

EL-SHARKAWY, M.A. Physiological characteristics of cassava tolerance to prolonged drought in the tropics: implications for breeding cultivars adapted to seasonally dry and semiarid environments. **Journal of Plant Physiology**, v.19, p.257-286, 2007.

ELSHIRE, R.J.; GLAUBITZ, J.C.; SUN, Q.; POLAND, J.A.; KAWAMOTO, K.; et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PloS One**, v.6, n.5, p.19379-19389, 2011.

ESUMA, W.; HERSELMAN, L.; LABUSCHAGNE, M.T.; RAMU, P.; LU, F.; et al. Genome-wide association mapping of provitamin A carotenoid content in cassava. **Euphytica**, v.212, n.1, p.97-110, 2016.

FAO. **Food and Agriculture Organization of the United Nations**. Save and Grow: Cassava a Guide to Sustainable Production Intensification. 2013. Disponível em: <http://www.fao.org/3/a-i2929o.pdf>. Acesso em: dez, 2017.

FAO. **Food and Agriculture Organization of the United Nations**. Food Outlook: Biannual Report on Global Food Markets. 2016. Disponível em: <http://www.fao.org/3/a-i6198e.pdf>. Acesso em: dez, 2017.

FARIAS, A.R.N.; SOUZA, L.S.; MATTOS, P.L.P.; FUKUDA, W.M.G. Aspectos socioeconômicos e agronômicos da mandioca. **Embrapa Informação Tecnológica**. v.1 p.326-354, 2006

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v.27, n.8, p.861-874, 2006.

FERGUSON, M.; RABBI, I.; KIM, D.J.; GEDIL, M.; LOPEZ-LAVALLE, L.A.B.; OKOGBENIN, E. Molecular markers and their application to cassava breeding: past, present and future. **Tropical Plant Biology**, v.5, n.1, p.95-109, 2012.

FERRARO, D.O.; RIVERO, D.E.; GHERSA, C.M. An analysis of the factors that influence sugarcane yield in northern Argentina using classification and regression trees. **Field Crops Research**, v.112, n.2, p.149-157, 2009.

FERREIRA, E.A.; SILVA, D.V.; BRAGA, R.R.; DE OLIVEIRA, M.C.; PEREIRA, G.A.M.; et al. Crescimento inicial da cultura da mandioca em sistema de policultivo. **Scientia Agraria Paranaensis**, v.13, n.3, p.219-226, 2014.

FLINT-GARCIA, S.A.; THORNSBERRY, J.M.; BUCKLER, E.S. Structure of linkage disequilibrium in plants. **Annual Review on Plant Biology**, v.54, p.357-374, 2003.

FREGENE, M.; ANGEL, F.; GOMEZ, R.; RODRIGUEZ, F.; CHAVARRIAGA, P.; et al. A molecular genetic map of cassava (*Manihot esculenta* Crantz). **Theoretical and Applied Genetics**, v.95, n.3, p.431-441, 1997.

FREGENE, M.; BERNAL, A.; DUQUE, M.; DIXON, A.; TOHME, J. AFLP Analysis of African cassava (*Manihot esculenta* Crantz) germplasm resistant to the cassava mosaic disease (CMD). **Theoretical and Applied Genetics**, v.100, p.678-685, 2000.

FUKUDA, W.M.G.; OLIVEIRA, R.D.; FIALHO, J.D.F.; CAVALCANTI, J.; CARDOSO, E.M.R.; et al. Germoplasma de mandioca (*Manihot esculenta* Crantz) no Brasil. **Revista Brasileira de Mandioca**, v.18, p.7-12, 2005.

GARCIA-HERNANDEZ, M.; BERARDINI, T.Z.; CHEN, G.; CRIST, D.; DOYLE, A.; et al. TAIR: a resource for integrated *Arabidopsis* data. **Functional & Integrative Genomics**, v.2, n.6, p.239–253, 2002.

GUIMARÃES, C.T.; MAGALHÃES, J.V.; LANZA, M.A.; SCHUSTER, I. Marcadores moleculares e suas aplicações no melhoramento genético. **Informe Agropecuário**, v.30, n.253, 2009.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. Elsevier, 2011.

HERSHEY, C.H. Cassava breeding-CIAT Headquarters. **Cassava Breeding and Agronomy Research in Asia**, v.1, p.99-116, 1988.

HOLLAND, J.B. Genetic architecture of complex traits in plants. **Current Opinion in Plant Biology**, v.10, n.2, p.156-161, 2007.

HUANG, G.; HUANG, G.B.; SONG, S.; YOU, K. Trends in extreme learning machines: A review. **Neural Networks**, v.61, p.32-48, 2015.

IBGE. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção Agrícola**. Disponível em: <ftp://ftp.ibge.gov.br/Producao_Agricola/Fasciculo_Indicadores_IBGE/estProdAgr_201704.pdf > Acesso em: dez, 2017.

IQUIRA, E.; HUMIRA, S.; FRANÇOIS, B. 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**, v.15, n.1, p.1-12, 2015.

ISLAMI, T.; GURITNO, B.; UTOMO, W.H. Performance of cassava (*Manihot esculenta* Crantz) based cropping systems and associated soil quality changes in the degraded tropical uplands of East Java, Indonesia. **Journal of Tropical Agriculture**, v.49, p.31-39, 2011.

JEHAN, T.; LAKHANPAUL, S. Single nucleotide polymorphism (SNP) methods and applications in plant genetics: a review. **Indian Journal of Biotechnology**, v.5, p.435-459, 2006.

KAYONDO, S.I.; DEL CARPIO, D.P.; LOZANO, R.; OZIMATI, A.; WOLFE, M.D.; et al. Genome-wide association mapping and genomic prediction for CBSD resistance in *Manihot esculenta*. **Scientific Reports**, v.8, n.1, p.1549-1583, 2018.

KIZITO, B.E.; RÖNNBERG-WÄSTLJUNG, A.C.; EGWANG, T.; GULLBERG, U.; FREGENE, M.; WESTERBERGH, A. Quantitative trait loci controlling cyanogenic glucoside and dry matter content in cassava (*Manihot esculenta* Crantz) roots. **Hereditas**, v.144, n.4, p.129-136, 2007.

LANDER, E.S.; LINTON, L.M.; BIRREN, B.; NUSBAUM, C.; ZODY, M.C.; et al. Initial sequencing and analysis of the human genome. **Nature**, v.409, n.6822, p.860-921, 2001.

LARA, A.C.C.; BICUDO, S.J.; BRACHTVOGEL, E.L.; DE ABREU, M.L.; CURCELLI, F. Melhoramento genético da cultura da mandioca (*Manihot esculenta* Crantz). **Revista Raízes e Amidos Tropicais**, v.4, n.1, p.54-64, 2008.

LIU, J.; ZHENG, Q.; MA, Q.; GADIDASU, K.K.; ZHANG, P. Cassava genetic transformation and its application in breeding. **Journal of Integrative Plant Biology**, v.53, n.7, p.552-569, 2011.

LOPES, M.S.; ARAUS, J.L.; VAN HEERDEN, P.D.; FOYER, C.H. Enhancing drought tolerance in C4 crops. **Journal of Experimental Botany**, v.62, n.9, p.3135-3153. 2011.

MAMMADOV, J.; AGGARWAL, R.; BUYYARAPU, R.; KUMPATLA, S. SNP markers and their impact on plant breeding. **International journal of plant genomics**, v.2012, p.387-398, 2012.

MARGULIES, M.; EGHOLM M.; ALTMAN, W.E.; ATTIYA, S.; BADE, J.S.; et al. Genome sequencing in open microfabricated high density picoliter reactors. **Nature**, v.437, n.7057, p.376-380, 2005.

MARÍN, C.J.A.; RAMÍREZ, H.; FREGENE, M. Genetic mapping and QTL analysis for carotenes in a S_1 population of cassava. **Acta Agronómica**, v.58, n.1, p.15-21, 2009.

MARMEY, P.; BEECHING, J.; HAMON, S.; CHARRIER, A. Evaluation of cassava (*Manihot esculenta* Crantz.) germplasm using RAPD markers. **Euphytica**, v.74, p.203-209, 1994.

MEHMOOD, T.; LILAND, K. H.; SNIPEN, L.; SOLVE, S. A Review of variable selection methods in partial least squares regression. **Chemometrics and Intelligent Laboratory Systems**, v.118, p.62-69, 2012.

MONGER, W.A.; SEAL, S.; COTTON, S.; FOSTER, G.D. Identification of different isolates of cassava brown streak virus and development of a diagnostic test. **Plant Pathology**, v.50, n.6, p.768-775, 2001.

MORANTE, N.; SÁNCHEZ, T.; CEBALLOS, H.; CALLE, F.; PÉREZ, J.C.; et al. Tolerance to postharvest physiological deterioration in cassava roots. **Crop Science**, v.50, n.4, p.1333-1338, 2010.

MUCHERINO, A.; PAPAJORGJI, P.; PARDALOS, P.M. A Survey of data mining techniques applied to agriculture. **Operational Research**, v.9, n.2, p.121-140, 2009.

MYLES, S.; PEIFFER, J.; BROWN, P.J.; ERSOZ, E.S.; ZHANG, Z.; COSTICH, D.E.; BUCKLER, E.S. Association mapping: critical considerations shift from genotyping to experimental design. **The Plant Cell**, v.21, n.8, p.2194-2202, 2009.

NZUKI, I.; KATARI, M.S.; BREDESON, J.V.; MASUMBA, E.; KAPINGA, F. et al. QTL mapping for pest and disease resistance in cassava and coincidence of some QTL with introgression regions derived from *Manihot glaziovii*. **Frontiers in Plant Science**, v.8, p.1168-1183, 2017.

OKOGBENIN, E.; SETTER, T.L.; FERGUSON, M.; MUTEGI, R.; CEBALLOS, H.; OLASANMI, B.; FREGENE, M. Phenotypic approaches to drought in cassava: review. **Frontiers in Physiology**, v.4, p.1-15, 2013.

OLSEN, K.M.; SCHAAL, B.A. Microsatellite variation in cassava (*Manihot esculenta*, Euphorbiaceae) and its wild relatives: further evidence for a Southern Amazonian origin of domestication. **American Journal of Botany**, v.88, p.131-142, 2001.

OLSEN, K.M. SNPs, SSRs and inferences on cassava's origin. **Plant Molecular Biology**, v.56, n.4, p.517-526, 2004.

PARK, S.J.; HWANG, C.S.; VLEK, P.L.G. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. **Agricultural Systems**, v.85, n.1, p.59–81, 2005.

PYPERS, P.; SANGINGA, J.M.; KASEREKA, B.; WALANGULULU, M.; VANLAUWE, B. Increased productivity through integrated soil fertility management in cassava–legume intercropping systems in the highlands of Sud-Kivu, DR Congo. **Field Crops Research**, v.120, n.1, p.76-85, 2011.

PROCHNIK, S.; MARRI, P.R.; DESANY, B.; RABINOWICZ, P.D.; KODIRA, C.; et al. The cassava genome: current progress, future directions. **Tropical Plant Biology**, v.5, n.1, p.88-94, 2012.

PUGNAIRE, F.I.; ENDOLZ, L.S.; PARDOS, J. Constraints by water stress on plant growth. **Handbook of Plant and Crop Stress**. v.2, p.271–283, 1999.

RABBI, I.Y.; KULEMBEKA, H.P.; MASUMBA, E.; MARRI, P.R.; FERGUSON, M. An EST-derived SNP and SSR genetic linkage map of cassava (*Manihot esculenta* Crantz). **Theoretical and Applied Genetics**, v.125, n.2, p.329-342, 2012.

RABBI, I.Y.; HAMBLIN, M.T.; KUMAR, P.L. GEDIL, M.A.; IKPAN, A.S.; JANNINK, J.L.; KULAKOW, P.A. High-resolution mapping of resistance to cassava mosaic geminiviruses in cassava using genotyping-by-sequencing and its implications for breeding. **Virus Research**, v.186, p.87-96, 2014.

RABBI, I.Y.; KULAKOW, P.A.; MANU-ADUENING, J.A.; DANKYI, A.A.; ASIBUO, J.Y.; et al. Tracking crop varieties using genotyping-by-sequencing markers: a case study using cassava (*Manihot esculenta* Crantz). **BMC Genetics**, v.16, n.1, p.115-126, 2015.

ROA, A.C.; MAYA, M.M.; DUQUE, M.C.; TOHME. J.; ALLEM, A.C.; BONIERBALE, M.W. AFLP Analysis of relationships among cassava and other *Manihot* species. **Theoretical and Applied Genetics**, v95, p.741-750, 1997.

RUß, G. Data mining of agricultural yield data: a comparison of regression models. **Industrial Conference on Data Mining**, v. 5633, p.24–37, 2009.

SANGER, F.; COULSON, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. **Journal of Molecular Biology**, v.94, n.3, p.441-448, 1975.

SCHÖLKOPF, B.; SMOLA, A.J.; WILLIAMSON, R.C.; BARTLETT, P.L. New support vector algorithms. **Neural Computation**, v.12, n.5, p.1207-1245, 2000.

SCHUSTER, S.C. Next-generation sequencing transforms today's biology. **Nature Methods**, v.5, n.1, p.16-20, 2008.

SOTO, J.C.; ORTIZ, J.F.; PERLAZA-JIMÉNEZ, L.; VÁSQUEZ, A.X.; LOPEZ-LAVALLE, L.A.B.; et al. A genetic map of cassava (*Manihot esculenta* Crantz) with integrated physical mapping of immunity-related genes. **BMC Genomics**, v.16, n.1, p.190-212, 2015.

SUBERE, J.O.Q.; BOLATETE, D.; BERGANTIN, R.; PARDALES, A.; BELMONTE, J.J.; MARISCAL, A.; SEBIDOS, R.; YAMAUCHI, A. Genotypic variation in responses of cassava (*Manihot esculenta* Crantz) to drought and rewatering: root system development. **Plant Production Science**, v.12, n.4, p.462-474, 2009.

TARDIEU, F. Any trait or trait-related allele can confer drought tolerance: just design the right drought scenario. **Journal of Experimental Botany**, v.63, n.1, p.25-31. 2012.

TUBEROSA, R.; SALVI, S. Genomics-based approaches to improve drought tolerance of crops. **Trends in Plant Science**, v.11, n.8, p.405-412, 2006.

TURYAGYENDA, L.F.; KIZITO, E.B.; FERGUSON, M.; BAGUMA, Y.; AGABA, M.; et al. Physiological and molecular characterization of drought responses and identification of candidate tolerance genes in cassava. **AoB Plants**, v.5, p.7-24, 2013a.

TURYAGYENDA, L.F.; KIZITO, E.B.; BAGUMA, Y.; OSIRU, D. Evaluation of Ugandan cassava germplasm for drought tolerance. **International Journal of Agriculture and Crop Sciences**, v.5, n.3, p.212-226, 2013b.

WANG, W.; VINOCUR, B.; ALTMAN, A. Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. **Planta**, v.218, n.1, p.1-14, 2003.

WANG, N.; WANG, Z.P.; LIANG, X.L.; WENG, J.F.; LV, X.L.; et al. Identification of loci contributing to maize drought tolerance in a genome-wide association study. **Euphytica**, v.210, n.2 p.165-179, 2016.

WEBER, V. S.; ARAUS, J. L.; CAIRNS, J. E.; SANCHEZ, C.; MELCHINGER, A. E.; ORSINI, E. Prediction of grain yield using reflectance spectra of canopy and

leaves in maize plants grown under different water regimes. **Field Crops Research**, v.128, p.82-90, 2012.

WEHNER, G.G.; BALKO, C.C.; ENDERS, M.M.; HUMBECK, K.K.; ORDON, F.F. Identification of genomic regions involved in tolerance to drought stress and drought stress induced leaf senescence in juvenile barley. **BMC Plant Biology**, v.15, n.1, p.125-140, 2015.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v.58, n.2, p.109-130, 2001.

ARTIGO 1

MODELOS DE PREDIÇÃO E SELEÇÃO DE CARACTERÍSTICAS AGRONÔMICAS E FISIOLÓGICAS PARA TOLERÂNCIA AO DÉFICIT HÍDRICO EM MANDIOCA

[1]Artigo a ser ajustado para posterior submissão ao Comitê Editorial do periódico científico Crop Science, em versão na língua inglesa.

Modelos de predição e seleção de características agronômicas e fisiológicas para tolerância ao déficit hídrico em mandioca

Resumo: O desenvolvimento de estratégias eficientes e acuradas de avaliação e predição do comportamento produtivo dos genótipos de mandioca (*Manihot esculenta* Crantz), pode reduzir o esforço gasto na fenotipagem de característica complexas como a tolerância ao déficit hídrico. O objetivo deste estudo foi selecionar características fenotípicas com elevada associação com a produtividade de raízes (PTR), além de estabelecer um modelo de predição do comportamento dos genótipos em condições de déficit hídrico. Foram avaliados 49 genótipos de mandioca em delineamento em blocos completos casualizados, com três repetições e em duas condições hídricas: com irrigação (controle) e sob déficit hídrico. As variáveis fisiológicas e agronômicas foram divididas em três grupos: Fisiol (todas as características fisiológicas); Fisiol+PPA (todos as características fisiológicas, com adição da produtividade de parte aérea) e Fisio+Agro (todos as características fisiológicas e agronômicas), e avaliados por seis diferentes técnicas de modelagem preditivas: *classification and regression trees* (CART), *artificial neural network* (ANN), *support vector machines* (SVM), *extreme learning machine* (ELM), *generalized linear model with stepwise feature selection* (GLMSS) e *partial least squares* (PLS). Outros três grupos foram formados pelos mesmos grupos, porém contendo apenas as características mais importantes para cada condição hídrica. As características mais importantes para predição da produtividade total de raízes (PTR) foram: Número de raízes por planta (NRP); Área abaixo da curva de progressão da expansão das folhas com base no índice de área foliar (AACP.IAF); Número de folhas mensurado no oitavo mês (NF.8) e Produtividade da parte aérea (PPA). A seleção das características mais importantes resultou no melhor ajuste dos modelos, sendo GLMSS; ELM; e PLS os modelos que apresentaram maior confiabilidade de predição de acordo com os valores de $r^2 > 0{,}75$ com *RMSE* variando entre 0,49 e 0,51.

Palavras chave: *Manihot esculenta* Crantz, produtividade de raízes, seca, performance agronômica.

Prediction models and selection of agronomic and physiological traits for tolerance to water deficit in cassava

Abstract: The development of efficient and accurate strategies for evaluating and predicting the root yield of cassava (*Manihot esculenta* Crantz), can reduce the effort spent on complex phenotyping such as tolerance to water deficit. The objective of this study was to select phenotypic traits in high association with fresh root yield (RoY), in addition to establish a prediction model of the performance of genotypes under water deficit conditions. A total of 49 cassava genotypes were evaluated in a complete randomized block design, with three replications and two water conditions: irrigation (control) and water deficit. The physiological and agronomic traits were divided into three groups: Fisol (all physiological traits); Fisiol + SP (all physiological traits, with addition of shoot yield) and Fisio + Agro (all the physiological and agronomic traits), and evaluated by six different predictive model: classification and regression trees (CART), artificial neural network (ANN), support vector machines (SVM), extreme learning machine (ELM), generalized linear model with stepwise feature selection (GLMSS) e partial least squares (PLS). Three other groups were formed by the same groups, but they only contained the most important traits for each water condition. The most important traits for predicting RoY were: number of roots per plant, leaf area index, number of leaves measured in the eighth month, and shoot yield. The selection of the most important traits resulted in the best adjustment of the models, being GLMSS, ELM, and PLS, the models that presented the highest reliability of prediction according to the values of $r^2 > 0.75$ with RMSE ranging from 0.49 to 0.51.

Keywords: *Manihot esculenta* Crantz, root yield, drought, agronomic performance.

INTRODUÇÃO

A mandioca (*Manihot esculenta* Crantz), possui uma ampla gama de usos na alimentação humana ou animal. É um alimento básico para mais de 800 milhões de pessoas, sendo a terceira fonte de calorias mais importante na América Latina, Ásia e África, perdendo apenas para o arroz e o milho (CEBALLOS et al., 2010; LIU et al., 2011; CIAT, 2017). Do ponto de vista industrial, o amido da mandioca possui aplicações na indústria têxtil, plástica, siderúrgica, papeleira, farmacêutica, alimentícia e de biocombustíveis (CEBALLOS et al., 2012; FAO, 2013). No Brasil, quarto produtor mundial, a mandioca é produzida principalmente por pequenos agricultores, representando uma fonte de segurança alimentar, devido a sua rusticidade e tolerância a estresses abióticos (MORANTE et al., 2010; CEBALLOS et al., 2011; OKOGBENIN et al., 2013; FAO, 2016).

A mandioca é uma cultura que apresenta boa adaptabilidade em diversos ambientes e ecossistemas (EL-SHARKAWY, 2012; OKOGBENIN et al., 2013). Essa plasticidade fenotípica também pode ser observada quanto à tolerância ao déficit hídrico, uma vez que a mandioca atinge produtividade de raízes relativamente elevada, mesmo quando exposta a baixos níveis de precipitação (BERGANTIN et al., 2004; EL-SHARKAWY, 2007; AINA et al., 2007; OKOGBENIN et al., 2013). A redução no rendimento das raízes depende da duração do estresse e do estágio de desenvolvimento em que a planta se encontra. O período crítico para o efeito mais severo do déficit hídrico na cultura estende-se do primeiro ao quinto mês após o plantio, que é o período correspondente a formação das raízes e do rápido crescimento foliar (OKOBEGNIN et al., 2013). Entretanto, apesar da redução do rendimento de raízes sob déficit hídrico, a mandioca possui elevada capacidade de recuperação e consegue atingir alta produtividade de biomassa após a reidratação, em virtude da rápida formação de folhas novas com elevadas taxas fotossintéticas, resultando em uma produção de raízes semelhante ao seu cultivo em condições hídricas ideais (EL-SHARKAWY, 2007).

Devido ao cultivo em condições adversas, a mandioca desenvolveu mecanismos fisiológicos para suportar períodos de seca prolongados e evitar a

desidratação (EL-SHARKAWY, 2007). Os principais mecanismos desenvolvidos foram a redução do crescimento dos ramos, área foliar reduzida por meio da senescência das folhas mais velhas, surgimento de folhas novas para realização da fotossíntese e elongação da raiz principal para captação de água em áreas mais profundas do solo. Outros mecanismos observados independentes da intensidade do estresse hídrico são o acúmulo do hormônio ácido abscísico (*abscisic acid* - ABA) nas folhas das plantas submetidas ao déficit de água e o fechamento estomático, resultando na manutenção do alto potencial hídrico nos tecidos (ALVES & SETTER, 2004; DUQUE & SETTER, 2013). Em relação ao ajustamento osmótico em condições de déficit hídrico, há um decréscimo na concentração de açúcares totais no interior da célula e um aumento na concentração de prolina, o que não é suficiente para ocasionar alteração no ajuste osmótico, que é controlado por sais de potássio (ALVES & SETTER, 2004). Todos esses mecanismos estão relacionados principalmente com a conservação da água nos tecidos de modo a conservar esse recurso escasso (EL-SHARKAWY, 2007; DUQUE & SETTER, 2013).

Embora a mandioca seja uma cultura com potencial tolerância à seca, as variedades cultivadas em áreas comerciais apresentam uma redução significativa no desempenho produtivo sob condições de déficit hídrico (OLIVEIRA et al., 2015). Por outro lado, existe uma enorme variabilidade genética na espécie para tolerância a este estresse, sobretudo nas regiões semiáridas do Brasil (OKOGBENIN et al., 2013; OLIVEIRA et al., 2017). Diversos acessos de mandioca já foram identificados e selecionados, para serem utilizados como parentais em programas de melhoramento, por apresentar alta produtividade sob condições de déficit hídrico (EL-SHARKAWY, 2012; OKOGBENIN et al., 2013; OLIVEIRA et al., 2017). Portanto, a existência de genes de tolerância à seca no germoplasma de mandioca é uma premissa importante para início das atividades de melhoramento visando a obtenção de novas variedades que sejam capazes de garantir elevados patamares de produtividade mesmo em condições de estresse hídrico prolongado. Entretanto, o passo seguinte relacionado à identificação dos genótipos tolerantes, geralmente é um processo tedioso, de alto custo, e geralmente muito sujeito as variações ambientais. Neste processo de fenotipagem para

tolerância à seca, é necessário que sejam investigadas quais características fisiológicas e agronômicas estão intimamente relacionadas à variação na resposta dos genótipos quando submetidos à restrição hídrica (LABAN et al., 2013).

Tendo em vista os custos elevados de fenotipagem para diversos atributos fisiológicos e agronômicos, a seleção de variáveis capazes de predizer a produtividade de raízes em condições de déficit hídrico, constitui uma ferramenta importante para auxiliar os programas de melhoramento genético no processo seletivo. Para realizar a predição de parâmetros de um conjunto de dados multivariados, geralmente utiliza-se modelos de regressão e classificação (ANDERSEN & BRO, 2010; MEHMOOD et al., 2012). Esses modelos, são eficazes para reduzir o número de variáveis avaliadas, por meio da seleção daquelas que mais contribuem para o modelo de predição, proporcionando o aumento da capacidade preditiva dos modelos (DAN et al., 2016; CHRISTENSON et al., 2016).

Diversas técnicas de modelagem preditivas têm sido utilizadas para aperfeiçoar a predição da produtividade agrícola, sendo que os modelos de classificação e regressão frequentemente utilizados são o *Artificial Neural Network* (ANN), *Classification and Regression Trees* (CART*), Generalized Linear Model with Stepwise Feature Selection* (GLMSS), *Support Vector Machines* (SVM) e *Partial Least Squares* (PLS) (PARK et al., 2005; MUCHERINO et al., 2009; RUß, 2009; WEBER et al., 2012; MEHMOOD et al., 2012). O modelo CART foi utilizado com alta eficiência na compreensão dos fatores que influenciam a produtividade da cana-de-açúcar (FERRARO et al., 2009), enquanto que o modelo PLS apresentou alta capacidade de predição da produtividade de grãos em variedades de milho submetidas à restrições hídricas (WEBER et al., 2012). Contudo, ainda não há relatos da utilização do modelo *Extreme Learning Machine* (ELM) para a predição da produtividade em vegetais, embora sua aplicação em outras áreas tem mostrado sua elevada capacidade de predição, em alguns casos superior aos modelos SVM e ANN (OLATUNJI et al., 2014; ABDULLAH et al., 2015; MOHAMMADI et al., 2015).

A precisão de um modelo de predição é dependente do conjunto de dados em análise, sendo necessária a avaliação de vários modelos para

determinar aquele que permite melhor ajuste e acurácia. De acordo com RUß (2009), o modelo SVM apresentou maior acurácia na predição da produtividade de trigo, em comparação com os modelos ANN e CART. Por outro lado, o modelo CART seguido pelo ANN apresentou maior predição da produtividade de grãos de milho, sob diferentes solos e condições de manejo, quando comparados com o modelo GLMSS (PARK et al., 2005). Portanto, tendo em vista que o maior interesse é que os modelos expressem a máxima capacidade de predição, a comparação entre modelos baseados em diferentes princípios matemáticos deve ser estudada.

Considerando que diferentes métodos de predição e classificação podem ser utilizados para predizer a produtividade agrícola com precisão, o presente estudo teve como objetivo selecionar as variáveis fisiológicas e agronômicas mais estreitamente associadas à produtividade de raízes em diferentes genótipos de mandioca, visando estabelecer um modelo de predição que possa contribuir na seleção de genótipos mais produtivos sob déficit hídrico, seja em bancos de germoplasma ou programas de melhoramento.

MATERIAL E MÉTODOS

Material vegetal, local e condições experimentais

Foram avaliados 49 genótipos pertencentes ao Banco Ativo de Germoplasma de Mandioca da Embrapa Semiárido e Embrapa Mandioca e Fruticultura, compostos por 25 variedades locais e 24 variedades melhoradas com histórico de tolerância à seca (Material suplementar, Tabela S1). Estes genótipos foram avaliados em experimentos com delineamento em blocos completos casualizados (DBCC), com três repetições e em duas condições hídricas: irrigado (controle) e sob déficit hídrico. Inicialmente, ambos os experimentos foram submetidos à irrigação até o quarto mês após o plantio, em seguida, a irrigação foi suspensa somente no experimento destinado à aplicação do estresse hídrico. A lâmina de água aplicada foi calculada em função da evapotranspiração da cultura, usando dados meteorológicos fornecidos pela estação meteorológica. As parcelas experimentais foram compostas de dez plantas, com espaçamento de 0,90 m entre linhas e 0,80

entre plantas. O plantio foi realizado com manivas de 16 cm, seguindo as recomendações e práticas agrícolas para a cultura.

Os experimentos foram realizados na Estação Experimental de Bebedouro, da Embrapa Semiárido, em Petrolina - PE (9°22’ de latitude Sul, 40°22' de longitude Oeste e altitude de 365,5 m), durante dois anos agrícolas (2012/2013 e 2013/2014). O local de avaliação apresenta clima semiárido e baixa precipitação pluviométrica anual. As condições climáticas do ano 2013 foram de precipitação anual média de 347,8 mm, umidade relativa do ar variando entre 48 e 61% e temperatura média entre 27,7 e 29,2 ºC. Em 2014, a precipitação anual média foi de 216,3 mm, umidade relativa do ar entre 55 e 67% e temperatura média variando entre 24,5 e 26,9 ºC (EMBRAPA SEMIÁRIDO, 2013; EMBRAPA SEMIÁRIDO, 2014). Portanto, as safras de 2012/13 e 2013/14 foram marcadas por condições meteorológicas com baixo volume de precipitação e consequente ocorrência severa de déficit hídrico, principalmente no período entre os meses de maio a outubro, nos dois anos agrícolas.

Variáveis fisiológicas e agronômicas analisadas

Para cada condição hídrica, foram realizadas avaliações para diversas características agronômicos e fisiológicos, que se iniciaram no quarto mês após o plantio, onde a irrigação foi suspensa no experimento destinado à aplicação do estresse hídrico. As características fisiológicas avaliadas foram:

1) Temperatura foliar no 4° (TF.4), 6° (TF.6), 9° (TF.9) e 10° (TF.10) mês após o plantio (MAP); e diferença entre a temperatura foliar e do ambiente mensuradas no 4° (DTFA.4), 6° (DTFA.6), 9° (DTFA.9) e 10° MAP (DTFA.10), após o plantio. A temperatura foliar foi mensurada utilizando um porômetro de difusão (modelo SC1 - Decagon) em folhas completamente expandidas, localizadas na parte superior da planta, exposta à radiação solar entre 09:00 e 11:00 horas;
2) Índice de área foliar (IAF, m^2 m^{-2}), mensurado no 5º e 6º MAP, utilizando o equipamento Accupar (modelo LP80 - Decagon), posicionando a linha de sensores junto à superfície do solo sob a copa da planta, exposta à radiação solar entre 09:00 e 11:00 horas; considerando a área abaixo da

curva de progressão da expansão das folhas com base no índice de área foliar (AACP.IAF), de acordo como o modelo de Campbell & Madden (1990);

3) Índice de retenção foliar (IRF, %), mensurado do 4° ao 12° MAP, com intervalo de 15 dias, por meio da porcentagem de folhas presentes na planta;
4) Índice relativo de clorofila (IRC, SPAD), mensurado no 4º (IRC.4), 7º (IRC.7) e 11º MAP (IRC.11), mensurado por meio do equipamento Clorofilog (Modelo CFL1030 - Falker), em folhas novas totalmente expandidas, pré-aclimatadas por 20 minutos à condição de escuro com "*dark leaf clips*";
5) Eficiência quântica potencial dos fotossistemas II (Fv/Fm), mensurada no 7° MAP, utilizando o equipamento Chlorophyll Fluorometer (Modelo OS30p+ - OptiScience), no período entre as 8:00 e 11:00horas, em folhas novas inteiramente expandidas, pré-aclimatadas por 20 minutos à condição de escuro com "*dark leaf clips*".

As características agronômicas avaliadas antes da colheita foram:

1) Diâmetro do caule (DC, cm), por meio de três mensurações, realizadas no 4° (DC.4), 7° (DC.7) e 11° MAP (DC.11), utilizando um paquímetro a 15 cm do solo;
2) Área abaixo da curva de progressão do crescimento das hastes com base no diâmetro do caule (AACP.DC) conforme Campbell & Madden (1990);
3) Número de folhas (NF), mensurado no 8° (NF.8) e 10° MAP (NF.10).

As características agronômicas avaliadas durante a colheita foram:

1) Altura da planta (AP, cm), do solo até broto apical;
2) Área abaixo da curva de progressão do crescimento das plantas com base na altura das plantas (AACP.AP) e considerando a taxa de crescimento relativo da altura da planta (AACP.TCR.AP);
3) Número de raízes por planta (NRP);

4) Produtividade da parte aérea (PPA, kg), por meio da pesagem da parte aérea (folhas, pecíolos e hastes) a partir do corte realizado a 10 cm da superfície do solo;
5) Produtividade de raízes (PTR, t ha^{-1}), comerciais e não-comerciais;
6) Teor de matéria seca da raiz (DMC, %), avaliada pelo método da balança hidrostática (KAWANO et al., 1987).

Obtenção dos BLUPs, herdabilidade e analise de deviance

Foi realizada uma análise de variância individual (características avaliadas somente em um ano agrícola) e conjunta (características avaliadas nos dois anos agrícolas) para cada condição hídrica, para obtenção dos componentes de variância e estimação dos BLUPs (*best linear unbiased prediction*). Os BLUPs da análise individual foram obtidos para cada característica, onde as observações fenotípicas Y_{ij} do genótipo *i* na repetição *j*, foi modelada pela equação $Y_{ij} = \mu + g_i + \epsilon_{ij}$, no qual, μ é a média geral, g_i é o efeito aleatório do genótipo *i*, e ϵ_{ij} é o efeito residual aleatório do genótipo *i* na repetição *j*. Para a análise conjunta, as observações fenotípicas Y_{ijk} do genótipo *i* na repetição *j* dentro do ambiente *k*, foi modelada pela equação $Y_{ij} = \mu + e_k + g_i + ({}^{r}/_{e})_{jk} + (g * e)_{ik} + \epsilon_{ijk}$, no qual, μ é a média geral, e_k é o efeito fixo do ano *k*; g_i é o efeito aleatório do genótipo *i*, $(r/e)_{jk}$ é o efeito aleatório da repetição *j* aninhada no ano *k*, $(g * e)_{ik}$ é o efeito aleatório da interação entre genótipos e ano e ϵ_{ijk} é o efeito residual aleatório do genótipo *i* na repetição *j* no ano *k*.

A estimação dos componentes de variância foi realizada pelo método REML (*Restricted Maximum Likelihood*), utilizando o pacote *lme4* do software R versão 3.4.4 (R CORE TEAM, 2018). A significância dos efeitos aleatórios foi obtida pelas análises de *deviance* (*Anadev*) usando o método REML. A significância do modelo completo e do modelo sem o efeito foi comparado pelo teste *qui-square*. A herdabilidade no sentido amplo (h^2) foi estimada de acordo com $h^2 = \frac{\sigma_G^2}{\sigma_F^2 + \sigma_E^2}$, onde σ_G^2 é a variância genotípica; σ_F^2 é a variância fenotípica e σ_E^2 é a variância ambiental, para características avaliadas em apenas um ano

agrícola, e $h^2 = \frac{\sigma_G^2}{(\sigma_G^2+\sigma_A^2+\frac{\sigma_{GxA}^2}{r}+\frac{\sigma_\varepsilon^2}{rxa})}$, onde σ_G^2 é a variância genotípica; σ_A^2 é a variância de ano; σ_{GxA}^2 é a variância da interação genótipo x ano; σ_ε^2 é a variância do erro entre parcelas; r é o número de repetições e a o número de anos.

Seleção de variáveis, modelos de predição e validação

Para a seleção de variáveis uteis para explicar os modelos de predição da produtividade de raízes, foram utilizados os seguintes modelos: *classification and regression trees* (CART), *artificial neural network* (ANN), *support vector machines* (SVM), *extreme learning machine* (ELM), *generalized linear model with stepwise feature selection* (GLMSS) e *partial least squares* (PLS). A seleção de variáveis foi realizada utilizando como critério a importância das variáveis (acurácia) dentro da estrutura dos modelos, utilizando a função *varImp* do pacote *caret* do software R versão 3.4.4 (R CORE TEAM, 2018). A importância das variáveis foi relativizada para 100% e as variáveis com mais de 50% de importância para explicar os modelos preditivos foram selecionadas. Exceção a este procedimento ocorreu nos modelos GLMSS, ELM, ANN e SVM, na qual não é possível inferir a importância das variáveis a partir da estrutura do próprio modelo. Deste modo, foi utilizado uma abordagem do tipo filtro, para buscar um subconjunto de variáveis capaz de minimizar o erro de predição. A classificação das variáveis foi realizada conforme a área abaixo da curva ROC (*receiver operating characteristic*). A regra trapezoidal foi usada para calcular a área sob a curva ROC, que em seguida foi utilizada como medida de importância da variável.

No modelo PLS a importância das variáveis foi baseada na relação entre a matriz X (variáveis dependentes) e a matriz Y (composta por variáveis independentes), permitindo analisar dados com desequilíbrio entre o número de variáveis e observações; com forte correlação e com elevados níveis de ruído. Esta técnica de regressão dá origem a um conjunto de parâmetros que contém informações sobre a estrutura e comportamento das matrizes X e Y. A seleção do número de variáveis a serem mantidas no modelo é realizada com base na avaliação de significância em termos preditivos de cada variável, onde

a inclusão de variáveis no modelo é interrompida quando essas variáveis deixam de ser significativas (WOLD et al., 2001; ANDERSEN & BRO, 2010).

Diferentemente de outros métodos, a análise CART uma técnica não paramétrica que não requer nenhuma pressuposição dos dados. O CART envolve o particionamento recursivo binário, que divide o conjunto original de dados e analisa as partições com base na razão de verossimilhança (LR) do qui-quadrado para testes de independência para cada possível divisão. O *p*-valor para cada teste qui-quadrado representa a probabilidade de obter um valor de qui-quadrado maior do que aquele encontrado pelo acaso. O critério utilizado para selecionar as partições leva em consideração a maximização da significância de cada variável dividido, ao invés apenas da estatística do teste. Cada variável candidata é classificada pela estatística *logworth* (log negativo dos p-valores ajustados para qui-quadrado) para identificar a divisão ideal para cada grupo de dados. Para variáveis contínuas, as divisões são construídas em torno de um valor de corte que maximiza a separação dos grupos, tendo como critério a soma dos quadrados das diferenças entre as médias dos dois grupos (AFONSO ET AL., 2012).

Para determinar o grupo de variáveis que pudessem gerar o modelo de predição mais acurado, as variáveis agronômicas e fisiológicas foram analisadas considerando três agrupamentos, para cada condição hídrica (irrigado e sob déficit hídrico). Inicialmente, foi realizada a importância das variáveis dos seguintes grupamentos: Fisiol – todas as características fisiológicas; Fisiol+PPA – todos as características fisiológicas, com adição da produtividade de parte aérea; Fisio+Agro – todos as características fisiológicas e agronômicas. Em seguida foi realizada a seleção das variáveis e mantido o mesmo agrupamento das variáveis para fins de predição, sendo: Fisiol-Sel – características fisiológicas mais importantes; Fisiol+PPA-Sel – características fisiológicas mais importantes, com adição da produtividade de parte aérea e Fisio+Agro-Sel – características fisiológicas e agronômicas mais importantes. Esses seis agrupamentos foram analisados com o intuito de verificar aquele que mais contribui no melhor ajuste dos modelos de predição da produtividade de raízes.

Após a determinação dos grupos de características, os modelos foram validados a partir dos dados de treinamento e utilizados para predizer a produtividade de raízes no conjunto de validação. Foi realizada uma validação cruzada (*cross-validation* – CV), com 3 repetições, 10-*folds* cada. Portanto, 90% das amostras foram usadas como população de treinamento e 10% como população de validação, em cada *fold*. O desempenho dos modelos de predição, para cada grupo de variáveis, foi avaliado com base na média dos valores de *root mean square error* (*RMSE*) e do coeficiente de regressão (r^2), obtidos em cada partição da CV.

O *RMSE* que é a medida da magnitude média dos erros estimados, foi utilizado para analisar os erros das estimativas. O *RMSE* possui valor positivo e quanto mais próximo de 0, maior será a qualidade dos valores estimados. O cálculo do *RMSE* foi realizado de acordo com a equação: $RMSE\sqrt{\frac{1}{n}\sum_{i=l}^{n}(E_i - O_i)^2}$, em que E_i e O_i são os valores estimados e observados, respectivamente, e n é o número de observações.

O r^2 varia entre 0 e 1, e quanto mais próximo de 1, melhor será o seu poder de explicação. O r^2 é descrito como a relação que pondera a proporção da variação total da variável dependente que é explicada pela variação da variável independente. É estimado considerando a equação: $r^2 = \frac{\sum_{i=l}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=l}^{n}(y_i - \bar{y}_i)^2}$, em que: $\sum_{i=l}^{n}(\hat{y}_i - \bar{y})^2$ corresponde à variação explicada, e $\sum_{i=l}^{n}(y_i - \bar{y})^2$ corresponde à variação não explicada pelos modelos. Todas as análises de predição foram realizadas utilizando-se o pacote *caret* do software R versão 3.4.4 (R CORE TEAM, 2018).

RESULTADOS

Componentes de variância e herdabilidade

De modo geral, a herdabilidade (h^2) no experimento irrigado variou de baixa (0,20 para IRF e AP) a alta (0,71 para PTR), enquanto no experimento de sequeiro a h^2 foi ainda menor sendo de magnitude baixa (0,20 para NF-10) a mediana (0,52 para IRC-7) (Tabela 1). Além disso, a maioria das

características avaliadas foram significativas pelo teste *qui-square*, com exceção das características fisiológicas TF e DTFA avaliados aos 9 e 10 meses, IRC-4 e Fv/Fm, para o experimento irrigado. Já para o experimento de sequeiro, todas as características agronômicas foram significativas pelo teste *qui-square*. Entretanto, para as características fisiológicas os genótipos de mandioca apresentaram diferenças significativas apenas para IRF e IRC avaliado aos 7 e 11 meses.

Tabela 1- Componentes de variância e herdabilidade no sentido amplo obtidos pela análise de variância individual e conjunta entre os anos 2013 e 2014 para cada condição hídrica para diversas características fisiológicas e agronômicas.

Características[1] /	Experimento Irrigado				Experimento Sequeiro			
Componentes	σ_G^2	σ_{GxA}^2	σ_ε^2	h^2	σ_G^2	σ_{GxA}^2	σ_ε^2	h^2
TF-4 / DTFA-4	1,58**	-	2,44	0,39	0,01^{ns}	-	6,42	
TF-6 / DTFA-6	1,39**	-	3,94	0,26	0,66^{ns}	-	4,24	
TF-9 / DTFA-9	0,41^{ns}	-	2,74		0,01^{ns}	-	4,47	
TF-10 / DTFA-10	0,01^{ns}	-	4,72		0,53^{ns}	-	3,91	
AACP.IAF	1506,63**	-	2111,51	0,42	4,28^{ns}	-	406,68	
IRF	29,55**	-	116,38	0,20	6,58**	-	20,30	0,24
IRC-4	3,01^{ns}	-	34,62		8,50^{ns}	-	47,66	
IRC-7	17,83**	-	12,37	0,59	17,06**	-	15,48	0,52
IRC-11	13,34**	-	13,18	0,50	8,41**	-	22,47	0,27
Fv/Fm	0,01 ^{ns}	-	0,01		0,01 ^{ns}	-	0,01	
DC-4	0,12**	-	0,14	0,46	0,06**	-	0,21	0,22
DC-7	0,13**	-	0,16	0,44	0,07**	-	0,20	0,26
DC-11	0,15**	-	0,14	0,53	0,08**	-	0,19	0,30
AACP.DC	625,25**	-	443,80	0,58	730,36**	-	882,06	0,45
AP	0,06**	-	0,07	0,45	0,04**	-	0,05	0,47
AACP.AP	2075,24**	-	1815,01	0,53	1357,97**	-	1874,24	0,42
AACP.TCR.AP	0,01**	-	0,04	0,23	0,01*	-	0,03	0,24
NF-8	4504,20**	-	2438,88	0,65	180,72*	-	687,48	0,21
NF-10	3153,16**	-	3447,99	0,48	63,27*	-	261,00	0,20
NRP	1,36*	0,51*	1,96	0,52	0,47*	0,46**	2,10	0,33
PPA	36,88*	18,31*	46,19	0,20	3,24**	8,78**	15,67	0,32
DMC	3,49*	0,05^{ns}	6,93	0,62	5,50**	9,70^{ns}	7,15	0,44
PTR	94,99*	55,87*	33,38	0,71	3,94**	8,65**	13,10	0,34

* e ** significativos a 5 e 1% de probabilidade e ns não significativo pelo teste *qui-square*; σ_G^2 – variância genotípica; σ_ε^2- variância do erro; h^2 - herdabilidade no sentido amplo; σ_{GxA}^2 – variância da interação genótipo x ano; (1) TF – Temperatura foliar; DTFA – Diferença entre a temperatura foliar e do ambiente; IAF – Índice de área foliar; IRF – Índice de retenção foliar; IRC – Índice relativo de clorofila; Fv/Fm – Eficiência quântica potencial dos fotossistemas II; DC – Diâmetro do caule; AP – Altura da planta; NF – Número de folhas; NRP – Número de raízes por planta; PPA – Produtividade de parte aérea; DMC – Teor de matéria seca; PTR – Produtividade total de raízes; TRC – taxa de crescimento relativo; AACP – Área abaixo da curva de progressão; e os números representam os meses de avaliação.

Importância e seleção de variáveis

Para os dados Fisiol (todas as características fisiológicas), nove das 14 características fisiológicas avaliadas para a condição irrigada foram selecionados por apresentarem importância relativa superior a 50% (AACP.IAF; IRC.4; IRC.7; DTFA.6; TF.6; IRC.11; TF.9; IRF e DTFA.4), considerando os métodos de classificação CART, PLS e Curva ROC. Dentre elas, a variável AACP.IAF foi a mais importante (100% de importância) entre todos os métodos de classificação, com exceção do método CART que considerou o IRC.7 (100% de importância) como a variável mais importante (Figura 1a). Na condição sob déficit hídrico, apenas quatro variáveis foram selecionadas, com valor de importância superior a 50% (AACP.IAF; TF.4; DTFA.4; IRC.7). Porém, assim como na condição irrigada, a característica AACP.IAF foi considerada como mais importante, diferindo apenas no método CART, que considerou a variável TF.4 como a mais importante (Figura 2a).

Para os dados Fisiol+PPA foram avaliadas 14 características fisiológicas além da produtividade da parte aérea (PPA), sendo 10 delas selecionadas com mais de 50% de importância relativa (PPA; AACP.IAF; IRC.4; IRC.7; DTFA.6; TF.6; IRC.11; TF.9; DTFA.9 e IRF) para a condição irrigada, nos métodos de classificação CART, PLS e Curva ROC. A característica PPA foi considerada como a mais importante, nos métodos de classificação utilizados, à exceção de PLS, que considerou AACP.IAF (Figura 1b). Nos experimentos com déficit hídrico foram selecionadas cinco características (AACP.IAF; PPA; TF.4; DTFA.4 e IRC.7) com importância superior a 50%. A variável AACP.IAF foi tida como a mais importante por todos os métodos de regressão utilizados (Figura 2b).

Para os dados Fisio+Agro, das 20 características fisiológicas e agronômicas avaliadas na condição irrigada, apenas 2 foram selecionados com acurácia superior a 50% (NRP e PPA), a característica NRP foi considerada

como mais importante para todos os métodos de regressão, com 100% de importância relativa para todos os métodos (Figura 1c). Na condição não irrigada, 4 características foram selecionadas com acurácia superior a 50% (NRP; AACP.IAF; NF.8 e PPA), dentre essas características, a NRP foi a mais importante para todos os modelos avaliados (100% de importância relativa) (Figura 2c).

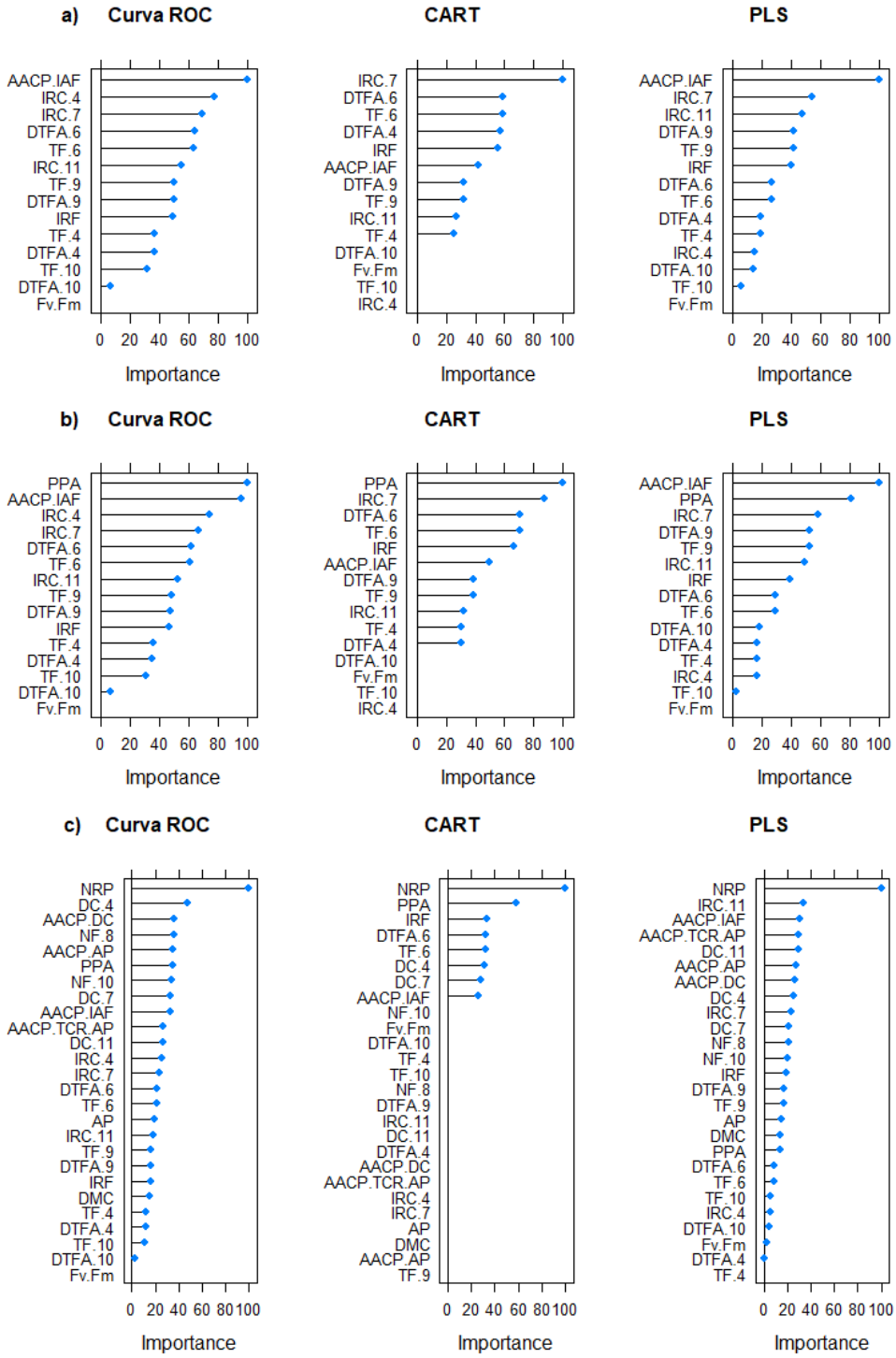

Figura 1. Importância das variáveis fisiológicas e agronômicas para a condição irrigada pelos métodos curva de ROC (*Receiver Operating Characteristics*), CART (*Classification and Regression Trees*) e PLS (*Partial Least Squares*); a) todas as características fisiológicas; b) todas as características fisiológicas, com adição da produtividade de parte aérea; e c) todas as características fisiológicas e agronômicas.

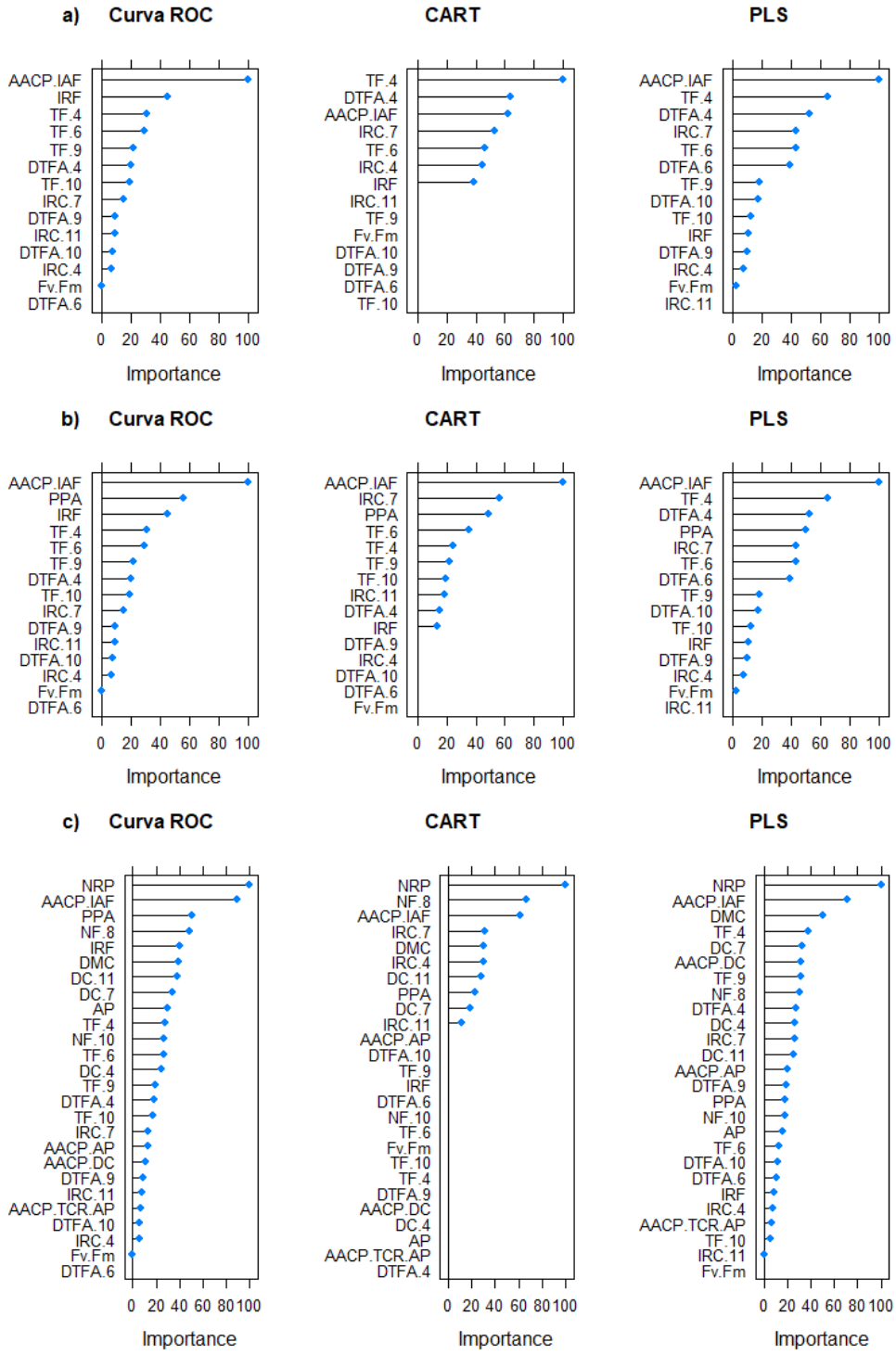

Figura 2. Importância das variáveis fisiológicas e agronômicas para a condição não irrigada pelos métodos curva de ROC (*Receiver Operating Characteristics*), CART (*Classification and Regression Trees*) e PLS (*Partial Least Squares*); a) todas as características fisiológicas; b) todas as características fisiológicas, com adição da produtividade de parte aérea; e c) todas as características fisiológicas e agronômicas.

As variáveis selecionadas para cada condição hídrica foram utilizadas para compor o Fisiol-Sel (apenas características fisiológicas), Fisiol+PPA-Sel (características fisiológicas + PPA), Fisio+Agro-Sel (todas as características fisiológicas e agronômicas) (Material suplementar, Figuras S1 e S2).

Comparação entre os modelos de predição para produtividade de raízes

A predição da PTR em experimentos irrigados, com base apenas nos dados fisiológicos foi de baixa magnitude, com variação de r^2 entre 0,31 e 0,41 e RMSE entre 0,86 e 1,07. A inclusão da característica PPA aumentou o r^2 para a maioria dos modelos de predição, à exceção do ANN e CART, porém a qualidade da predição ainda foi bastante baixa (r^2 entre 0,33 e 0,46 e RMSE entre 0,82 e 1,04). Esta mesma tendência ocorreu ao se incluir as outras características agronômicas (altura da planta, área abaixo da curva de progressão do crescimento das plantas, número de raízes por planta, teor de matéria seca da raiz) na análise conjunta com os dados fisiológicos. Neste caso, o r^2 aumentou bastante (variação de 0,5 a 0,72), enquanto o RMSE reduziu (variação de 0,62 a 0,80). Portanto, a inclusão de parâmetros agronômicos de crescimento e de componentes de produção foram de fundamental importância para melhorar a capacidade preditiva dos modelos (Figura 3).

Ainda nos experimentos irrigados, verificou-se que a implementação dos modelos de predição apenas nas variáveis selecionadas de maior importância não resultou em aumento significativo do r^2 quando se utilizou apenas as características fisiológicas (Fisiol-Sel) ou características fisiológicas além da PPA (Fisiol+PPA-Sel) (Figura 3). No entanto, quando todas as características selecionadas foram analisadas de forma conjunta (Fisio+Agro-Sel), houve um

aumento importante do r^2 para a maioria dos modelos. Isso foi confirmado pela menor variação entre os valores observados e preditos para produtividade total de raiz e maior inclinação da reta de regressão (Figura 4). Exceção ocorreu com o modelo SVM, cujo r^2 reduziu no modelo completo de 0,72 para 0,52 no modelo apenas com as variáveis selecionadas. Isto possivelmente ocorreu porque alguma variável resposta importante para este modelo deve ter sido descartada. Além disso, uma observação importante é que de modo geral o RMSE dos modelos que selecionaram as variáveis explicativas mais importantes, foi menor do que os modelos completos.

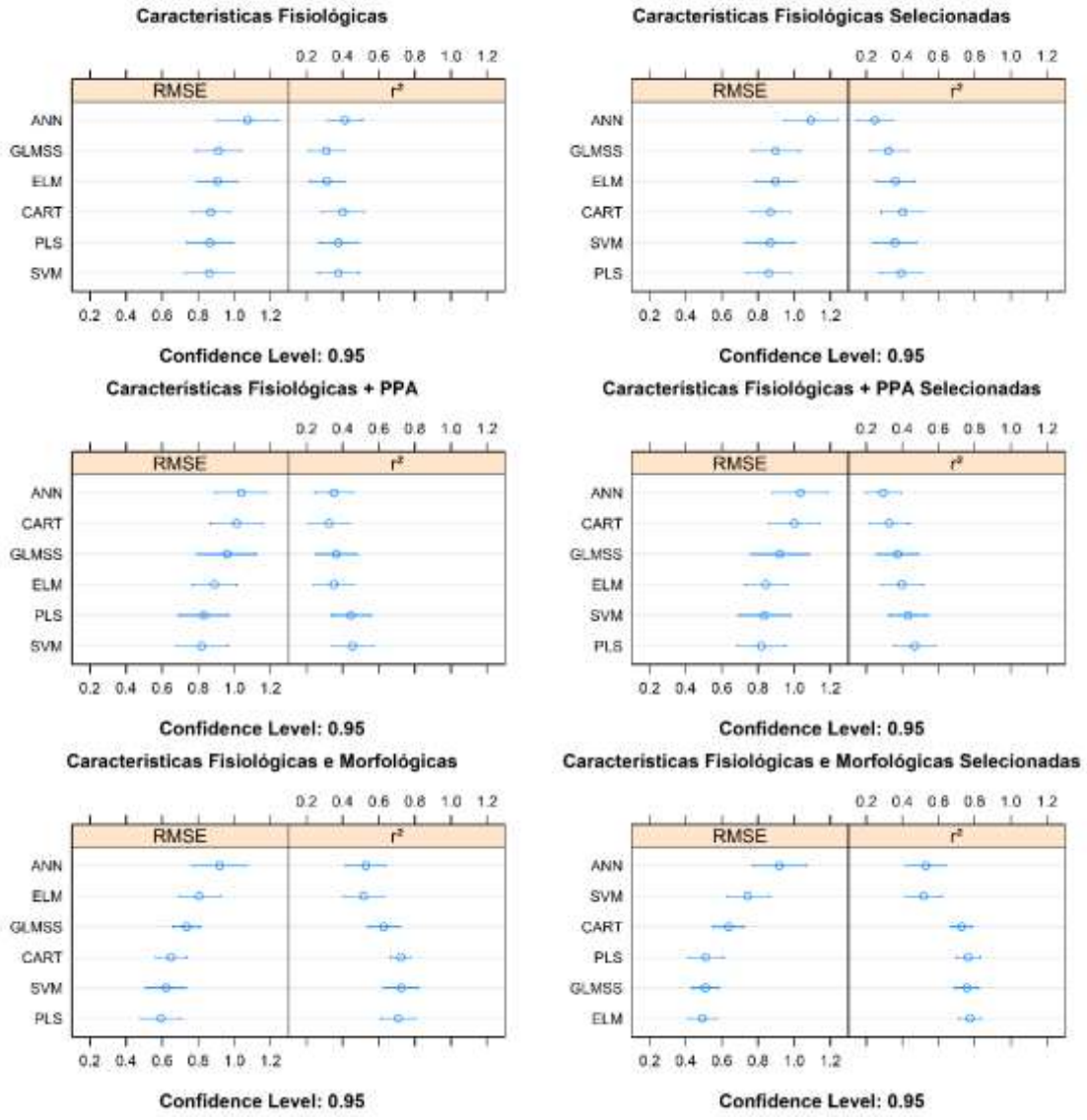

Figura 3. Comparação dos métodos de predição ANN (*Artificial Neural Network*); GLMSS (*Generalized Linear Model with Stepwise Feature Selection*);

CART (*Classification and Regression Trees*); ELM (*Extreme Learning Machine*); PLS (*Partial Least Squares*) e SVM (*Support Vector Machines*), para predição da produtividade de raízes de mandioca, utilizando variáveis fisiológicas e agronômicas em experimento irrigado.

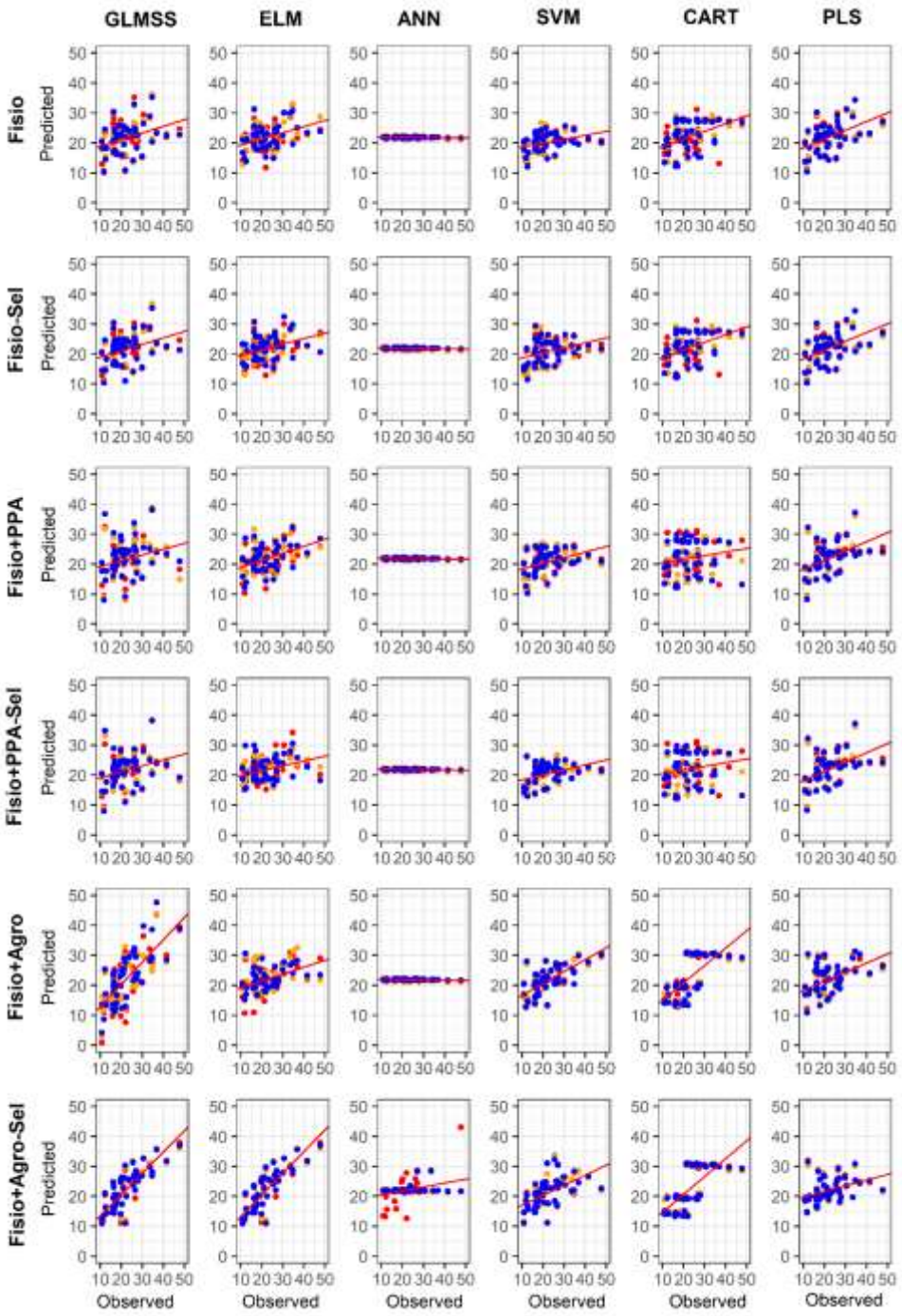

Figura 4. Regressão entre os valores observados e preditos obtidos nos esquemas de validação cruzada para produtividade de raízes em experimento irrigado, utilizando variáveis fisiológicas e agronômicas com base nos modelos de predição ANN (*Artificial Neural Network*), GLMSS (*Generalized Linear Model with Stepwise Feature Selection*), CART (*Classification and Regression Trees*), ELM (*Extreme Learning Machine*), PLS (*Partial Least Squares*) e SVM (*Support Vector Machines*).

De forma similar aos experimentos com irrigação, a inclusão da característica PPA nos experimentos submetidos ao déficit hídrico não resultou em aumento significativo do r^2 para a maioria dos modelos, à exceção do CART (r^2 de 0,21 e 0,32, nos dados Fisiol e Fisiol+PPA, respectivamente) (Figura 5). Em contrapartida, a inclusão dos dados agronômicos aumentou o r^2 em todos os modelos de predição (variação de 0,35 a 0,45) e reduziu o RMSE (à exceção do modelo ANN). Quando se aplicou os modelos de predição apenas nas características fisiológicas selecionadas, verificou-se um ligeiro aumento do r^2 e redução no RMSE, em comparação com o modelo com todas as características fisiológicas. Esta mesma tendência ocorreu quando se adicionou a característica PPA e se analisou todas as características fisiológicas e agronômicas simultaneamente (Figura 5). Neste caso, os modelos SVM, PLS e GLMSS apresentaram melhor acurácia utilizando as variáveis do grupo Fisio+Agro-Sel, além de apresentar um melhor ajuste entre os valores observados e preditos para produtividade de raiz em relação a reta de regressão (Figuras 5 e 6).

Em termos comparativos, ficou evidente que o estabelecimento de modelos de predição apenas com base nas variáveis de maior importância pode resultar em melhoria da capacidade preditiva (menor RMSE e maior r^2). Em termos médios, o aumento do r^2 foi de 11 e 24% nos experimentos irrigados e com déficit hídrico, respectivamente. Outro aspecto importante é que a capacidade preditiva (r^2) da PTR em mandioca variou de 0,52 a 0,77 nos experimentos irrigados e com seleção de variáveis, enquanto que nos experimentos com déficit hídrico a variação foi de 0,41 a 0,58 (Figuras 3 e 5).

Isto indica que, na média geral, a predição dos experimentos irrigados foi 33% maior em comparação com os de déficit hídrico.

De modo geral, foi possível identificar modelos bastante promissores para predição da PTR em mandioca nas duas condições hídricas avaliadas. De todos os modelos avaliados o PLS foi o que apresentou maior r^2 e menor RMSE, independentemente da seleção de variáveis e da condição hídrica sendo o modelo de predição mais estável.

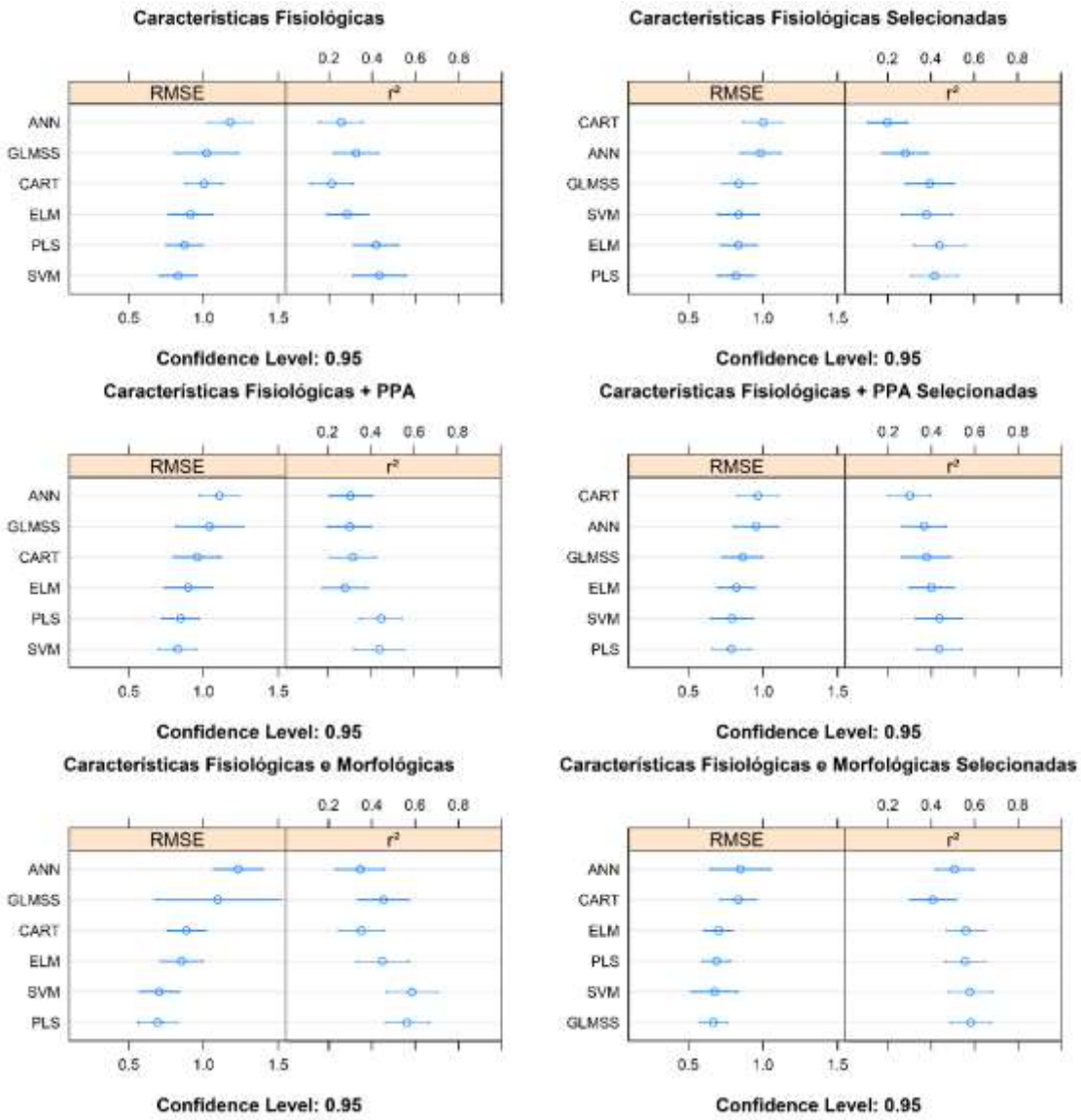

Figura 5. Comparação dos métodos de predição ANN (*Artificial Neural Network*); GLMSS (*Generalized Linear Model With Stepwise Feature Selection*); CART (*Classification And Regression Trees*); ELM (*Extreme*

Learning Machine); PLS (*Partial Least Squares*) e SVM (*Support Vector Machines*), para predição da produtividade de raízes de mandioca, utilizando variáveis fisiológicas e agronômicas em experimento submetido ao déficit hídrico.

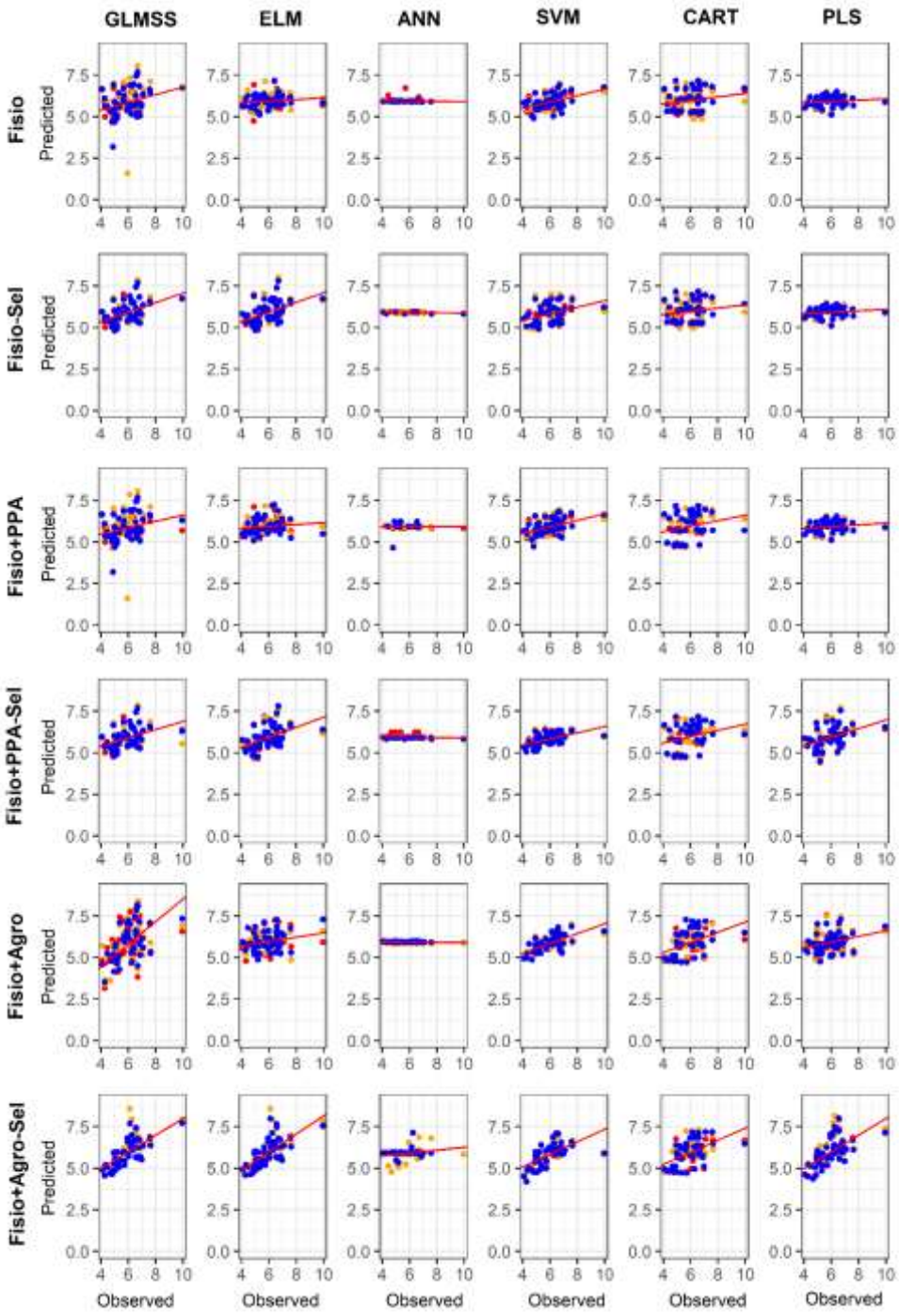

Figura 6. Regressão entre os valores observados e preditos obtidos nos esquemas de validação cruzada para produtividade de raízes em experimento de sequeiro, utilizando variáveis fisiológicas e agronômicas com base nos modelos de predição ANN (*Artificial Neural Network*), GLMSS (*Generalized Linear Model with Stepwise Feature Selection*), CART (*Classification and Regression Trees*), ELM (*Extreme Learning Machine*), PLS (*Partial Least Squares*) e SVM (*Support Vector Machines*).

DISCUSSÃO

Componentes de variância e herdabilidade para análise da tolerância à seca na mandioca

Com exceção de algumas características fisiológicas, os valores da h^2 na condição de irrigada foram de média a alta, porém na condição de sequeiro houve uma redução na h^2 para a maioria das características avaliadas. O decréscimo observado nas estimativas de h^2 é frequentemente associado às condições severas de estresse hídrico ao qual o experimento foi submetido, bem como ao material genético utilizado nos experimentos que foi composto por variedades tolerantes e suscetíveis ao déficit hídrico (FARFAN et al., 2015; OLIVEIRA et al., 2017). Os resultados observados no presente estudo corroboram os relatos de outros autores que avaliaram experimentos considerando condições irrigadas e de sequeiro, na qual as estimativas de h^2 foram menores para características avaliadas em condições de déficit em culturas como feijão comum (HINKOSSA et al., 2013), batata (CABELLO et al., 2014) e milho (BEYENE et al., 2015), bem como em mandioca (OLIVEIRA et al., 2015).

Seleção de variáveis para estabelecer um modelo de predição acurado

Um modelo de predição acurado da produtividade de raízes em condições de déficit hídrico, com base em características fisiológicas, proporcionaria uma economia de tempo e recursos com a fenotipagem, tendo em vista que as principais avaliações agronômicas são realizadas no final do ciclo da cultura, o que torna o processo de seleção muito lento e dispendioso (OKOGBENIN et al., 2013). Todavia, a tolerância ao estresse hídrico em

mandioca tem sido explicada pela existência combinada de mecanismos fisiológicos para evitar/tolerar a desidratação (OKOGBENIN et al., 2013), além da manutenção da capacidade fotossintética das folhas em ambientes com déficit hídrico prolongado (EL-SHARKAWY, 2007).

Apesar dos relatos das associações prévias entre características agronômicas *versus* tolerância ao déficit hídrico em mandioca, os resultados deste estudo demonstram que o uso de variáveis fisiológicas isoladamente não foi eficiente para predizer a produtividade de raízes em mandioca. A ausência de correlação entre produtividade de raízes e algumas características fisiológicas avaliadas após a indução do déficit hídrico, a exemplo da taxa relativa de expansão foliar, índice relativo de clorofilas e a eficiência quântica potencial dos fotossistemas II (Fv/Fm), conforme previamente relatado por Aidar et al. (2015), pode explicar o fato destas características em conjunto serem pobres preditoras da produtividade em mandioca. Em trigo, Lopes et al., (2012) também observaram que o teor de clorofila dentre os diferentes ambientes avaliados, raramente apresentou alguma contribuição significativa para explicar a produtividade de grãos.

Adicionalmente, a análise de características fisiológicas e agronômicas de forma conjunta, resultou em aumento de 76% e 47% no valor médio do r^2 nos ensaios irrigados e sob déficit, respectivamente quando características agronômicas avaliadas no final de ciclo como: altura da planta; Área abaixo da curva de progressão do crescimento das plantas; número de raízes por planta; teor de matéria seca da raiz e produtividade da parte aérea, foram incluídas no modelo juntamente com as características agronômicas. Este aumento foi ainda maior (106%) quando se procedeu a predição apenas com base nas características mais importantes para explicar a produtividade de raízes no experimento irrigado. Quando se realizou a seleção de características nos experimentos com déficit hídrico para todas as variáveis, o aumento do r^2 em relação apenas ao uso de características agronômicas foi semelhante ao modelo sem seleção de variáveis (45%). Por outro lado, mesmo possuindo elevada correlação com a produtividade de raízes (AIDAR et al., 2015; TUMUHIMBISE et al., 2015; SILVA et al., 2016), a inclusão apenas da produtividade da parte aérea na predição conjunta com as características

agronômicas, de modo geral, não resultou em aumentos expressivos no r^2 e na redução do RSME.

O uso de características agronômicas e fisiológicas conjuntamente, tem sido uma alternativa eficiente para predizer a produtividade de grãos sob estresse hídrico, em diferentes culturas como o arroz; a cevada e trigo (ALLAH et al., 2010; VAEZI et al., 2010; LOPES et al., 2012; MOHAMMADI et al., 2013). Em trigo, características agronômicas e fisiológicas explicaram cerca de 27% da variação para produtividade em condições de déficit hídrico (LOPES et al., 2012).

A seleção direta para produtividade em ambientes sob estresse é dificultada devido à grande variação climática e da interação genótipo × ambiente, que resulta em baixa herdabilidade para muitas características produtivas. Por outro lado, um conjunto de características fisiológicas e agronômicas mais estritamente correlacionadas com maior eficiência no uso da água e alta produtividade em ambientes sob déficit hídrico apresentam grande potencial para serem utilizadas na seleção de genótipos em condições de estresse hídrico (RICHARDS et al., 2002; RICHARDS, 2006).

Possivelmente, a análise conjunta de características fisiológicas e agronômicas foi mais eficiente para a predição da produtividade de raízes, devido à integração da base fisiológica da tolerância à seca com componentes agronômicos mais diretamente associados a parâmetros produtivos (OKOGBENIN et al., 2013). Após a seleção das variáveis fisiológicas e agronômicas mais importantes pelos métodos de classificação e regressão, o número de raízes por planta e a produtividade de parte aérea (Fisio+Agro-Sel) apresentaram maior importância para a predição da produtividade, em ambiente irrigado. De fato, o número de raízes por planta possui alta herdabilidade e correlação com a produtividade de raízes de mandioca, em ambientes normais de irrigação (AVIJALA et al., 2015; OLIVEIRA et al., 2015). No entanto, para a condição sob déficit hídrico, além destas duas características, AACP.IAF e NF.8 também foram consideradas importantes e assim inseridas no grupo Fisio+Agro-Sel.

Em ambientes com restrição hídrica, os modelos de predição devem considerar o índice de área foliar e características que indicam a capacidade de

biomassa dos genótipos, como produtividade da parte aérea, número de folhas, e número total de raízes. Estudos vem demonstrando que características relacionadas a biomassa da parte aérea e da raiz, diâmetro de raiz e densidade de ramificação, apresentam potencial para serem usados como preditores da eficiência no uso da água e de nutrientes no campo, para realizar a seleção de genótipos de mandioca precocemente, durante o estágio de crescimento juvenil (ADU et al., 2018). No entanto, estudos adicionais serão necessários para relacionar as características da raiz de mandioca em estágio juvenil com o desempenho de plantas adultas cultivadas em campo com relação à tolerância seca, eficiência do uso de nutrientes e produtividade. Para condições favoráveis de cultivo, nossos resultados indicam que o uso apenas de variáveis agronômicas é capaz de resultar em adequada predição da produtividade de raízes.

Desempenho dos modelos de classificação e predição para produtividade de raízes

Os métodos de classificação PLS, CART e curva ROC foram eficientes em determinar conjuntamente, o grupo de variáveis com maior importância para os modelos de predição, a partir das características fisiológicas e agronômicas. De modo geral esta seleção prévia das características de maior peso, resulta em redução no tempo necessário para a análise dos dados, por simplificar os modelos, além de apresentar um efeito prático de extrema importância que é a priorização de variáveis a serem coletadas no campo, sobretudo visando reduzir os custos da fenotipagem.

Os modelos CART, ANN, SVM, GLMSS e PLS, com exceção do ELM, têm sido frequentemente utilizados na predição agrícola (PARK et al., 2005; MUCHERINO et al., 2009; RUß, 2009; WEBER et al., 2012; MEHMOOD et al., 2012). Determinar qual modelo de predição possibilita mais capacidade preditiva na seleção de genótipos mais produtivos sob déficit hídrico, é de suma importância para programas de melhoramento genético visando maior tolerância a seca em mandioca. Dentre os modelos avaliados, os que apresentaram maior capacidade preditiva ($r^2 > 0{,}75$), melhor ajuste dos entre os valores observados e preditos a reta de regressão e menor variação foram o

GLMSS, ELM e PLS, para o Fisio+Agro-Sel (variáveis fisiológicas e agronômicas selecionadas) na condição irrigada (Figura 4). Além disso, os valores do RMSE foram baixo (entre 0,49 e 0,51) nestes modelos, indicando menores desvios dos dados experimentais para a predição da produtividade de raízes. Já nos experimentos sob déficit hídrico para o mesmo conjunto de dados (Fisio+Agro-Sel), os modelos ELM, PLS, GLMSS e SVM apresentaram a maior capacidade preditiva ($r^2 > 0{,}56$), embora com magnitudes intermediarias em relação aos experimentos irrigados. Provavelmente os valores intermediários de r^2 em condição de déficit hídrico são devidos principalmente, à grande variação climática em ambientes submetidos a estresses abióticos.

Os resultados do presente trabalho corroboram os relatos de RUß (2009), na qual demonstrou que o modelo SVM foi mais acurado em relação aos modelos ANN e CART, sendo recomendado para fins de predição agrícola. Por outro lado, o modelo ELM vem sendo bastante utilizado em pesquisas relacionadas a hidrologia e climatologia, para a predição de dados de evapotranspiração (ABDULLAH et al., 2015; YIN et al., 2017), temperatura diária do ponto de condensação (MOHAMMADI et al., 2015), e predição do índice de seca (DEO & SAHIN, 2015). Esse modelo apresenta a vantagem de ser simples em sua aplicação, com alta velocidade de análise e possuir uma boa capacidade de generalização (ABDULLAH et al., 2015), além de ser, geralmente mais eficiente quando comparado a outros modelos, como o ANN e SVM (OLATUNJI et al., 2014; MOHAMMADI et al., 2015; DEO & SAHIN, 2015).

Adicionalmente o modelo PLS foi bastante interessante como um dos modelos de maior estabilidade da capacidade preditiva, independente do regime hídrico. De fato, o PLS tem amplos usos em diversas culturas, como arroz, soja, trigo, milho e cevada com a finalidade de predição do rendimento agrícola (HANSEN et al., 2002; LIN et al., 2012; WEBER et al., 2012; CHRISTENSON et al., 2016). Em um estudo similar com cultura do milho, considerando condições de déficit hídrico e irrigada, o modelo PLS foi eficaz em predizer a produtividade de grãos e em explicar a variabilidade entre ambientes com diferentes condições hídricas (WEBER et al., 2012). Apesar do modelo ANN ter um bom desempenho em outros estudos, mostrando-se superior a modelos com base em regressão linear múltipla para predição da

produtividade em diversas espécies (KAUL et al., 2005; JI et al., 2007), seu uso na predição da produtividade de raízes de mandioca, juntamente com o modelo CART, apresentou menor capacidade preditiva em relação aos outros modelos. Nossos resultados indicam que os modelos ELM, PLS, GLMSS, foram portanto, os mais promissores para predição da produtividade de raízes de mandioca.

Perspectivas para o melhoramento de mandioca

Em um programa de melhoramento genético para a seleção de genótipos sob condições de estresse, é essencial estabelecer protocolos de fenotipagem baseados em características geneticamente associadas à produtividade sob estresse, que possuam alta herdabilidade, facilidade e agilidade de mensuração no campo, e ainda possuam estabilidade ao longo dos períodos de avaliação. A seleção de variáveis mais informativas possibilita a obtenção de uma economia de tempo e recursos com a fenotipagem de um número reduzido de características capazes de predizer a produtividade de raízes com maior eficiência. Assim, o presente estudo buscou selecionar características fisiológicas e agronômicas mais estritamente relacionadas à produtividade de raízes em diferentes genótipos de mandioca. Portanto, as características fisiológicas mais importantes para a predição do potencial produtivo dos genótipos de mandioca avaliados foram Área abaixo da curva de progressão da expansão das folhas com base no índice de área foliar (AACP.IAF) e Número de folhas mensurado no oitavo mês (NF.8); enquanto que as características agronômicas mais importantes foram Número de raízes por planta (NRP) e Produtividade da parte aérea (PPA).

A capacidade de predição da produtividade tanto em condições irrigadas quanto sob estresse hídrico, pode ser utilizada para selecionar genótipos superiores para futuros cruzamentos, aumentando assim o ganho de seleção por unidade de tempo. Portanto, os modelos GLMSS; PLS e ELM se mostraram promissores como técnicas preditivas acurada e com potencial para predição da produtividade de raízes em genótipos de mandioca. No entanto, os

resultados obtidos no presente estudo baseiam-se em um local e período especifico (2012/13 e 2013/14), devendo ser complementado com outros estudos considerando vários locais e anos de avaliação, a fim de avaliar a robustez e a estabilidade do poder preditivo dos modelos aqui estabelecidos.

CONCLUSÃO

Embora as variáveis fisiológicas sejam importantes indicadores da tolerância à seca na mandioca, quando avaliadas isoladamente, não foram suficientes para predizer a produtividade de raízes na cultura, sendo necessária a inclusão de dados agronômicos obtidos em períodos tardios de avaliação dos experimentos, para o estabelecimento de modelos de predição mais acurados.

As características produtividade da parte aérea, número de raízes, número de folhas e índice de área foliar apresentaram alta importância nos ambientes de sequeiro, enquanto que em ambientes irrigados apenas as duas primeiras tiveram maior importância relativa. Dentre as diferentes características analisadas nos modelos de predição avaliados, os modelos ELM, PLS e GLMSS, apresentaram maior potencial para predição da produtividade na cultura da mandioca. Apesar de não ter sido possível treinar os modelos apenas com características fisiológicas ou agronômicas avaliadas precocemente, os resultados apresentados contribuem para a seleção de genótipos com melhor desempenho no rendimento de raízes, tanto na condição irrigada quanto em condição não irrigada, com economia de tempo na fenotipagem dos dados de campo.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDULLAH, S.S.; MALEK, M.A.; ABDULLAH, N.S.; KISI, O.; YAP, K.S. Extreme learning machines: a new approach for prediction of reference evapotranspiration. **Journal of Hydrology**, v.527, p.184-195, 2015.

ADU, M.O.; ASARE, P.A.; ASARE-BEDIAKO, E.; AMENORPE, G.; ACKAH, F.K.; AFUTU, E.; AMOAH, M.N.; YAWSON, D.O. Characterising shoot and root

system trait variability and contribution to genotypic variability in juvenile cassava (*Manihot esculenta* Crantz) plants. **Heliyon**, v.4, n.6, p.1-24, 2018.

AFONSO, A.M.; EBELL, M.H.; GONZALES, R.; STEIN. J.; GENTON, B.; SENN, N. The use of classification and regression trees to predict the likelihood of seasonal influenza. **Family Practice**, v.29, n.6, p.671–677, 2012.

AIDAR, S.T.; MORGANTE, C.V.; CHAVES, A.R.M.; COSTA NETO, B.P.; VITOR, A.B.; MARTINS, D.R.P.S.; SILVA, R.; CRUZ, J.L.; OLIVEIRA, E.J. Características fisiológicas, produção total de raízes e de parte aérea em acessos de *Manihot esculenta* em condições de déficit hídrico. **Revista Brasileira de Geografia Física Número Especial IV SMUD**, v.8, p.685-696, 2015.

AINA, O.O.; DIXON, A.G.; AKINRINDE, E.A. Effect of soil moisture stress on growth and yield of cassava in Nigeria. **Pakistan Journal of Biological Sciences: PJBS**, v.10, n.18, p.3085-3090, 2007.

ALLAH, A.A.A.; AMMAR, M.H.; BADAWI, A.T. Screening rice genotypes for drought resistance in Egypt. **Journal of Plant Breeding and Crop Science**, v.2, n.7, p.205-215, 2010.

ALVES, A.A.C.; SETTER, T.L. Abscisic acid accumulation and osmotic adjustment in cassava under water deficit. **Environmental and Experimental Botany**, v.51, n.3, p.259-271, 2004.

ANDERSEN, C. M.; BRO, R. Variable selection in regression - a tutorial. **Journal of Chemometrics**, v.24, n.11-12, p.728-737, 2010.

AVIJALA, M.F.; BHERING, L.L.; PEIXOTO, L.A.; CRUZ, C.D.; CARNEIRO, P.C.S.; CUAMBE, C.E.; ZACARIAS, A. Evaluation of cassava (*Manihot esculenta* Crantz) genotypes reveals great genetic variability and potential

selection gain. **Australian Journal of Crop Science**, v.9, n.10, p.940-947, 2015.

BERGANTIN, R.V.; YAMAUCHI, A.; PARDALES JR, J.R.; BOLATETE JR, D.M. Screening cassava genotypes for resistance to water deficit during crop establishment. **Philippine Journal of Crop Science**, v.29, n.1, p.29-39, 2004.

BEYENE, Y.; SEMAGN, K.; MUGO, S.; TAREKEGNE, A.; BABU, R.; et al. Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. **Crop Science**, v.55, n.1, p.154-163, 2015.

CABELLO, R.; MONNEVEUX, P.; BONIERBALE, M.; KHAN, M.A. Heritability of yield components under irrigated and drought conditions in andigenum potatoes. **American Journal of Potato Research**, v.91, n.5, p.492-499, 2014.

CAMPBELL, C.L.; MADDEN, L.V. **Introduction to Plant Disease Epidemiology**. John Wiley & Sons, 1990.

CEBALLOS, H.; OKOGBENIN, E.; PÉREZ, J.C.; LÓPEZ-VALLE, L.A.B.; DEBOUCK, D. Cassava. In: root and tuber crops. **Springer**, p.53-96, 2010.

CEBALLOS, H.; RAMIREZ, J.; BELLOTTI, A. C.; JARVIS, A.; ALVAREZ, E. Adaptation of cassava to changing climates. **Crop Adaptation to Climate Change**, p.411-425, 2011.

CEBALLOS, H.; KULAKOW, P.; HERSHEY, C.; Cassava breeding: current status, bottlenecks and the potential of biotechnology tools. **Tropical Plant Biology**, v.5, p.73-87, 2012.

CHRISTENSON, B.S.; SCHAPAUGH, W.T.; AN, N.; PRICE, K.P.; PRASAD, V.; FRITZ, A.K. Predicting soybean relative maturity and seed yield using canopy reflectance. **Crop Science**, v.56, n.2, p. 625-643, 2016.

CIAT. **International Center for Tropical Agriculture**, 2017. Disponível em: <http://ciat.cgiar.org/what-we-do/breeding-better-crops/rooting-for-cassava/>. Acesso em: dez, 2017.

DAN, Z.; HU, J.; ZHOU, W.; YAO, G.; ZHU, R.; ZHU, Y.; HUANG, W. Metabolic prediction of important agronomic traits in hybrid rice (*Oryza sativa* L.). **Scientific Reports**, v.6, p.1-9, 2016.

DEO, R.C.; ŞAHIN, M. Application of the extreme learning machine algorithm for the prediction of monthly effective drought index in eastern Australia. **Atmospheric Research**, v.153, p.512-525, 2015.

DUQUE, L.O.; SETTER, T.L. cassava response to water deficit in deep pots: root and shoot growth, aba, and carbohydrate reserves in stems, leaves and storage roots. **Tropical Plant Biology**, v.6, n.4, p.199-209, 2013.

EL-SHARKAWY, M.A. Physiological characteristics of cassava tolerance to prolonged drought in the tropics: implications for breeding cultivars adapted to seasonally dry and semiarid environments. **Journal of Plant Physiology**, v.19, p.257-286, 2007.

EL-SHARKAWY, M.A. Stress-tolerant cassava: the role of integrative ecophysiology-breeding research in crop improvement. **Open Journal of Soil Science**, v.2, n.02, p.162-186, 2012.

EMBRAPA SEMIÁRIDO. **Centro de Pesquisa Agropecuária do Trópico Semiárido**. Dados meteorológicos de 2013. Disponível: http://www.cpatsa.embrapa.br:8080/servicos/dadosmet/ceb-anual.html. Acesso em: dez, 2017.

EMBRAPA SEMIÁRIDO. **Centro de Pesquisa Agropecuária do Trópico Semiárido**. Dados meteorológicos de 2014. Disponível:

http://www.cpatsa.embrapa.br:8080/servicos/dadosmet/ceb-anual.html. Acesso em: dez, 2017.

FAO. **Food and Agriculture Organization of the United Nations**. Food outlook: biannual report on global food markets, 2016. Disponível em: <http://www.fao.org/3/a-i6198e.pdf>. Acesso em: dez, 2017.

FAO. **Food and Agriculture Organization of the United Nations**. Save and grow: cassava a guide to sustainable production intensification, 2013. Disponível em: http://www.fao.org/3/a-i2929o.pdf. Acesso em: dez, 2017.

FARFAN, I.D.B.; LA FUENTE, G.N.; MURRAY, S.C.; ISAKEIT, T.; HUANG, P.C.; et al. Genome wide association study for drought, aflatoxin resistance, and important agronomic traits of maize hybrids in the sub-tropics. **Plos One**, v.10, n.2, p.0117737, 2015.

FERRARO, D.O.; RIVERO, D.E.; GHERSA, C.M. An analysis of the factors that influence sugarcane yield in northern Argentina using classification and regression trees. **Field Crops Research**, v.112, n.2, p.149-157, 2009.

HANSEN, P.M.; JØRGENSEN, J.R.; THOMSEN, A. Predicting grain yield and protein content in winter wheat and spring barley using repeated canopy reflectance measurements and partial least squares regression. **The Journal of Agricultural Science**, v.139, n.3, p.307-318, 2002.

HINKOSSA, A.; GEBEYEHU, S.; ZELEKE, H. Generation mean analysis and heritability of drought resistance in common bean (*Phaseolus vulgaris* L.). **African Journal of Agricultural Research**, v.8, n.15, p.1319-1329, 2013.

JI, B.; SUN, Y.; YANG, S.; WAN, J. Artificial neural networks for rice yield prediction in mountainous regions. **The Journal of Agricultural Science**, v.145, n.3, p.249-261, 2007.

KAUL, M.; HILL, R.L.; WALTHALL, C. Artificial neural networks for corn and soybean yield prediction. **Agricultural Systems**, v.85, n.1, p.1-18, 2005.

KAWANO, K.; FUKUDA, W.M.G.; CENPUKDEE, U. Genetic and environmental effects on dry matter content of cassava root 1. **Crop Science**, v.27, n.1, p.69-74, 1987.

LABAN, T.F.; KIZITO, E.B.; BAGUMA, Y.; OSIRU, D. Evaluation of Ugandan cassava germplasm for drought tolerance. **International Journal of Agriculture and Crop Sciences**, v.5, n.3, p.212-226, 2013.

LIN, W.S.; YANG, C.M.; KUO, B.J. Classifying cultivars of rice (*Oryza sativa* L.) based on corrected canopy reflectance spectra data using the orthogonal projections to latent structures (O-PLS) method. **Chemometrics and Intelligent Laboratory Systems**, v.115, p.25-36, 2012.

LIU, J.; ZHENG, Q.; MA, Q.; GADIDASU, K.K.; ZHANG, P. Cassava genetic transformation and its application in breeding. **Journal of Integrative Plant Biology**, v.53, n.7, p.552-569, 2011.

LOPES, M.S.; REYNOLDS, M.P.; JALAL-KAMALI, M.R.; MOUSSA, M.; FELTAOUS, Y.; TAHIR, I.S.A.; BARMA, N.; VARGAS, M.; MANNES, Y.; BAUM, M. The yield correlations of selectable physiological traits in a population of advanced spring wheat lines grown in warm and drought environments. **Field Crops Research**, v. 128, p. 129-136, 2012.

MEHMOOD, T.; LILAND, K. H.; SNIPEN, L.; SOLVE, S. A review of variable selection methods in partial least squares regression. **Chemometrics and Intelligent Laboratory Systems**, v.118, p.62-69, 2012.

MOHAMMADI, K.; SHAMSHIRBAND, S.; MOTAMEDI, S.; PETKOVIĆ, D.; HASHIM, R.; GOCIC, M. Extreme learning machine based prediction of daily

dew point temperature. **Computers and Electronics in Agriculture**, v.117, p.214-225, 2015.

MOHAMMADI, R.; HEIDARI, B.; HAGHPARAST, R. Traits associated with drought tolerance in spring durum wheat (*Triticum turgidum* L. var. *durum*) breeding lines from international germplasm. **Crop Breeding Journal**, v.3, n.2, p.87-98, 2013.

MORANTE, N.; SÁNCHEZ, T.; CEBALLOS, H.; CALLE, F.; PÉREZ, J.C.; EGESI, C.; CUAMBE, C.E.; ESCOBAR, A.F.; ORTIZ, D.; CHAVEZ, A.L; FREGENE, M. Tolerance to postharvest physiological deterioration in cassava roots. **Crop Science**, v.50, n.4, p.1333-1338, 2010.

MUCHERINO, A.; PAPAJORGJI, P.; PARDALOS, P. M. A survey of data mining techniques applied to agriculture. **Operational Research**, v.9, n.2, p.121-140, 2009.

OKOGBENIN, E.; SETTER, T.L.; FERGUSON, M.; MUTEGI, R.; CEBALLOS, H.; OLASANMI, B.; FREGENE, M. Phenotypic approaches to drought in cassava: review. **Frontiers in Physiology**, v.4, p.1-15, 2013.

OLATUNJI, S.O.; SELAMAT, A.; ABDULRAHEEM, A. A hybrid model through the fusion of type-2 fuzzy logic systems and extreme learning machines for modelling permeability prediction. **Information Fusion**, v.16, p.29-45, 2014.

OLIVEIRA, E.J.; AIDAR, S.T.; MORGANTE, C.V.; CHAVES, A.R.M.; CRUZ, J.L.; COELHO FILHO, M.A. Genetic parameters for drought-tolerance in cassava. **Pesquisa Agropecuária Brasileira**, v.50, n.3, p.233-241, 2015.

OLIVEIRA, E.J.; MORGANTE, C.V.; AIDAR, S.T.; CHAVES, A.R.M.; ANTONIO, R.P.; CRUZ, J.L.; COELHO FILHO, M.A. Evaluation of cassava germplasm for drought tolerance under field conditions. **Euphytica**, v.213, n.8, p.188-208, 2017.

PARK, S.J.; HWANG, C.S.; VLEK, P.L.G. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. **Agricultural Systems**, v.85, n.1, p.59–81, 2005.

R CORE TEAM. R: a language and environment for statistical computing. **R Foundation for Statistical Computing**. Disponível em: <URL https://www.R-project.org/>, 2018.

RICHARDS, R.A.; REBETZKE, G.J.; CONDON, A.G.; VAN-HERWAARDEN, A.F. Breeding opportunities for increasing the efficiency of water use and crop yield in temperate cereals. **Crop Science**, v.42, n.1, p.111-121, 2002.

RICHARDS, R.A. Physiological traits used in the breeding of new cultivars for water-scarce environments. **Agricultural Water Management**, v.80, n.1-3, p.197-211, 2006.

RUß, G. Data mining of agricultural yield data: a comparison of regression models. **Industrial Conference on Data Mining**, p.24–37, 2009.

SILVA, R.S.; MOURA, E.F.; FARIAS-NETO, J.T.; SAMPAIO, J.E. Genetic parameters and agronomic evaluation of cassava genotypes. **Pesquisa Agropecuária Brasileira**, v.51, n.7, p.834-841, 2016.

TUMUHIMBISE, R.; SHANAHAN, P.; MELIS, R.; KAWUKI, R. Genetic variation and association among factors influencing storage root bulking in cassava. **The Journal of Agricultural Science**, v.153, n.7, p.1267-1280, 2015.

VAEZI, B.; BAVEI, V.; SHIRAN, B. Screening of barley genotypes for drought tolerance by agro-physiological traits in field condition. **African Journal of Agricultural Research**, v.5, n.9, p.881-892, 2010.

WEBER, V.S.; ARAUS, J.L.; CAIRNS, J.E.; SANCHEZ, C.; MELCHINGER, A. E.; ORSINI, E. Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes. **Field Crops Research**, v.128, p.82-90, 2012.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v.58, n.2, p.109-130, 2001.

YIN, Z.; FENG, Q.; YANG, L.; DEO, R.C.; WEN, X.; SI, J.; XIAO, S. Future projection with an extreme-learning machine and support vector regression of reference evapotranspiration in a mountainous inland watershed in north-west China. **Water**, v.9, n.11, p.880, 2017.

MATERIAL SUPLEMENTAR

Tabela S1 - 49 acessos de mandioca avaliados sob condição irrigada e não irrigada.

Genótipos	Tipo	Reação*	Razão de seleção	Origem País/Estado
9624-09	Melhorado	D	Alta retenção de folhas	Brasil/Bahia
BGM-0089	Variedade Local	D	Alta retenção de folhas	Colômbia/Valle
BGM-0096	Variedade Local	D	Coleção Semiárido	Brasil/-
BGM-0116	Variedade Local	T	Coleção Semiárido	Brasil/Bahia
BGM-0163	Variedade Local	D	Coleção Semiárido	Brasil/Bahia
BGM-0279	Variedade Local	D	Alta retenção de folhas	Brasil/Bahia
BGM-0331	Melhorado	D	Alta retenção de folhas	Colômbia/Valle
BGM-0360	Melhorado	D	Alta retenção de folhas	Colômbia/Valle
BGM-0541	Variedade Local	D	Alta retenção de folhas	Brasil/Bahia
BGM-0598	Variedade Local	T	Alta retenção de folhas	Brasil/Rio Grande do Sul
BGM-0785	Variedade Local	D	Alta retenção de folhas	Brasil/Bahia
BGM-0815	Variedade Local	D	Coleção Semiárido	Brasil/Alagoas
BGM-0818	Variedade Local	D	Coleção Semiárido	Brasil/ Sergipe
BGM-0856	Variedade Local	D	Coleção Semiárido	Brasil/Sergipe
BGM-0876	Variedade Local	S	Alta retenção de folhas	Brasil/Pará
BGM-0908	Variedade Local	S	Alta retenção de folhas	Colômbia/Valle
BGM-1171	Variedade Local	D	Alta retenção de folhas	Brasil/Pará
BGM-1195	Variedade Local	D	Alta retenção de folhas	Brasil/ -
BGM-1482	Variedade Local	D	Coleção Semiárido	Brasil/Bahia
BGM-2020	Variedade Local	D	Alta retenção de folhas	Brasil/Bahia
Branquinha	Variedade Local	D	Variedade produtiva	Brasil/Pernambuco
BRS A. Burro	Melhorado	T	Tolerante à seca	Brasil/Piauí
BRS Dourada	Melhorado	D	Variedade produtiva	Brasil/Bahia
BRS Formosa	Melhorado	T	Tolerante à seca	Brasil/Bahia
BRS G. Ovo	Melhorado	T	Tolerante à seca	Brasil/Amazonas
BRS Kiriris	Melhorado	T	Tolerante à seca	Brasil/Bahia
Cacau	Variedade Local	S	Alta retenção de folhas	Brasil/Pernambuco
Cachimbo	Variedade Local	S	Alta retenção de folhas	Brasil/Pernambuco
Do Céu	Variedade Local	T	Tolerante à seca	Brasil/Pernambuco
E. Ladrão	Variedade Local	T	Tolerante à seca	Brasil/Piauí
Eucalipto	Variedade Local	D	Alta retenção de folhas	Brasil/Paraná
GCP-001	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-009	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-014	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-020	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-025	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-043	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-046	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-095	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-128	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-179	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-190	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-194	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-227	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-374	Melhorado	T	Tolerante à seca	Colômbia/Valle
Mani Branca	Melhorado	D	Alta retenção de folhas	Brasil/Paraná
NG-310	Melhorado	D	Alta retenção de folhas	Brasil/Distrito Federal
Paulo Rosa	Variedade Local	S	Alta retenção de folhas	Brasil/Bahia
Sacai	Variedade Local	T	Tolerante à seca	Brasil/Bahia

*D – Desconhecido; S – Susceptível; T – Tolerante;

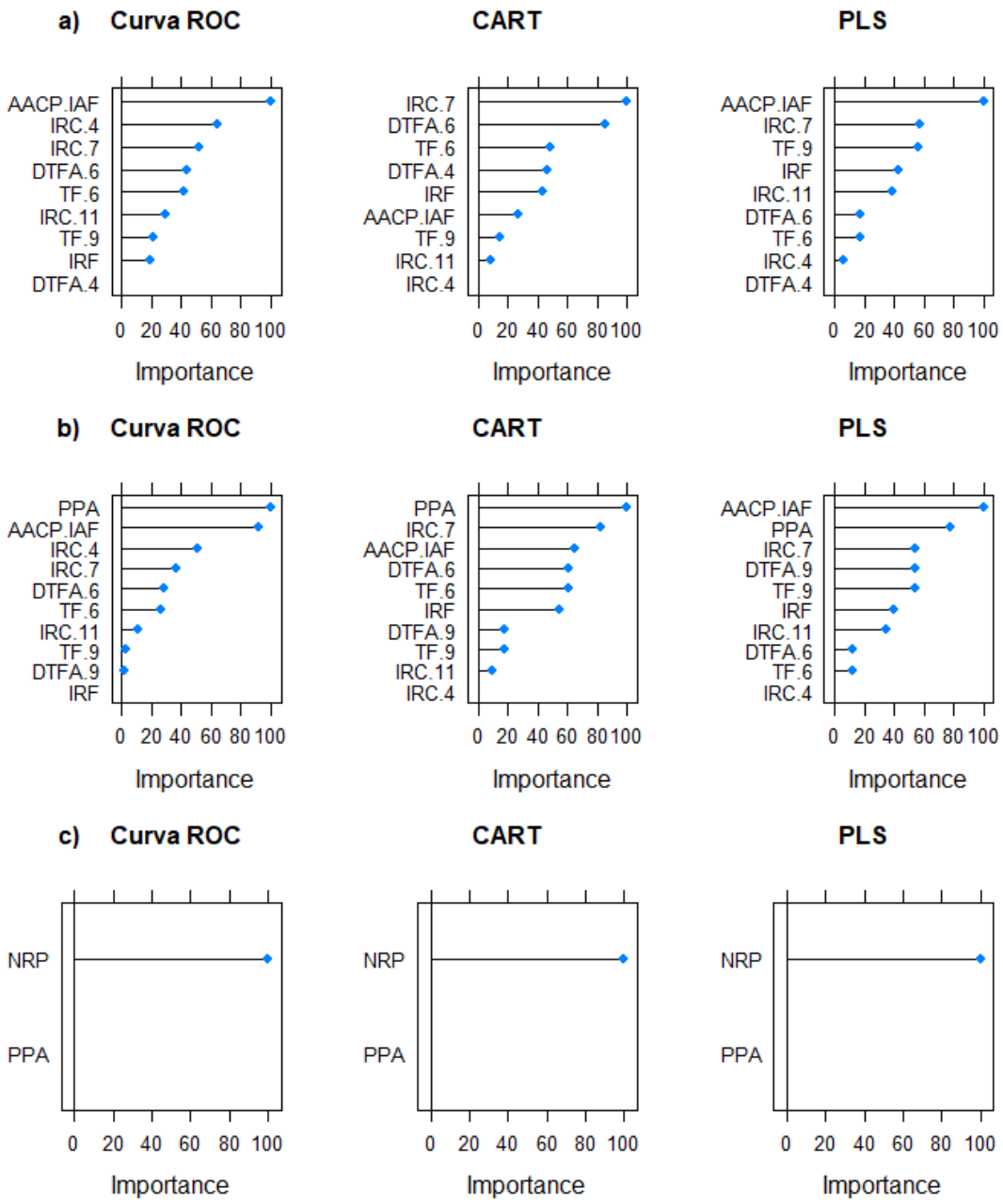

Figura S1 – Importância das variáveis fisiológicas e agronômicas selecionadas, para a condição irrigada pelos métodos curva de ROC (*Receiver Operating Characteristics*), CART (*Classification and Regression Trees*) e PLS (*Partial Least Squares*); a) características fisiológicas mais importantes; b) características fisiológicas mais importantes, com adição da produtividade de parte aérea; e c) características fisiológicas e agronômicas mais importantes).

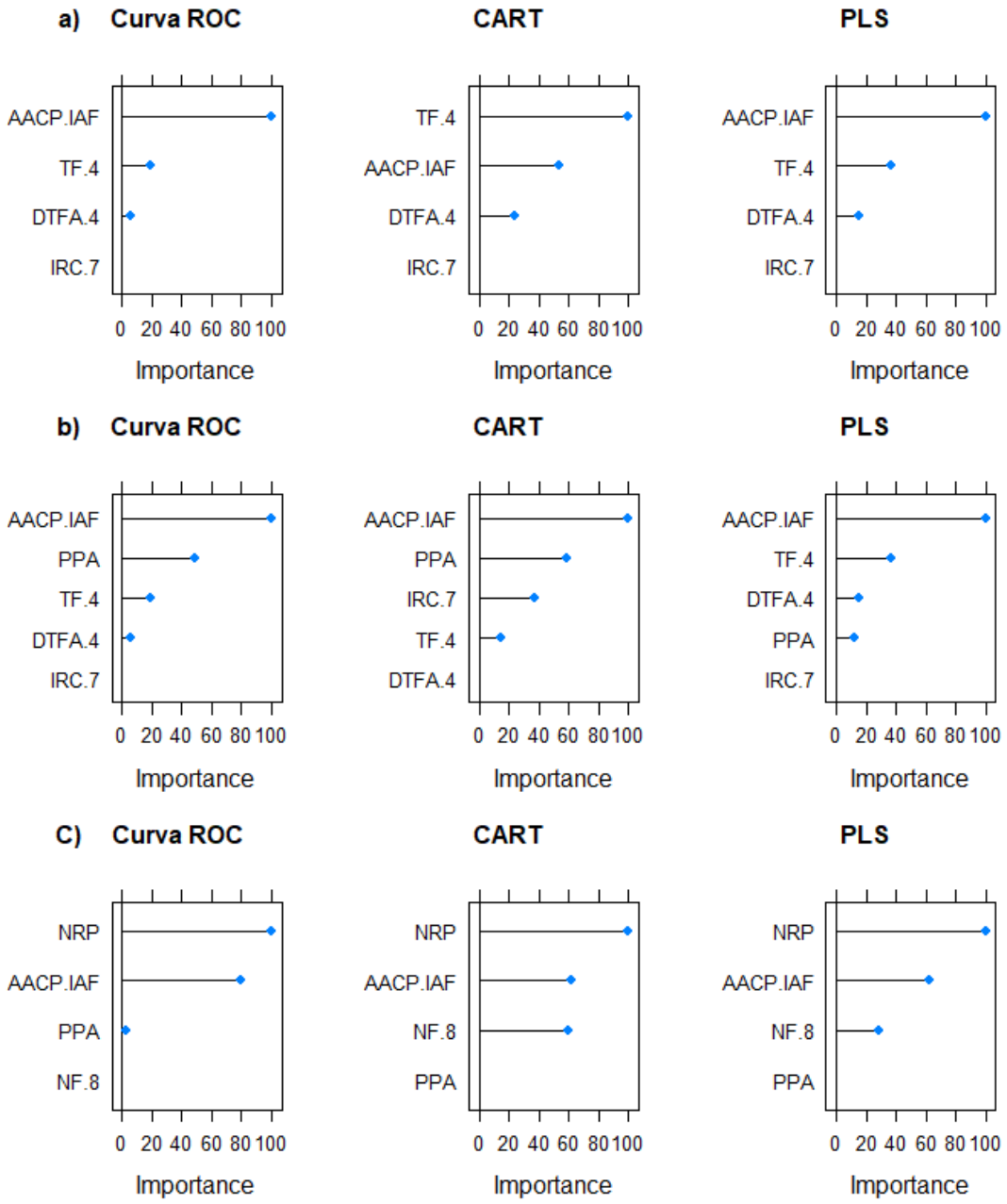

Figura S2- Importância das variáveis fisiológicas e agronômicas selecionadas, para a condição não irrigada pelos métodos curva de ROC (*Receiver Operating Characteristics*), CART (*Classification and Regression Trees*) e PLS (*Partial Least Squares*); a) características fisiológicas mais importantes; b) características fisiológicas mais importantes, com adição da produtividade de parte aérea; e c) características fisiológicas e agronômicas mais importantes).

ARTIGO 2

ASSOCIAÇÃO GENÔMICA AMPLA (GWAS) PARA TOLERÂNCIA AO DÉFICIT HÍDRICO EM MANDIOCA

[2]Artigo a ser ajustado para posterior submissão ao Comitê Editorial do periódico científico PlosOne, em versão na língua inglesa.

Associação genômica ampla (GWAS) para tolerância ao déficit hídrico em mandioca

Resumo: Apesar de ser considerada uma espécie tolerante à seca, a mandioca (*Manihot esculenta* Crantz) tem seu crescimento e produtividade reduzidos em ambiente de estresse hídrico prolongado. Portanto, o objetivo deste estudo foi identificar regiões genômicas associadas ao déficit hídrico em mandioca por meio da associação genômica ampla (GWAS). Foram avaliados 49 genótipos de mandioca em duas condições hídricas: irrigada (controle) e sob déficit hídrico. As características avaliadas foram: produtividade total de raízes (PTR), da parte aérea (PPA) e de amido (PAMD), teor de matéria seca nas raízes (DMC), índice de tolerância a seca (DTI) e índice de estabilidade da tolerância a seca (DTSI). A GWAS foi realizada utilizando o modelo linear misto múltiplo (MLMM), com inclusão da matriz de parentesco e estrutura populacional, para cada condição hídrica e característica. As estimativas de herdabilidade no sentido amplo (h^2) foram bastante variáveis em função das características em análise e dos ambientes. Foram identificadas 54 associações marcador-fenótipo para as características avaliadas, sendo 48 SNPs distribuídos em todos os 18 cromossomos da mandioca. Além disso, foi possível identificar marcadores específicos e estáveis ao longo dos ambientes. Os SNPs identificados estão próximos a 120 transcritos, dos quais 90 foram previamente descritos e 24 possuem anotação funcional conhecida. Alguns desses transcritos estão relacionados a proteínas envolvidas com a tolerância a seca, tais como: domínio Apetala 2 (AP2), proteína potenciadora de oxigênio do fotossistema II, zíper de leucina e fator de transcrição BZIP. Existe grande potencial de aplicação destes SNPs na seleção assistida por marcadores para o desenvolvimento de novas variedades de mandioca tolerantes ao déficit hídrico.

Palavras-chave: *Manihot esculenta*, mapeamento associativo, SNP, tolerância à seca.

Genome-wide association study (GWAS) for drought-tolerance in cassava

Abstract: Cassava (*Manihot esculenta* Crantz) is considered a drought tolerant species, however, its growth and productivity are reduced in environments with prolonged water stress. Therefore, the objective of this study was to identify genomic regions associated to water deficit in cassava using genome-wide association study (GWAS). A total of 49 cassava genotypes were evaluated in two water conditions: irrigated (control) and under water deficit. The evaluated characteristics were: fresh root yield (RoY), shoot yield (ShY), and starch yield (StY), dry matter content in the roots (DMC), dry tolerance index (DTI) and dry tolerance stability index (DTSI). GWAS was performed using the multiple mixed linear model (MLMM), including kinship matrix and population structure for each hydric condition for each characteristic. The estimates of heritability in the broad sense (h^2) were variable in function of the traits and the environments analyzed We identified 54 marker-phenotype associated with the traits evaluated, in which 48 SNPs were distributed in all 18 cassava chromosomes. In addition, we identified specific and stable markers throughout the environments. The identified SNPs are close to 120 transcripts, of which 90 have been previously described and 24 have known functional annotation. Some of these transcripts are related to proteins involved with drought tolerance, such as: Apetala 2 domain (AP2), oxygen potentiating protein of photosystem II, leucine zipper, and transcription factor BZIP. There is great potential for the application of these SNPs in the selection of markers for the development of new varieties of cassava tolerant to water deficit.

Keywords: *Manihot esculenta*, associative mapping, SNP, drought tolerance.

INTRODUÇÃO

A mandioca (*Manihot esculenta* Crantz) possui enorme importância socioeconômica por ser a terceira maior fonte de alimento, perdendo apenas para o arroz e o milho, constituindo a base alimentar de cerca de 800 milhões de pessoas, principalmente em regiões tropicais na América Latina, Ásia e África (CEBALLOS et al., 2010; LIU et al., 2011; CIAT, 2017). A mandioca é cultivada especialmente por pequenos agricultores, na maioria localizados em países tropicais situados na região equatorial, entre 30˚ norte e 30˚ sul do equador, e altitude variando de 0 a 2000 metros, sob precipitação anual de 500 mm (em zonas semiáridas) a mais de 2000 mm (zonas ecológicas úmidas), o que indica sua ampla adaptabilidade a diversos ambientes e ecossistemas de cultivo (EL-SHARKAWY, 2012; OKOGBENIN et al., 2013). Essa plasticidade fenotípica também se reflete na sua tolerância ao déficit hídrico, uma vez que a mandioca é capaz de apresentar produtividades expressivas, em comparação com outras culturas, mesmo em condições de baixa precipitação e em solos de baixa fertilidade (EL-SHARKAWY, 2007; OKOGBENIN et al., 2013).

O déficit hídrico é considerado o estresse abiótico com maior impacto na agricultura, por interferir diretamente no crescimento e desenvolvimento das plantas (CATTIVELLI et al., 2008). O cultivo da mandioca em condições de déficit hídrico aliado a estratégias inadequadas de manejo, não utilização de defensivos e insumos agrícolas, bem com o uso de variedades com baixo potencial de produção são fatores que reduzem o potencial produtivo da cultura (OLIVEIRA et al., 2015). Esse cenário é observado em regiões semiáridas do Nordeste brasileiro, já que o rendimento médio da raiz é de 9,5 $t.ha^{-1}$ em comparação com os 23,6 $t.ha^{-1}$ obtidos por alguns genótipos sob condições experimentais de estresse hídrico (IBGE, 2016; OLIVEIRA et al., 2015). Além disso, El-Sharkawy (2012), verificou que a mandioca possui adaptabilidade a prolongados períodos de estresse hídrico, e em muitas situações com razoável produtividade de raízes após o final do período de estresse. Por exemplo, mesmo sendo cultivada em ambientes com déficit hídrico e baixas temperaturas durante o inverno, algumas cultivares de mandioca apresentam rendimento radicular de até 66 $t.ha^{-1}$ na bacia do rio Limpopo na África do Sul

(OGOLA & MATHEWS, 2011). Portanto, existe uma importante variabilidade genética na espécie cultivada (*M. esculenta*) que pode permitir o aumento do potencial produtivo de mandioca para o cultivo em regiões semiáridas, por meio do melhoramento e seleção de genótipos mais tolerantes ao déficit hídrico.

A tolerância ao déficit hídrico é uma característica quantitativa complexa, regulada por vários genes, o que dificulta o processo de seleção em condições de campo (OKOGBENIN et al., 2013). Por outro lado, a identificação de regiões genômicas envolvidas na resposta ao estresse, o entendimento do controle genético e o desenvolvimento de ferramentas de seleção assistida por marcadores, pode contribuir para melhorar o processo de seleção fenotípica e com isso reduzir o tempo necessário para o desenvolvimento de novas variedades de mandioca com tolerância ao déficit hídrico.

Alguns esforços foram e continuam sendo empreendidos com o objetivo de identificar regiões ligadas a características complexas em mandioca, por meio do mapeamento de QTLs (*Quantitative Trait Loci*) (MASUMBA et al., 2017; SEDANO et al., 2017). Informações sobre QTLs que regulam a resposta ao déficit hídrico podem ser utilizados para elucidar a base fisiológica da tolerância à seca e auxiliar na seleção de genótipos com maior produtividade em condições de estresse hídrico (TUBEROSA e SALVI, 2006). Entretanto, a análise de QTL possui algumas limitações, onde apenas um número reduzido de QTLs com maior efeito são detectados, em contraste com a natureza poligênica da variação genética total observada para a maioria das características quantitativas (DEKKERS, 2004). Em muitos casos, esta limitação é devida ao fato de que somente a diversidade alélica que segrega entre os parentais das populações segregantes pode ser detectada, e pelo limite na resolução do mapeamento genético como resultado do reduzido número de recombinantes produzidos nas populações segregantes.

Em termos de cobertura genômica, os avanços recentes nas tecnologias genômicas tem permitido a realização de genotipagem em larga escala com base em marcadores SNP (*Single Nucleotide Polymorphism*), em um processo rápido e acessível a diversas espécies. Com uso da informação genômica em larga escala foi possível ampliar o uso de estratégias baseadas no desequilíbrio de ligação (LD), como a associação genômica ampla (GWAS –

Genome wide association studies), para um maior refinamento no mapeamento de características de interesse em nível populacional. A GWAS explora os eventos de recombinação histórica que ocorreram naturalmente através de várias gerações para mapear QTLs (ROSENBERG et al., 2010; KORTE e FARLOW, 2013). De modo geral, a GWAS supera algumas das limitações da análise convencional de QTLs, e ainda permite a identificação dos fenótipos de interesse, fornece *insights* sobre a arquitetura genética da característica e sugere potenciais candidatos para mutagênese e transgenia. Além disso, permite a escolha de parentais para análises de QTLs sendo, portanto, uma estratégia complementar para o mapeamento mais refinado de características quantitativas.

Na cultura do milho, por exemplo, marcadores com associação significativa a tolerância ao déficit hídrico, foram identificados, validados, desenvolvidos e utilizados com eficiência para seleção de genótipos tolerantes (HAO et al., 2010; LIU et al., 2013). Em diversas culturas, esta mesma abordagem tem sido utilizada no entendimento da tolerância ao déficit hídrico com base em diversas características fisiológicas (WEHNER et al., 2015), coeficiente de tolerância à deficiência hídrica (MA et al., 2016), altura de plantas e características de florescimento (FARFAN et al., 2015) e produtividade de grãos (PANTALIÃO et al., 2016). Além disso, a GWAS tem sido uma ferramenta eficiente na identificação de regiões genômicas ligadas a outros estresses abióticos, como a eficiência no uso do nitrogênio (MOROSINI et al., 2017).

Adicionalmente como a tolerância ao estresse hídrico possui forte interação genótipo x ambiente (GxA), os estudos de GWAS com base em experimentos multiambientes propiciam uma vantagem adicional, por possibilitar a detecção de regiões associadas à genes expressos em condições e ambientes específicos, reduzindo com isto o ruído de fatores não genéticos de ambiente, tornando assim os resultados mais robustos e confiáveis (MATHEWS et al., 2008; MALOSETTI et al., 2013, FARFAN et al., 2015; GUTIÉRREZ et al., 2015). Entretanto, de acordo com nosso conhecimento, estudos de GWAS para entendimento do efeito de estresses abióticos ainda não foram explorados na cultura da mandioca. Portanto, o objetivo do presente

estudo foi avaliar um painel de diversidade de acessos de mandioca, em duas condições hídricas para a identificação de regiões genômicas associadas ao déficit hídrico em mandioca por meio da GWAS.

MATERIAL E MÉTODOS

Dados fenotípicos

Para a GWAS foi analisado um grupo constituído de 49 genótipos de mandioca contrastantes, sendo 25 variedades locais (coletadas em regiões semiáridas) e 24 variedades melhoradas com histórico de tolerância à seca, por terem sido selecionadas em condições de seca (Material suplementar, Tabela S1). As variedades de mandioca foram avaliadas em dois anos agrícolas (2012/2013 e 2013/2014), e em duas condições hídricas, com irrigação (IN) e sob déficit hídrico (DH). Em ambas as condições, foi utilizado um delineamento em blocos completos casualizados (DBCC), com três repetições, com dez plantas por parcela (duas linhas com 5 plantas), espaçamento de 0,90 m entre linhas e 0,80 entre plantas. O plantio foi realizado com manivas de 16 cm, seguindo as recomendações e práticas agrícolas para a cultura.

Até o quarto mês após o plantio, todos os seis blocos foram submetidos à irrigação por gotejamento (4 $L.h^{-1}$), sendo que a lâmina de água aplicada foi calculada em função da evapotranspiração da planta, estimada pelos dados meteorológicos fornecidos pela estação meteorológica da Embrapa Semiárido instalada no campo experimental. Em seguida, a irrigação foi suspensa até a colheita (12º mês após o plantio), nos três blocos destinado à aplicação do estresse hídrico e mantida nos outros três blocos.

Os experimentos foram realizados na Estação Experimental de Bebedouro, da Embrapa Semiárido, Petrolina - PE (9°22’ de latitude Sul, 40°22' de longitude Oeste e altitude de 376 m). O local de avaliação apresenta clima semiárido e foi escolhido por apresentar baixa precipitação pluviométrica anual. As variações climáticas durante os períodos avaliados foram monitoradas pela estação meteorológica instalada no campo experimental. O ano 2013 apresentou precipitação anual média de 347,8 mm, umidade relativa do ar variando entre 48 e 61% e temperatura média entre 27,7 e 29,2 ºC. Em 2014, a

precipitação anual média foi de 216,3 mm, umidade relativa do ar entre 55 e 67% e temperatura média variando entre 24,5 e 26,9 ºC (EMBRAPA SEMIÁRIDO, 2013; EMBRAPA SEMIÁRIDO, 2014). As safras de 2012/13 e 2013/14 foram marcadas por condições meteorológicas com baixo volume de precipitação, evidenciando a ocorrência de seca, principalmente no período entre os meses de maio a outubro.

As colheitas foram realizadas no 12º mês após o plantio e em seguida foram avaliados as características de produtividade da parte aérea (PPA em $t.ha^{-1}$), determinada pela pesagem da parte aérea das plantas, a partir do corte realizado a 10 cm da superfície do solo; produtividade total de raízes (PTR em $t.ha^{-1}$), determinada pela pesagem de todas as raízes da planta; teor de matéria seca nas raízes (DMC em %), e produtividade de amido (PAMD em $t.ha^{-1}$), obtida pela multiplicação do teor de amido e produtividade total de raízes frescas. Tanto o teor de matéria seca nas raízes quanto o teor de amido foram obtidos com o peso especifico das raízes de acordo com Kawano et al. (1987).

Dados genotípicos

O DNA do painel de diversidade de mandioca foi extraído utilizando-se o protocolo CTAB (brometo de cetiltrimetilamônio) descrito por Doyle e Doyle (1987), com algumas modificações, à exemplo do aumento da concentração de 2-mercaptoetanol a 0,4% e adição de polivilpirrolidona (PVP). A quantificação foi realizada em gel de agarose 1,0% (p/v) corado com brometo de etídeo (1,0 $mg.mL^{-1}$) utilizando como padrão uma série de concentrações do fago Lambda (Invitrogen). O DNA genômico foi ajustado para concentração final de 20 $ng.ul^{1}$.

Em seguida, as amostras de DNA foram genotipadas no *Genomic Diversity facility* – Universidade de Cornell, por GBS (*Genotyping by Sequencing*). De forma resumida, o DNA foi digerido usando a enzima de restrição *Ape*KI para preparação das bibliotecas seguindo o protocolo descrito por Elshire et al. (2011). A ligação entre o adaptador *Ape*KI-cut e o DNA genômico foi realizada após a digestão das amostras e o sistema multiplex foi realizado para o sequenciamento utilizando o *Genome Analyzer 2000* (Illumina, Inc., San Diego, CA). Os dados genômicos foram submetidos ao controle de qualidade com a remoção de marcadores com *Call Rate* ≥ 0,80 e menor

frequência alélica (MAF<0.05). Em seguida os dados foram imputados utilizando software Beagle (BROWNING & BROWNING, 2009). Após o controle de qualidade a matriz de marcadores foi composta por 25.597 SNPs.

Obtenção dos valores genotípicos preditos e índices de seleção para estudo da GWAS

As variedades de mandioca foram avaliadas em quatro ambientes considerando o ano agrícola de 2012/2013 na condição irrigada (IN13) e não irrigada (DH13) e ano agrícola de 2013/2014 na condição irrigada (IN14) e não irrigada (DH14). Os BLUPs (*Best Linear Unbiased Prediction*) para as análises de GWAS, em cada condição hídrica e característica, foram obtidos pelas análises dentro de cada ambiente e em multiambientes.

As observações fenotípicas Y_{ij} do genótipo *i* na repetição *j*, dentro de ambientes foram ajustadas de acordo com a equação (1): $Y_{ij} = \mu + g_i + r_j + \epsilon_{ij}$, no qual, μ é a média geral, g_i é o vetor do efeito aleatório do genótipo *i*, r_j é o vetor dos efeitos fixo de repetição *j* e ϵ_{ij} é o efeito residual aleatório do genótipo *i* na repetição *j*. Para o modelo multiambientes, as observações fenotípicas Y_{ijk} do genótipo *i* na repetição *j* dentro do ambiente *k*, foi modelada pela equação (2): $Y_{ij} = \mu + e_k + g_i + (r/e)_{jk} + (g * e)_{ik} + \epsilon_{ijk}$, no qual, μ é a média geral, e_k é o efeito fixo do ambiente *k*; g_i é o efeito aleatório do genótipo *i*, $(r/e)_{jk}$ é o efeito aleatório da repetição *j* aninhada no ambiente *k*, $(g * e)_{ik}$ é o efeito aleatório da interação entre genótipos e ambientes e ϵ_{ijk} é o efeito residual aleatório do genótipo *i* na repetição *j* no ambiente *k*. A obtenção dos BLUPs e estimação dos componentes de variância, obtida pelo REML (*Restricted Maximum Likelihood*), e a herdabilidade no sentido amplo (h^2), foram estimadas pelo pacote *gdata* do software R versão 3.4.4 (R CORE TEAM, 2018).

Os BLUPs de todas as características foram utilizados para estimar os índices de seleção de tolerância a seca (DTI) (FERNANDEZ, 1992) e índice de estabilidade da tolerância a seca (DTSI) (BOUSLAMA & SCHAPAUGH, 1984). Os índices foram calculados a partir das equações (3) $DTI = \frac{Y_s \times Y_p}{(\bar{Y}_p)^2}$ e (4)

$DTSI = \frac{Y_S}{Y_p}$, onde, Y_s e Y_p são as características de um determinado genótipo sob condição de seca e irrigado, respectivamente, e $\bar{y}_p$ é a média de todos os genótipos para uma determinada característica sob condição irrigada.

A h^2 das característica por ambiente foi estimada de acordo com $h^2 = \frac{\sigma_G^2}{\sigma_F^2+\sigma_E^2}$, onde σ_G^2 é a variância genotípica; σ_F^2 é a variância fenotípica e σ_E^2 é a variância ambiental. Considerando, cada condição hídrica, a h^2 foi estimada pela expressão $h^2 = \frac{\sigma_G^2}{(\sigma_G^2+\sigma_A^2+\frac{\sigma_{GxE}^2}{r}+\frac{\sigma_\varepsilon^2}{rxa})}$, onde σ_G^2 é a variância genotípica; σ_A^2 é a variância ambiental; σ_{GxA}^2 é a variância da interação genótipo x ambiente; σ_ε^2 é a variância do erro entre parcelas; r é o número de repetições e a o número de ambientes.

Análise do desequilíbrio de ligação, matriz de parentesco e estrutura populacional

O desequilíbrio de ligação (LD) foi estimado utilizando os coeficientes de correlação (r^2) entre os pares de loci em cada cromossomo e em seguida o padrão de distribuição para todo o genoma foi visualizado, a partir de gráficos gerados pelo pacote *LDheatmap* do software R versão 3.4.4 (R CORE TEAM, 2018). Para investigar o declínio do LD, os valores de r^2 foram plotados em função da distância genética, em pares de base (pb), utilizando regressão não linear para ajuste, de acordo com Weisberg (2005).

Um *heatmap* mostrando o parentesco entre os genótipos foi gerado usando o método de VanRaden (2008), implementado no pacote GAPIT (*Genome Association and Prediction Integrated Tool*) (LIPKA et al., 2012), do software R versão 3.4.4 (R CORE TEAM, 2018). A estrutura populacional foi estimada, utilizando o software fastStructure (v1.0) de agrupamento e estratificação (RAJ et al., 2014). O algoritmo fastStructure determina o número de grupos (K) que melhor explica a estrutura da população. Neste caso, múltiplas opções de K foram testadas (1 a 10), para determinar o número ideal de grupos que melhor expliquem a estrutura da população.

Associação genômica (GWAS)

Foram utilizadas três entradas fenotípicas para cada característica nas análises de GWAS, para garantir que os QTLs fossem identificados considerando ambientes específicos, multiambientes e índices de tolerância à seca e de estabilidade da tolerância a seca. As variáveis utilizadas foram: 1) BLUPs para cada característica por ambiente (Equação 1); 2) BLUPs obtidos pelo modelo multiambiente (Equação 2) e 3) e BLUPs para cada característica considerando índices de tolerância à seca e de estabilidade da tolerância a seca, que levam em consideração os dados irrigados e de sequeiro conjuntamente (Equaçoes 3 e 4).

As análises de GWAS foram realizadas utilizando o modelo linear misto múltiplo (MLMM), implementado no pacote FarmCPU (*Fixed and random model Circulating Probability Unification*) (LIU et al., 2016), do software R versão 3.4.4 (R CORE TEAM, 2018). O algoritmo *Efficient Mixed Model Association* (EMMA), está implementado no pacote para reduzir o tempo computacional na estimação dos componentes de variância para cada marcador (KANG et al., 2008).

O modelo MLMM, é dividido em duas partes: um modelo de efeito fixo (FEM) e um modelo de efeito aleatório (REM) que foram usados iterativamente. Inicialmente, o REM estima os marcadores associados como covariáveis para realizar o controle de falsos positivos e os utilizam para obter a matriz de parentesco genômica (*Kinship*). Esses marcadores associados são considerados como pseudo QTNs (*Quantitative Trait Nucleotide*) e em seguida, o FEM testa todos os marcadores, um por vez, juntamente com a matriz de parentesco, usada como covariável para controlar falsos positivos e falsos negativos. Em cada iteração, o *p-value* dos marcadores de teste e os marcadores associados são unificados. A matriz de parentesco genômica e a estrutura populacional foram utilizadas como covariáveis pelo modelo MLMM, conforme proposto por Yu et al. (2006).

Foi utilizada uma correção para controlar múltiplos testes, onde, as associações significantes entre os marcadores e o fenótipo foram estimadas

utilizando o -$\log_{10}$ *p-value* e a correção de Bonferroni para múltiplos testes, aos níveis de 1 e 5% de significância.

Além disso, foi estimada a variância dos SNPs significativos por meio da equação: $\widehat{\sigma^2_{SNP}} = \widehat{a^2}.p(1-p)$, onde $\hat{a}$ é o efeito estimado do SNP, e p é a menor frequência alélica (MAF) (ZHANG et al., 2010). O efeito do SNP foi expresso em termos de proporção da variância genética explicada pelo marcador.

Anotação *in silico* de SNPs

As funções biológicas putativas de SNPs significativos foram determinadas por meio do alinhamento das sequências dos SNP com proteínas relacionadas à tolerância ao déficit hídrico utilizando uma janela de 20 kb, na base de dados Phytozome (http://www.phytozome.net), por meio do blastx.

RESULTADOS

Componentes de variância e herdabilidade

De modo geral, a herdabilidade (h^2) nos diferentes experimentos foi elevada, sobretudo na condição irrigada em 2013, cuja variação foi de 0,64 (DMC) a 0,80 (PTR e PAMD), enquanto na condição de sequeiro neste mesmo ano agrícola os valores de h^2 foram menores, com variação de 0,30 (PPA) a 0,60 (DMC) (Tabela 1).

Na condição irrigada em 2014, a h^2 das quatro características agronômicas foram parecidas com 2013, com variação de 0,50 (DMC) a 0,81 (PTR). Similarmente ao ano de 2013, a h^2 na condição de sequeiro foi menor que nos experimentos irrigados para a maioria das características, à exceção de DMC cujas estimativas foram mais elevadas nesta última condição (h^2=0,81). Na análise multiambiente na condição irrigada, a h^2 variou entre baixa (0,20 para PPA) a alta (0,69 para PTR), enquanto na condição de sequeiro as estimativas de h^2 variaram de baixa (0,26 para PPA) a mediana (0,47 para PTR). À exceção da característica PPA, as estimativas de h^2 foram maiores na condição irrigada.

Tabela 1- Componentes de variância e herdabilidade no sentido amplo para as análises individual por ambiente e conjunta para cada condição hídrica para produtividade total de raízes (PTR), produtividade parte aérea (PPA), teor de matéria seca (DMC) e produtividade de amido (PAMD).

Componentes / Características	Experimento irrigado 2013				Experimento irrigado 2014			
	PTR	PPA	DMC	PAMD	PTR	PPA	DMC	PAMD
σ_G^2	204,35	38,77	6,82	15,71	95,47	71,09	3,35	6,77
σ_ε^2	51,36	18,91	3,89	3,86	21,79	66,42	3,32	1,75
h^2	0,80	0,67	0,64	0,80	0,81	0,52	0,50	0,79
	Experimento sequeiro 2013				Experimento sequeiro 2014			
σ_G^2	20,23	4,34	15,59	1,29	9,07	20,14	15,84	0,32
σ_ε^2	15,80	10,14	10,46	0,97	7,86	20,17	3,64	0,37
h^2	0,56	0,30	0,60	0,57	0,53	0,50	0,81	0,46
	Experimento irrigado – Multiambiente				Experimento sequeiro - Multiambiente			
σ_G^2	94,99	36,88	4,48	7,15	3,94	3,24	5,50	0,19
σ_{GxA}^2	55,87	18,31	0,58	4,10	8,65	8,78	9,70	0,54
σ_A^2	22,39	137,49	0,78	1,82	0,18	4,25	4,61	0,04
σ_ε^2	33,38	46,19	3,50	2,59	13,10	15,67	7,15	0,67
h^2	0,69	0,20	0,56	0,68	0,47	0,26	0,39	0,40

σ_G^2 – variância genotípica; σ_ε^2- variância do erro; h^2 - herdabilidade no sentido amplo; σ_{GxA}^2 – variância da interação genótipo x ambiente; σ_A^2 – variância ambiental.

Entre todas as características avaliadas, as menores correlações genéticas (variação de 0,22 a 0,29) foram identificadas nos experimentos de sequeiro nos anos agrícolas de 2013 e 2014 (DH13 e DH14) (Tabela 2), enquanto que as maiores correlações genéticas foram identificadas nos ambientes irrigados nos dois anos de cultivo (IN13 e IN14). Por outro lado, correlações genéticas de magnitude mediana foram identificadas nos experimentos irrigados e de sequeiro em 2013 (IN13 × DH13), cuja variação foi

de 0,39 (DMC) a 0,58 (PPA). Além disso, as estimativas de covariâncias entre ambientes foram superiores a zero em todas as características avaliadas.

Tabela 2- Correlação genética (diagonal superior) e covariâncias (diagonal inferior) entre experimentos irrigados e de sequeiro nos anos agrícolas de 2013 e 2014 para características agronômicas em mandioca.

Característica	Experimento	IN13	IN14	DH13	DH14
PTR	IN13		0,64	0,57	0,36
	IN14	83,20		0,41	0,37
	DH13	26,48	13,13		0,24
	DH14	13,46	9,29	2,17	
PPA	IN13		0,56	0,58	0,38
	IN14	24,97		0,56	0,49
	DH13	4,23	5,22		0,22
	DH14	8,87	14,44	1,06	
DMC	IN13		0,74	0,39	0,51
	IN14	2,83		0,43	0,56
	DH13	3,00	2,31		0,29
	DH14	4,59	3,48	3,70	
PAMD	IN13		0,65	0,53	0,36
	IN14	6,25		0,33	0,33
	DH13	1,89	0,78		0,24
	DH14	0,65	0,40	0,10	

IN13 e IN14: experimentos irrigados nos anos de 2013 e 2014, respectivamente; DH13 e DH14: experimentos de sequeiro nos anos de 2013 e 2014, respectivamente; PTR – produtividade total de raízes; PPA – produtividade da parte aérea; DMC – teor de matéria seca nas raízes; PAMD – produtividade de amido.

Estrutura populacional

De acordo com a matriz de parentesco genômico existe um baixo grau de parentesco entre os 49 genótipos de mandioca analisados (Figura 1). No entanto, é possível observar a presença de dois grupos genéticos bastante distintos, de acordo com a análise de similaridade genética (Figura 1). O baixo parentesco entre a maioria dos acessos pode ser vantajoso para ações de

melhoramento genético, tendo em vista a possibilidade desses genótipos apresentarem alelos alternativos para tolerância ao déficit hídrico. No entanto, observamos algumas exceções, com maior similaridade genética entre as variedades locais BGM-0279 vs Engana Ladrão e BGM-0163 vs BGM-0876, e entre os genótipos melhorados BGM-0360 vs 9624-09, GCP-095 vs GCP-009 e GCP-014 vs Mani Branca, cujo parentesco variou de 0,51 a 0,57.

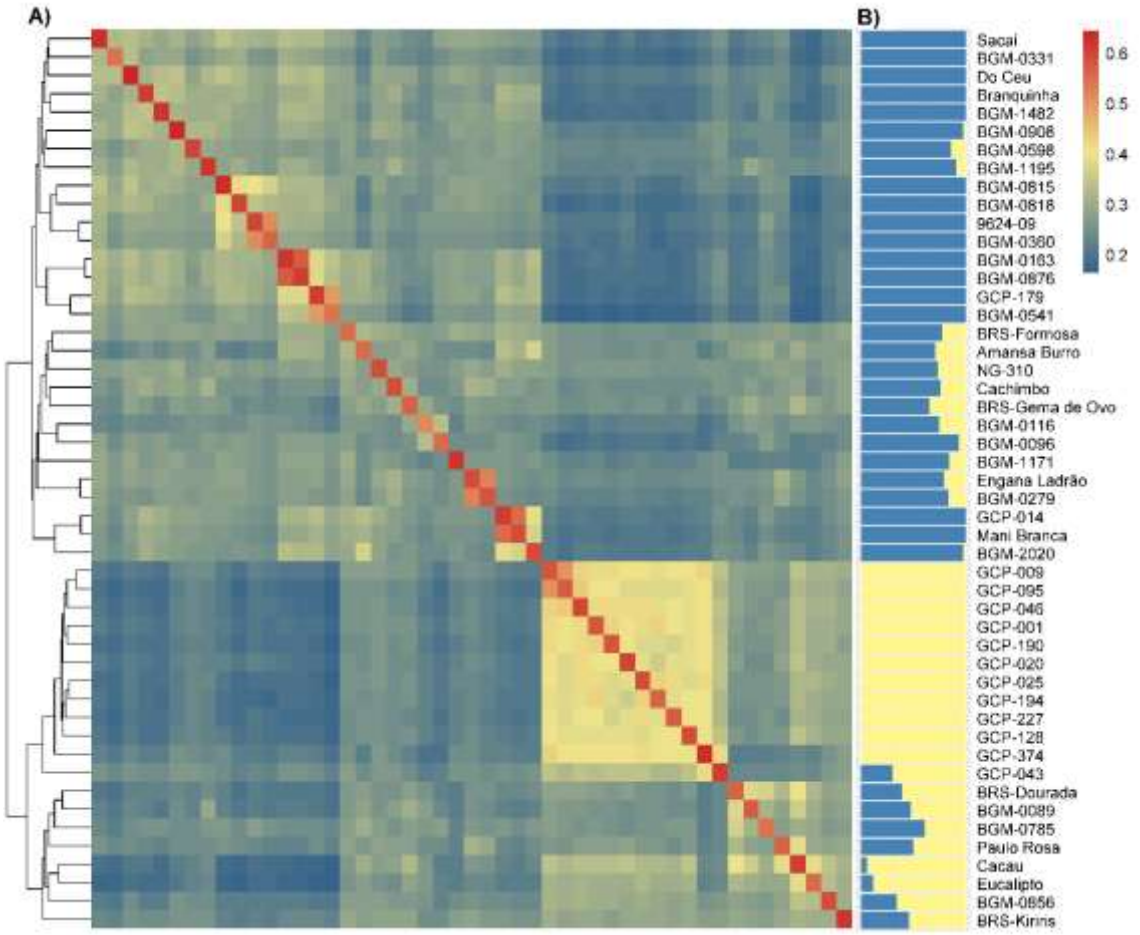

Figura 1 - A) *Heatmap* da matriz de parentesco genômico obtida pelo método de VanRaden (2008) com base na análise de 25.597 SNPs; e B) Estrutura populacional estimada pelo método fastStructure com K=2 nos 49 genótipos de mandioca avaliados para tolerância ao déficit hídrico.

O algoritmo fastStructure foi utilizado para inferir a estrutura da população, por meio de estimativas da proporção de ancestralidade entre os acessos. O número de grupos que melhor explicou a estrutura populacional foi K=2, de acordo com o algoritmo fastStructure (Figura 1). Esse resultado

corrobora os dados da matriz de parentesco genômico e análise de similaridade genética (dendograma), que estruturou os genótipos com base na origem geográfica e tolerância ao déficit hídrico. O grupo identificado pela cor azul, consistiu no agrupamento da maioria dos acessos, incluindo variedades melhoradas e genótipos considerados variedades locais. O grupo em amarelo foi constituído em sua maioria por genótipos derivados de cruzamentos entre parentais contrastantes para tolerância ao déficit hídrico, e que foram previamente selecionados em condições semiáridas no Nordeste do Brasil. Exceção, ocorreu apenas nos genótipos GCP-014 e GCP-179, que ficaram alocados no agrupamento em azul.

Análise do desequilíbrio de ligação

Um total de 25.597 SNPs (média de 1.422 SNPs por cromossomo) foi utilizado nos cálculos de desequilíbrio de ligação (LD). A média geral do r^2 foi de 0,047 com variação de 0,00 à 1,00. A extensão do decaimento do LD demonstrou diferenças em nível cromossômico, apresentando uma média cromossômica de 4,76% de pares de SNPs com r^2 > 0,20. O cromossomo 1 apresentou a maior porcentagem de SNPs em LD com r^2 > 0,20 (8,23%); já os cromossomos 17 e 10 apresentaram a menor porcentagem de pares de SNPs com r^2 > 0,20 (3,58 e 3,59%, respectivamente). As estimativas médias de r^2 em nível cromossômico variaram de 0,041 (cromossomo 10) à 0,055 (cromossomo 18) (Material Suplementar S2).

O padrão de decaimento do LD, em relação à distância entre os marcadores, foi investigado considerando todos os 18 cromossomos da mandioca e no geral, a distribuição mostrou um rápido decaimento do LD em função do aumento da distância física. Observamos um rápido decaimento do LD ao longo das distâncias físicas entre locos, com alcance próximo à 2000 pb (Figura 2).

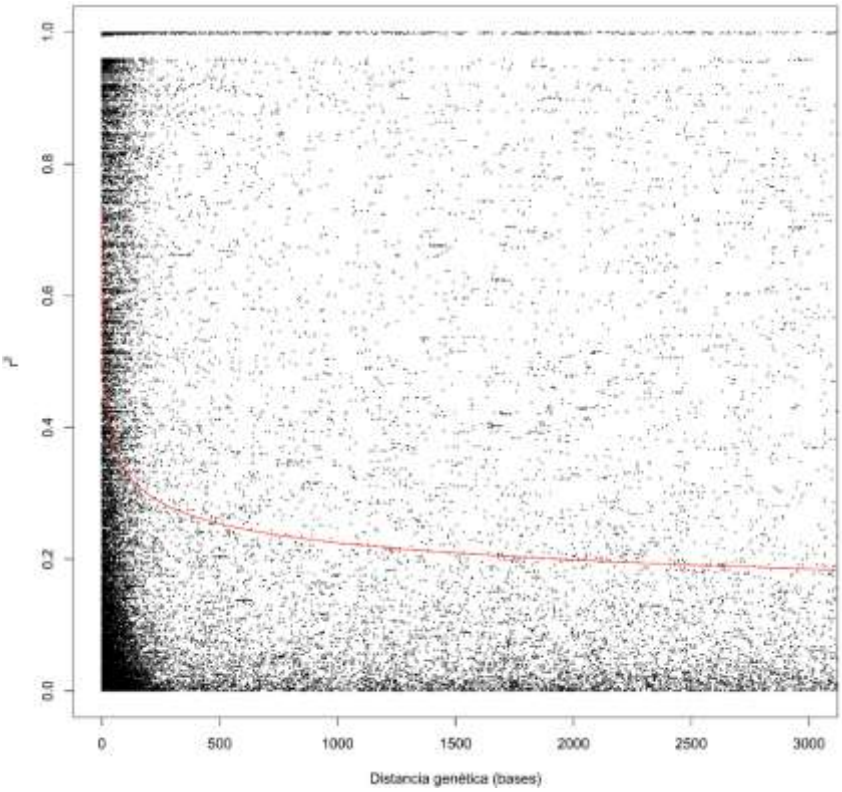

Figura 2 - Padrão de decaimento do desequilíbrio de ligação em relação à distância entre marcadores SNPs, avaliado em 49 genótipos de mandioca e considerando a análise conjunta de todos os cromossomos da mandioca.

Análise de associação genômica considerando ambientes individuais e multiambiente

Nas análises considerando as condições de sequeiro e irrigado, 23 SNPs foram identificados para as quatro características avaliadas (DMC, PAMD, PTR e PPA), em ambientes individuais e multiambiente (Tabela 3). Para a condição irrigada (2013 e 2014) foram identificados SNPs com associação significativa apenas para a característica DMC, portanto esses SNPs foram significativos apenas em ambientes sem estresse hídrico e estão localizados em seis cromossomos diferentes (1, 9, 12, 13, 15 e 17) (Figura 3). No experimento Irrigado 2013 a associação mais significativa ocorreu para o SNP S12_25684461 (P-valor (-log10) = 7,89), com efeito negativo -1,64 e baixa variância (0,288). Já no experimento Irrigado 2014, a associação mais significativa foi observada para o SNP S9_21799548 (P-valor (-log10) = 9,92), com efeito positivo de 1,19 e variância de 0,318 (Tabela 3).

Tabela 3 - Marcadores SNPs (*Single Nucleotide Polymorphism*) associados ao teor de matéria seca, produtividade total de raízes, amido e parte aérea nas raízes em experimentos irrigados de sequeiro, conduzidos em 2013 e 2014, identificados via *genome wide association studies* (GWAS).

Ambiente	SNP	Cr[1]	Posição (pb)	MAF[2]	P-valor (-log10)[y]	Efeito	Var (SNP)
Teor de matéria seca							
Irrigado (2013)	S1_28193620	1	28193620	0,47	6,08*	1,18	0,344
	S12_25684461	12	25684461	0,12	7,89**	-1,64	0,288
	S17_3366403	17	3366403	0,34	6,08*	-1,14	0,290
Irrigado (2014)	S9_21799548	9	21799548	0,34	9,92**	1,19	0,318
	S13_26038260	13	26038260	0,42	6,46**	-0,62	0,092
	S15_9075007	15	9075007	0,47	6,00*	-0,71	0,125
Sequeiro (2014)	S4_5993266	4	5993266	0,41	7,82**	-1,03	0,254
	S7_20574171	7	20574171	0,47	7,56**	2,70	1,819
	S11_4654140	11	4654140	0,49	17,57**	20,20	101,992
	S13_22348129	13	22348129	0,06	6,28**	-1,48	0,126
	S13_23968102	13	23968102	0,16	6,72**	2,11	0,609
Produtividade total de raízes							
Sequeiro (2013)	S1_18594607	1	18594607	0,49	8,57**	-8,49	17,993
	S2_4054519	2	4054519	0,21	7,26**	1,27	0,274
	S3_5509517	3	5509517	0,11	8,01**	2,61	0,679
	S15_7659343	15	7659343	0,19	8,00**	2,29	0,818
	S16_5925755	16	5925755	0,14	8,32**	-1,88	0,431
Produtividade de amido							
Sequeiro (2013)	S1_18594607	1	18594607	0,49	9,50**	-2,70	1,817
	S2_4054519	2	4054519	0,21	10,46**	0,52	0,046
	S3_3056452	3	3056452	0,36	6,48**	-0,42	0,040
	S4_3558714	4	3558714	0,49	7,76**	0,38	0,036
	S14_6078279	14	6078279	0,09	13,38**	1,43	0,170
Produtividade da parte aérea							
Sequeiro (2014)	S4_21615445	4	21615445	0,10	8,16**	4,44	1,809
Sequeiro (Multi)	S4_21615445	4	21615445	0,10	7,49**	1,13	0,117

[1] Cromossomo; [y] SNPs com associação significativa a * 5% e ** 1% pelo teste de Bonferroni; [2]Menor Frequência Alélica.

Além dos marcadores associados a DMC nos ambientes sob irrigação normal, outros cinco SNPs foram associados a esta característica em condições de estresse. Dentre os cinco SNPs associados a DMC nos experimentos de sequeiro 2014, três estão localizados nos cromossomos 4, 7, 11 e dois no cromossomo 13 (Figura 3). O SNP S11_4654140 foi o mais

significativo (P-valor (-log10) = 17,57), apresentando um alto efeito positivo (20,20) e, consequentemente, alta variância (101,99) (Tabela 3). No entanto, nenhum SNP apresentou associação estável com base na análise multiambientes para DMC.

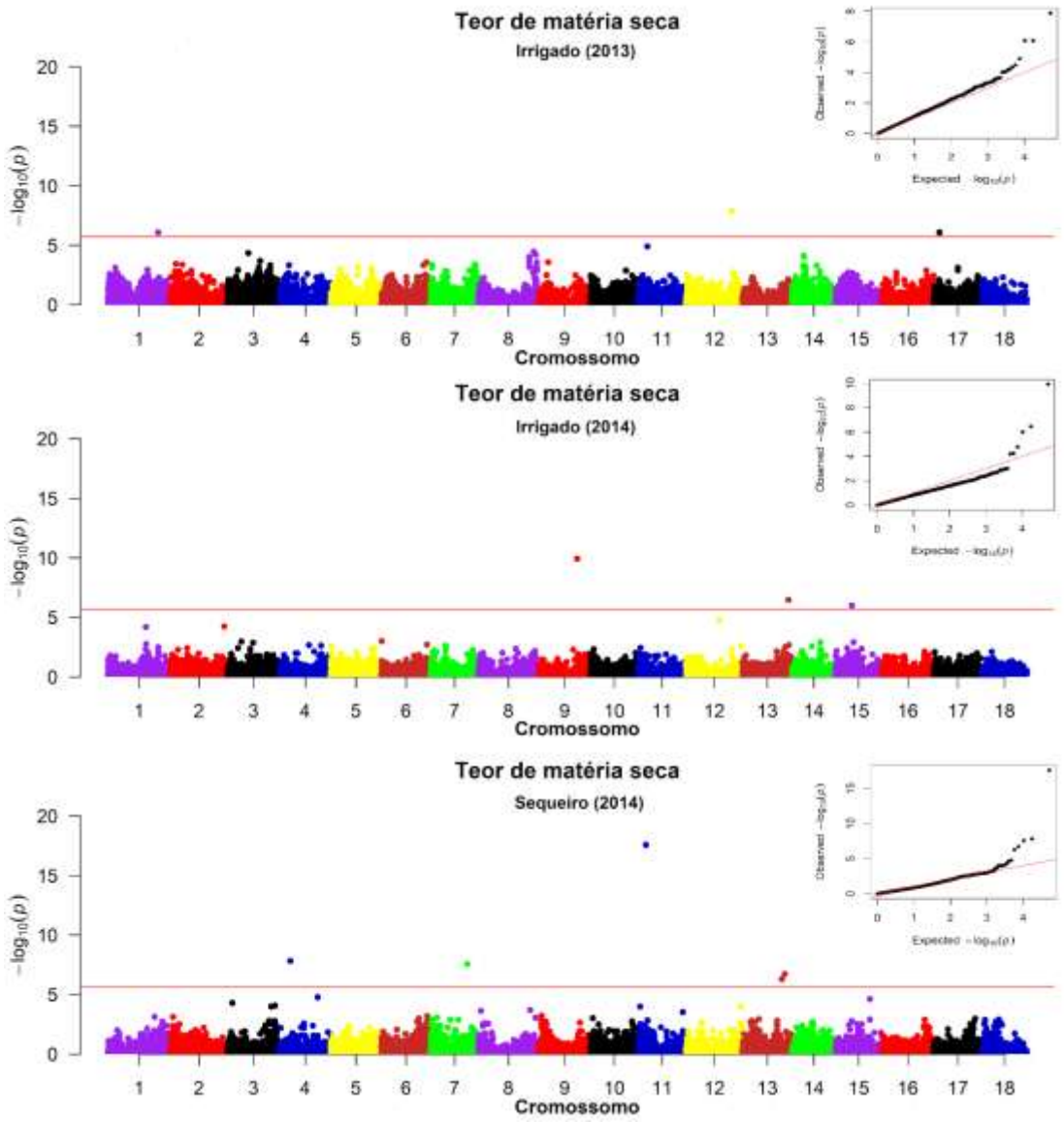

Figura 3- Gráfico Manhattan plot indicando os SNPs associados à teor de matéria seca nas raízes, em 49 genótipos de mandioca avaliados em experimentos irrigado e de sequeiro, conduzidos em 2013 e 2014. A localização dos SNPs em cada cromossomo e o teste de associação ($-\log_{10}(p)$) estão representados no eixo x e y, respectivamente. A linha vermelha indica o

nível de correção de Bonferroni (p<0,05). O gráfico acima refere-se ao quantile-quantile (QQ) dos p-valores observados e esperados da análise de associação.

Para o experimento de sequeiro avaliado em 2013, foram identificados 8 SNPs associados a PTR e PAMD, sendo que dois SNPs localizados nos cromossomos 1 e 2 (S1_18594607 e S2_4054519) apresentaram associação comum para duas características. O SNP S1_18594607 apresentou efeito negativo para ambas características (-8,49 e -2,70; com variância de 17,99 e 1,82, para PTR e PAMD, respectivamente) e o SNP S2_4054519 apresentou efeito positivo para ambas características (1,27 e 0,52, com variância de 0,27 e 0,05, para PTR e PAMD, respectivamente) (Tabela 3).

Outros SNPs significativos para PTR foram localizados nos cromossomos 3, 15 e 16, sendo que o p-valor destes SNPs variou de 8,00 a 8,32 (Tabela 3). Já para PAMD, foram identificados SNPs específicos localizados nos cromossomos 3, 4 e 14 com p-valor variando de 6,48 a 13,38 (Tabela 3 e Figura 4). Dentre os SNPs mais importantes para PAMD, o S14_6078279 apresentou a maior significância (P-valor (-log10) = 13,38), com efeito positivo (1,43) e baixa variância (0,17).

Para a característica PPA o SNP S4_21615445 localizado no cromossomo 4, apresentou associação especifica e estável entre os diferentes ambientes (Figura 4). Esse SNP foi associado a PPA no experimento de sequeiro avaliado no ano 2014 e nas análises multiambiente. Além disso, apresentou efeito positivo nos dois ambientes 4,44 (2014) e 1,13 (multiambiente), embora tenha apresentado maior variância (1,81) no ambiente 2014 (Tabela 3).

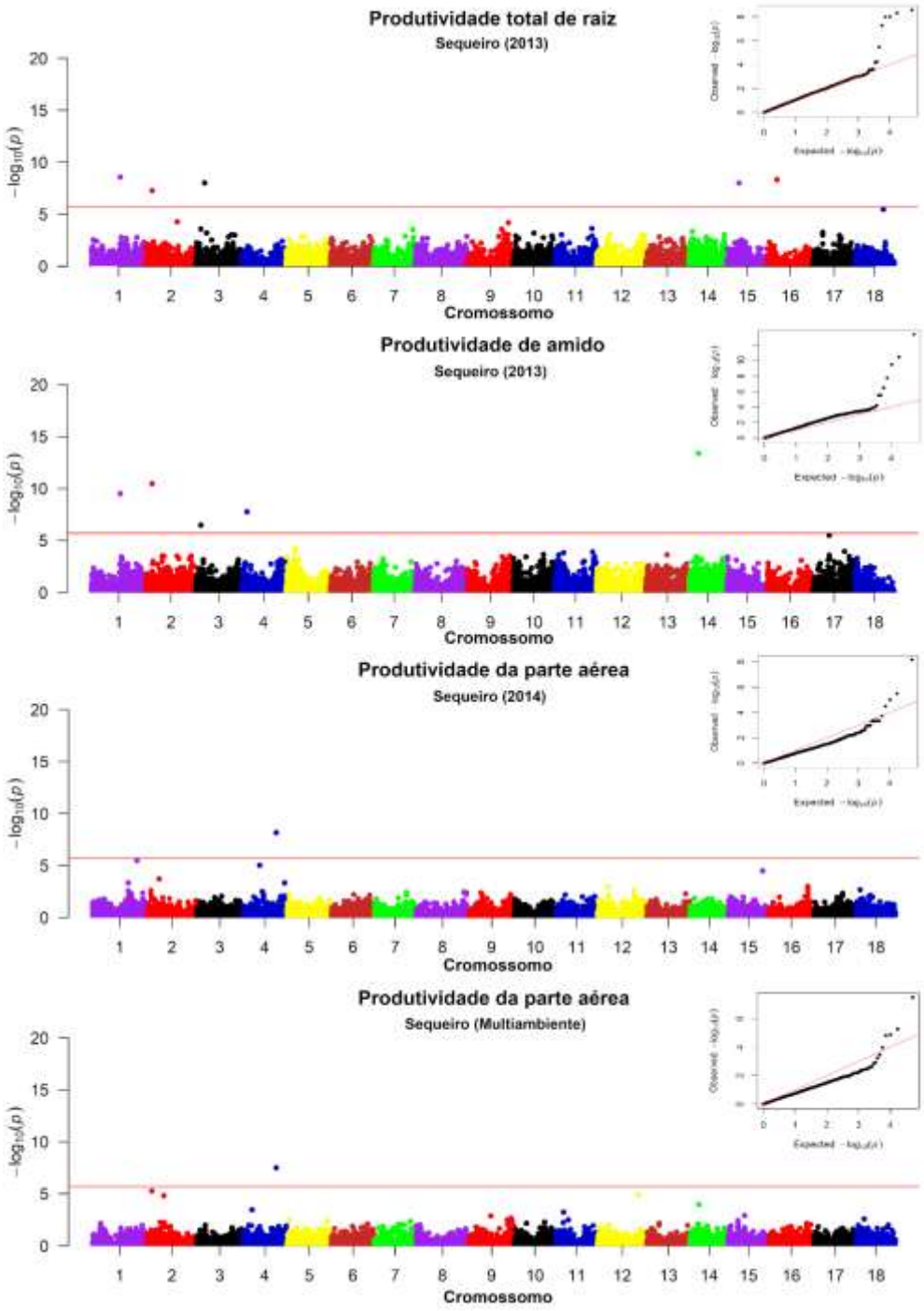

Figura 4- Gráfico Manhattan plot indicando os SNPs associados à produtividade de raízes, amido e parte aérea de 49 genótipos de mandioca avaliados em experimento de sequeiro, conduzidos em 2013; 2014 e em multiambientes. A localização dos SNPs em cada cromossomo e o teste de associação ($-\log_{10}(p)$) estão representados no eixo x e y, respectivamente. A linha vermelha indica o nível de correção de Bonferroni ($p<0,05$). O gráfico

acima refere-se ao quantile-quantile (QQ) dos p-valores observados e esperados da análise de associação.

Análise de associação genômica utilizando índices de seleção

Os índices de seleção consideraram os dados dos experimentos irrigados e de sequeiro conjuntamente, em cada ano de avaliação (2013 e 2014), e considerando a análise conjunta de anos. O intuito em utilizar esses índices de seleção relacionados ao estresse hídrico, foi o de identificar regiões genômicas estritamente relacionadas a tolerância a seca e a estabilidade de rendimento.

Foram identificados 14 SNPs com associação significativa para o índice de tolerância a seca (DTI) (Figura 5 e Tabela 4). Esses marcadores foram localizados em 12 cromossomos, comprovando a natureza poligênica da tolerância a seca em mandioca. Dentre esses marcadores, quatro foram associados a PAMD (com variância entre 0,75 a 1,86 e efeito positivo, com exceção do SNP S10_20405553) e um SNP associado à PPA, o mesmo identificado nas análises de ambientes específicos e multiambiente (com variância 0,363 e efeito positivo 1,99). Nove SNPs apresentaram associação significativa do DTI para DMC, sendo que nos experimentos em 2014, todos os cinco SNPs apresentam efeito positivo, enquanto que na análise multiambiente os outros quatro SNPs apresentaram efeito negativo e variância quase nula (variação entre 0,001 e 0,003).

Para o índice de estabilidade da tolerância a seca (DTSI), foram identificados 17 SNPs com associação significativa para as quatro características (DMC, PAMD, PTR e PPA), sendo que todos foram identificados nos experimentos de 2014. Dois deles foram identificados para PTR, localizados nos cromossomos 10 e 15; quatro SNPs foram identificados para PPA, sendo dois localizados no cromossomo 5 e os outros dois nos cromossomos 6 e 12; oito SNPs significativamente associados à DMC, localizados nos cromossomos 3, 6, 7, 11, 12 e 17; e três SNPs associados a PAMD, localizados nos cromossomos 6, 10 e 16 (Figura 6 e Tabela 5). Apesar dos altos p-valores (-log10 = 22,50), a variância destes SNPs foi nula ou baixa (variação entre 0,00 e 0,09) e, consequentemente os efeitos também foram

baixos, com variação entre -0,13 (S12_552437) a 0,60 (S11_4654140) (Tabela 5).

Tabela 4 – Marcadores SNPs (*Single Nucleotide Polymorphism*) associados ao índice de tolerância a seca para as características produtividade de amido, produtividade da parte aérea e teor de matéria seca avaliados nos anos agrícolas de 2013 e 2014, e em multiambientes, identificados via *genome wide association studies* (GWAS).

Ambiente	SNP	Cr[1]	Posição (pb)	MAF[2]	P-valor (-log10)[y]	Efeito	Var (SNP)
2013	***Produtividade de amido***						
	S4_3015131	4	3015131	0,05	6,71**	6,19	1,858
	S10_20405553	10	20405553	0,11	6,18*	-2,74	0,749
	S14_6039650	14	6039650	0,11	6,67**	4,00	1,592
	S18_2494095	18	2494095	0,07	6,49**	4,01	1,065
2014	***Produtividade da parte aérea***						
	S4_21615445	4	21615445	0,1	5,81*	1,99	0,363
	Teor de matéria seca						
	S5_1282566	5	1282566	0,05	8,99**	0,26	0,003
	S11_4654140	11	4654140	0,49	12,46**	1,05	0,274
	S14_23407800	14	23407800	0,31	6,03*	0,10	0,002
	S15_3918053	15	3918053	0,21	7,03**	0,10	0,002
	S16_21272865	16	21272865	0,44	9,80**	0,11	0,003
Multiambiente	***Teor de matéria seca***						
	S7_24121759	7	24121759	0,36	5,99*	-0,05	0,001
	S8_32094464	8	32094464	0,42	9,49*	-0,08	0,002
	S9_10268397	9	10268397	0,42	10,03*	-0,09	0,002
	S12_25167807	12	25167807	0,16	7,34*	-0,08	0,001

[1] Cromossomo; [y] SNPs com associação significativa a * 5% e ** 1% pelo teste de Bonferroni; [2]Menor Frequência Alélica.

Tabela 5 - Marcadores SNPs (*Single Nucleotide Polymorphism*) associados ao índice de estabilidade da tolerância a seca (DTSI) para as características produtividade total de raízes (PTR), produtividade da parte aérea (PPA), teor de matéria seca (DMC) e produtividade de amido (PAMD) avaliado em 2014.

SNP	Cr[1]	Posição (pb)	MAF[2]	P-valor (-log10)[y]	Efeito	Var (SNP)
Produtividade total de raízes						
S10_24780638	10	24780638	0,11	11,72**	0,49	0,024
S15_19883928	15	19883928	0,05	7,36**	0,53	0,013
Produtividade da parte aérea						
S5_5857490	5	5857490	0,23	7,55**	-0,09	0,001
S5_7261808	5	7261808	0,37	5,81*	-0,07	0,001
S6_18511652	6	18511652	0,05	7,12**	0,16	0,001
S12_552437	12	552437	0,06	6,08*	-0,13	0,001
Teor de matéria seca						
S3_27980858	3	27980858	0,09	12,24**	0,06	0,000
S3_26997330	3	26997330	0,05	7,97**	-0,06	0,000
S6_24608698	6	24608698	0,31	6,41**	0,02	0,000
S7_3279854	7	3279854	0,07	6,07*	0,03	0,000
S11_4654140	11	4654140	0,49	22,50**	0,60	0,089
S11_2699771	11	2699771	0,49	8,43**	-0,03	0,000
S12_5681456	12	5681456	0,44	9,95**	-0,05	0,001
S17_7612220	17	7612220	0,30	6,21*	0,03	0,000
Produtividade de amido						
S6_18122549	6	18122549	0,18	6,54**	-0,11	0,002
S10_24426293	10	24426293	0,13	12,28**	0,29	0,010
S16_21413646	16	21413646	0,42	6,01*	-0,08	0,002

[1] Cromossomo; [y] SNPs com associação significativa a * 5% e ** 1% pelo teste de Bonferroni; [2]Menor Frequência Alélica.

Anotação *in silico* dos SNPs

Os 54 SNPs significativamente associados a ambientes com e sem estresse hídrico e aos índices de tolerância e estabilidade da tolerância à seca, encontram-se próximos a 121 transcritos previamente identificados e disponíveis na base de dados Phytozome (http://www.phytozome.net) (Material suplementar, Tabelas S5, S6, S7 e S8). Dentre esses transcritos, 22 encontram-se relacionados aos SNPs identificados em condições irrigadas,

sendo que dos seis SNPs para esta condição hídrica, quatro deles encontram-se inseridos dentro da região gênica dos transcritos Manes.01G182700.1, Manes.17G012100.1, Manes.13G132400.1 e Manes.15G120300.1. Os transcritos que apresentam anotação funcional conhecida (6), encontram-se relacionados às proteínas quinases associadas a repetições ricas em leucinas, *Scarecrow-like* 32, proteínas potenciadores de oxigênio do fotossistema II e proteínas com funções na regulação transcricional de RNA polimerase II, além do domínio BTB/POZ e associado à GTPases Rop (Material Suplementar S5).

Nos experimentos de estresse hídrico foram identificados uma maior quantidade de transcritos, sendo 31 no total, relacionados a 14 SNPs. Desses SNPs, sete encontram-se inseridos na região codificante dos transcritos Manes.02G053200.1, Manes.03G058400.1, Manes.15G102800.1, Manes.14G074700.1, Manes.07G087500.1, Manes.11G048600.1 e Manes.04G077900.1 (Material Suplementar S6). Dentre os transcritos, nove apresentaram anotação funcional conhecida, relacionados as proteínas LURP, Nup188 e a zíper de leucina associada ao homeodomínio (domínio de 60 aminoácidos que permite ligação ao DNA e atua como fator de transcrição), e aos domínios B3, Apetala 2 (AP2) e CRM (*splicing* de RNA de cloroplasto e maturação ribossômica) (Material Suplementar S6).

Já para o índice de tolerância à seca, nove transcritos contiveram na sua sequência SNPs significativos, enquanto outros 22 transcritos foram localizados próximos aos SNPs significativos, com distância variando entre 103 a 12.762 pb. Dentre esses transcritos, três apresentam anotação funcional conhecida, relacionados as proteínas Nup188, proteínas quinases associadas a repetições ricas em leucinas e a proteínas de repetição pentatricopeptídica (PPR) (Material Suplementar S7). Dos 15 SNPs associados ao índice de estabilidade da tolerância à seca, 12 encontram-se inseridos dentro da região codificante de 12 transcritos. Além disso, 26 transcritos foram identificados próximos aos SNPs, com distância física variando entre 60 a 13.628 pb. No geral, esses transcritos possuem anotação funcional relacionados a repetições ricas em leucina (LRR), ao fator de transcrição BZIP e as proteínas zíper de leucina e quinase, além do domínio Gnk2 (Material Suplementar S8).

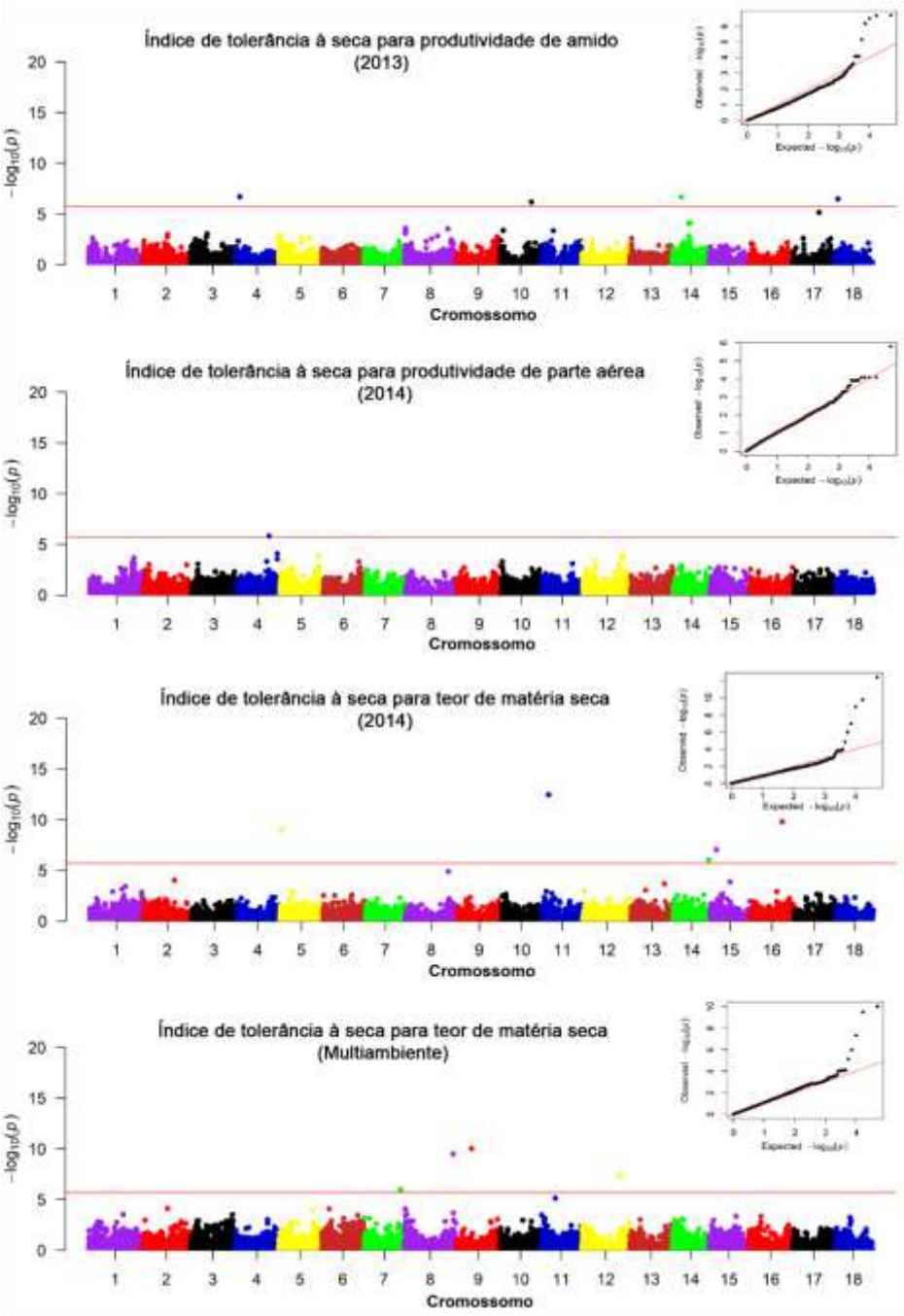

Figura 5- Gráfico Manhattan plot indicando os SNPs associados ao índice de tolerância a seca para produtividade de amido e parte aérea, além do teor de matéria seca em 49 genótipos de mandioca avaliados em experimentos conduzidos em 2013; 2014 e em multiambiente. A localização dos SNPs em cada cromossomo e o teste de associação ($-\log_{10}(p)$) estão representados no eixo x e y, respectivamente. A linha vermelha indica o nível de correção de Bonferroni ($p<0,05$). O gráfico acima refere-se ao quantile-quantile (QQ) dos p-valores observados e esperados da análise de associação.

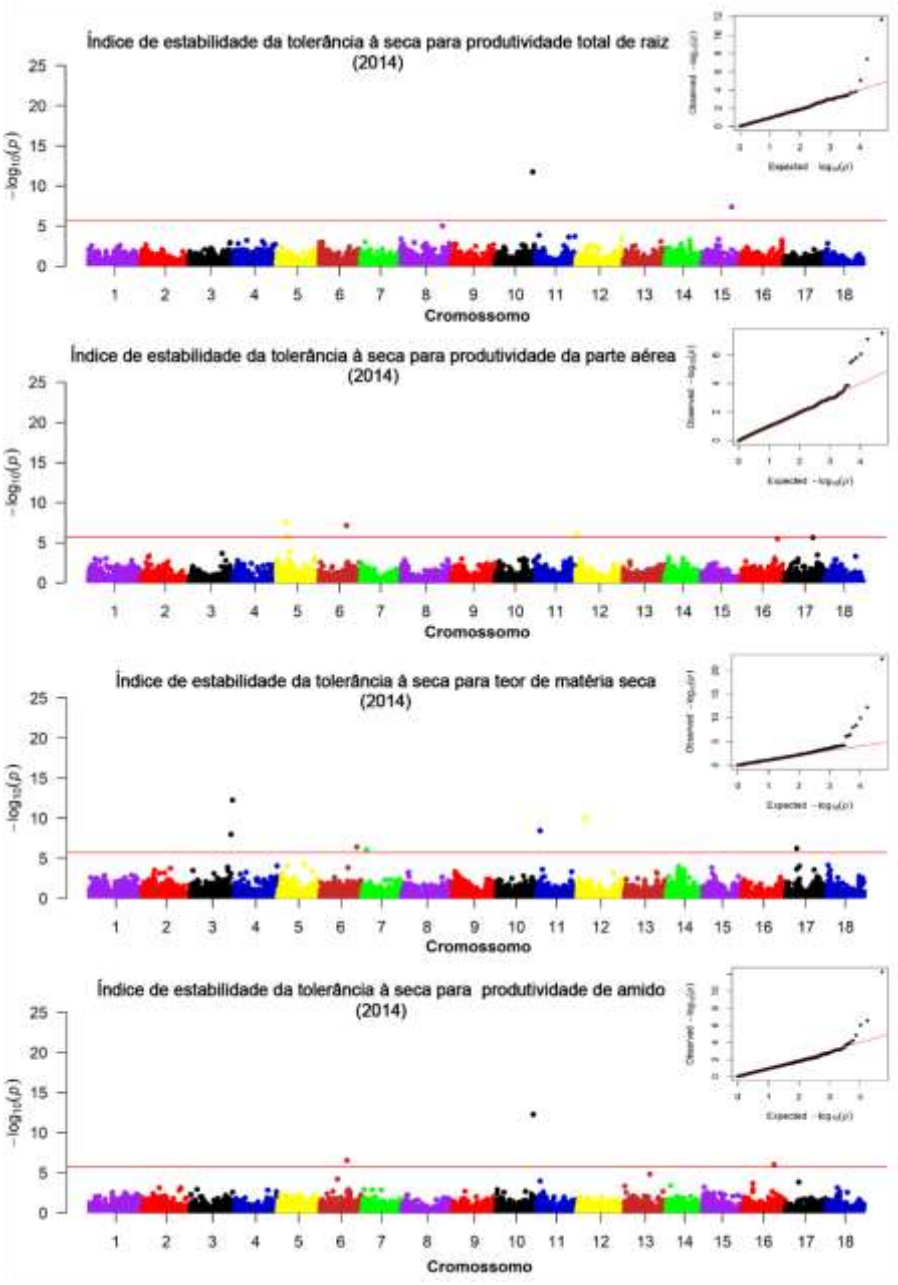

Figura 6- Gráfico Manhattan plot indicando os SNPs associados ao índice de estabilidade da tolerância à seca para produtividade de raízes, amido e parte aérea, além do teor de matéria seca em 49 genótipos de mandioca avaliados em experimento conduzido em 2014. A localização dos SNPs em cada cromossomo e o teste de associação ($-\log_{10}(p)$) estão representados no eixo x e y, respectivamente. A linha vermelha indica o nível de correção de Bonferroni ($p<0,05$). O gráfico acima refere-se ao quantile-quantile (QQ) dos p-valores observados e esperados da análise de associação.

DISCUSSÃO

Importância das características fenotípicas para análise da tolerância à seca na mandioca

O estresse hídrico induzido nos ensaios de 2013 e 2014 foi severo o suficiente para provocar alterações importantes nos componentes de variância e h^2 nas análises por ambiente e por condição hídrica (Tabela 1). Os valores da h^2 na condição de irrigação normal foram de médio a alto, porém na condição de sequeiro estes valores foram reduzidos para a maioria das características agronômicas dentro de ambiente, bem como na analise multiambiente. Este decréscimo das estimativas de h^2 é frequentemente relacionado às condições de estresse ao qual o experimento foi submetido, bem como ao material genético utilizado, que no presente estudo também foi composto por variedades suscetíveis ao déficit hídrico (FARFAN et al., 2015; OLIVEIRA et al., 2017). Resultados similares foram encontrados em experimentos considerando diferentes condições hídricas, na qual as estimativas de h^2 foram menores para características avaliadas em condições de déficit em culturas como feijão comum (HINKOSSA et al., 2013), batata (CABELLO et al., 2014) e milho (BEYENE et al., 2015).

A correlação genética avaliada neste trabalho foi maior entre os experimentos irrigados nos anos de 2013 e 2014 em comparação com os ambientes de Sequeiro 2013 e 2014. Correlações genéticas mais elevadas são esperados entre experimentos avaliados em condições ambientais similares, pois existe uma tendência na indução de respostas correlatas nos genótipos, resultando em fortes correlações genéticas (MALOSETTI et al., 2013). A baixa correlação genética entre os experimentos de sequeiro nos anos de 2013 e 2014, pode estar relacionada as diferenças nas condições climáticas nestes dois anos de cultivo, considerando que houve uma maior precipitação média anual em 2013 (347,8 mm) em relação ao ano de 2014 (216,3 mm), além de alterações na temperatura média e umidade relativa do ar (EMBRAPA SEMIÁRIDO, 2013; EMBRAPA SEMIÁRIDO, 2014).

Associação genômica entre genes e respostas ao déficit hídrico na mandioca

Com o custo do sequenciamento reduzido, os estudos de GWAS têm sido rotineiramente realizados para explorar a variação alélica associada a características de interesse agronômicos. A aplicação deste método permite a identificação de regiões em blocos de haplotipos que possibilitam associações precisas entre marcadores moleculares e características fenotípicas de interesse (HAN & HUANG, 2013). No arroz, 13 SNPs foram associados à produtividade sob condições de seca, cujas regiões genômicas estavam próximas a 30 genes que apresentam anotação funcional e, dentre esses, 10 possuem anotação funcional relacionada à seca e/ou tolerância ao estresse abiótico. Dentre os SNPs identificados, 2 apresentaram potencial para realização de ensaios TaqMan, para serem usados em PCR de rotina (PANTALIÃO et al., 2016). Já na cultura do milho, foram encontrados 42 SNPs associados a 33 genes, dos quais três foram co-localizados em regiões de QTL relacionados à seca (XUE et al., 2013).

No método GWAS, estas associações são realizadas de maneira bastante confiável, tendo em vista que possíveis associações espúrias ocasionadas devido à relação de parentesco entre os genótipos são reduzidas com base na inclusão da matriz de parentesco genômica e estrutura populacional (FLINT-GARCIA et al., 2005). No presente trabalho houve uma concordância entre a estruturação populacional em dois grupos pelo fastStructure e pela matriz de parentesco genômico. Portanto, foi possível estruturar os genótipos com base na origem geográfica, tolerância ao déficit hídrico e classificação específica em termos de melhoramento, já que a maioria dos genótipos derivados de cruzamentos entre parentais contrastantes para tolerância ao déficit hídrico permaneceram no mesmo grupo. Por outro lado, em um estudo GWAS para resistência à podridão radicular em mandioca, 263 acessos foram agrupados em quatro grupos, não apresentando relação com as características agronômicas e a origem destes acessos, devido à presença de hibridizações históricas e recentes relacionados aos genótipos inclusos nos diferentes grupos (BRITO et al., 2017).

A extensão do desequilíbrio de ligação (LD), é outro fator determinante da eficiência da GWAS. O LD é realizado com base na avaliação da associação não aleatória entre os marcadores SNPs par a par em LD significativo, sendo que o parâmetro r^2 tem sido um dos mais utilizadas para avaliar o LD (MANGIN et al., 2011; FLIT-GARCIA et al., 2003; LIPKA et al., 2015). O decaimento do LD nos 49 genótipos de mandioca ocorreu rapidamente com o aumento da distância física entre os locos no genoma (LD com r^2=0.2, próximo a 2000 pb). Estudos anteriores com mandioca e batata, espécies de propagação vegetativa, vem demonstrando a ocorrência de uma baixa extensão no LD nestas espécies (ESUMA et al., 2016; BRITO et al., 2017; STICH et al., 2013). Para que sejam encontradas associações significativas entre locos marcadores e fenótipos de interesse se faz necessário utilizar uma grande densidade de marcadores. No entanto, mesmo sendo identificado um baixo LD (r^2= 0,20 próximo de 1.320 pb), importantes associações significativas para o conteúdo de carotenoides foram relatadas na mandioca (ESUMA et al., 2016).

Conhecer a magnitude do LD é importante, pois essa informação estabelece a quantidade de marcas necessárias para a realização de diversos estudos, incluindo os estudos de associação e de seleção assistida (GRENIER et al., 2015; ESUMA et al., 2016). Além disso, o LD pode diferir entre populações da mesma espécie, sendo influenciado pelo tamanho e complexidade do genoma, padrões de recombinação do genoma, estrutura populacional, sistema reprodutivo da espécie, processos de domesticação e melhoramento (CHING et al., 2002; MATHER et al., 2007; WÜRSCHUM et al., 2013). Por exemplo, populações oriundas de melhoramento, apresentam LD com alcance de 100-500 kb em linhagens endogâmicas de milho (CHING et al., 2002). Em espécies cultivadas de arroz (*O. japonica*) proveniente de regiões temperadas o LD foi mais elevado (>500 kb), em comparação com variedades provenientes de regiões tropicais (150 kb) e de espécies de *O. indica* (75 kb) (MATHER et al., 2007). Esses resultados indicam que os processos de melhoramento e a domesticação podem interferir fortemente no comportamento do LD.

Em plantas alógamas, a exemplo do milho, o LD depende muito do tipo de população estudada, mas basicamente possui baixa extensão, devido ao maior tamanho do genoma e a alta taxa de recombinação, além da grande mobilidade existente no genoma desta espécie (pela ação de transposons e retrotransposons) (GUPTA et al., 2005). Assim, a extensão do LD em milho é tida como baixa, com rápido decaimento ao longo das distâncias físicas entre locos e com alcance variando, na maioria das populações, entre 1,0 e 10,0 kb (LU et al., 2012; TRUNTZLER et al., 2012). Além disso, o decaimento é mais rápido em regiões propensas à recombinação (*hotspots*), como observado por Würschum et al. (2013), na cultura do trigo, em que foi verificada diferenças nas medidas do LD ao longo do genoma.

Além dos fatores mencionados anteriormente que afetam o poder de detecção da análise GWAS, a densidade de marcadores é outro fator crucial (LIPKA et al., 2015). A extensão do LD determina a densidade de marcadores necessária para uma resolução de mapeamento eficiente. Se o LD decai em uma distância curta, a resolução do mapeamento tende a ser elevada, entretanto, é necessária uma grande densidade de marcadores moleculares, amplamente distribuídos no genoma, para potencializar a resolução do mapeamento. Por outro lado, se o LD se estender por uma longa distância, a resolução de mapeamento tende a ser baixa, mas um número relativamente pequeno de marcadores será necessário (GRADY et al., 2011; ZHU et al., 2008). No presente estudo utilizou-se uma densidade média de marcadores SNPs (25.597 SNPs), que mesmo após o controle de qualidade dos dados genômicos, apresentou uma ampla representatividade do genoma da mandioca, com uma distribuição de 1 SNP a cada 22.1 Mb.

A captação de efeitos genéticos associados a determinadas características agronômicas, também depende do tamanho e da representatividade da população de mapeamento. Uma população com tamanho reduzido pode levar a falsos positivos (BRACHI et al., 2011), principalmente ao se tratar de uma característica multigênica complexa, tal como a tolerância à seca. No entanto, o uso de genótipos contrastantes, com diferentes níveis de tolerância a seca, e com origens diferentes, propicia uma estrutura populacional diversa, compensando o tamanho populacional reduzido

e garantindo a diversidade da população avaliada. Embora a população de mapeamento avaliada no presente estudo seja considerada pequena, os genótipos utilizados compõem um amplo painel de diversidade genética para a tolerância ao déficit hídrico, sendo composta por variedades com diferentes níveis de tolerância a seca e origens diversas. Em outras culturas, como milho e *Arabidopsis*, a ampla diversidade genética capturada pelo painel de genótipos estudados, ainda que apresentando tamanho populacional reduzido, foi eficiente na detecção de regiões genômicas relacionadas à eficiência do uso de nitrogênio e a floração (no milho e em *Arabidopsis*, respectivamente) (ATWELL et al., 2010; MOROSINI et al., 2017). Algumas dessas regiões já foram mencionadas na literatura, evidenciando a efetividade dessa análise e a importância dessas regiões para o controle genético destas características (ATWELL et al., 2010).

Neste painel de 49 genótipos de mandioca foi possível identificar 54 associações marcador-fenótipo, sendo 48 SNPs distribuídos entre todos os cromossomos da mandioca. Quatorze SNPs estão associados aos fenótipos em condição de estresse hídrico, a maioria com efeito positivo sobre o fenótipo, 14 associados a tolerância a seca e 17 associados a estabilidade da tolerância à seca. O SNP S4_21615445 foi associado à característica PPA tanto no ambiente 2014 quanto no multiambiente em condições de sequeiro, bem como foi associado ao índice de tolerância a seca para a mesma característica. Possivelmente a identificação destas dezenas de regiões genômicas associadas a características agronômicas relacionados à tolerância à seca, ocorreu devido à análise multiambiente dos experimentos. Por outro lado, isso resultou na instabilidade na detecção de SNPs entre as condições avaliadas (irrigada e sequeiro). Nos experimentos irrigados foram detectados SNPs significativos apenas para DMC, porém houve uma consistência nos resultados, pelo fato de haver SNPs comuns nos dois anos de avaliação.

Nos experimentos irrigados, houve uma alta correlação genética (0,74) entre os anos de avaliação, além de uma maior coincidência no ranqueamento dos genótipos, em comparação com os experimentos de sequeiro, que apresentam uma baixa correlação genética entre os anos de avaliação, com variação de 0,22 (PPA) a 0,29 (DMC), além da menor coincidência no

ranqueamento dos genótipos (Material Suplementar S3 e S4). Esses fatos associados à elevada influência ambiental, podem ajudar a explicar a ocorrência apenas de marcas específicas por ambiente, na condição de sequeiro. No entanto, para PPA, o SNP S4_21615445 apresentou efeito específico e estável entre os ambientes de sequeiro. Para PPA, houve uma baixa variância da interação genótipo × ambiente e pouca influência dos anos de avaliação nos experimentos de sequeiro. Portanto, estudos GWAS considerando diferentes condições de estresse, possibilitam a identificação de efeitos alélicos em ambientes favoráveis e sobre estresse (MILLET et al., 2016). De fato, a resistência de variedades de milho à aflatoxina, apresentou baixa consistência dos QTLs associados à resistência a este caractere entre os diferentes ambientes avaliados (FARFAN et al., 2015).

Anotação *in silico* de SNPs

A identificação de regiões genômicas associadas à tolerância ao déficit hídrico contribuirá para o entendimento dos fatores genéticos envolvidos na tolerância a este estresse abiótico, além de contribuir no aprofundamento do conhecimento sobre as regiões genômicas e proteínas relacionadas. Dentre os transcritos relacionados aos SNPs identificados, alguns estão envolvidos na produção de proteínas envolvidas na tolerância a seca, como aquelas potenciadoras de oxigênio do fotossistema II, da proteína zíper de leucina associada ao homeodomínio do fator de transcrição BZIP e ao domínio Apetala 2 (AP2) (LICAUSI et al., 2013; PANTALIÃO et al., 2016; GE et al., 2012). Dentre estes, o domínio Apetala 2 (AP2), relatado na cultura do arroz, está relacionado a fatores de transcrição que desempenham papel no crescimento e desenvolvimento, bem como em respostas a estímulos ambientais (LICAUSI et al., 2013; PANTALIÃO et al., 2016).

Na cultura do trigo foi identificado um crescente acúmulo da proteína potenciadora de oxigênio do fotossistema II durante o desenvolvimento dos grãos em alguns genótipos submetidos a condições de estresse hídrico. Portanto, essa proteína pode estar relacionada com uma maior resistência à seca nessas plantas (GE et al., 2012). Situação similar foi relatada na cultura

da mandioca, onde essa proteína foi expressa em genótipos cultivados em condições de estresse hídrico (LOKKO et al., 2007).

A proteína zíper de leucina é um regulador de crescimento e de resposta à tolerância a seca em plantas. Essas proteínas são subunidades do complexo de proteínas TORC2 (rictor-mTOR), que controlam o crescimento e a proliferação de células em grande parte das plantas (JACINTO et al., 2006). A proteína Lzipper-MIP pertence à família bZIP (*basic leucine zippers*) possuindo uma série de genes envolvidos na resposta à seca em algumas culturas, à exemplo do arroz e *Arabidopsis* (UNO et al., 2000; XIANG et al.; 2008). No arroz os genes OsbZIP16 e OsbZIP23 são os principais responsáveis por conferir a tolerância a seca (XIANG et al.; 2008; CHEN, et al., 2012). Em *Arabidopsis* foi relatado que a presença do zíper de leucina associado ao homeodomínio constitui uma família de fatores de transcrição que é regulada em nível transcricional pela disponibilidade de água e ácido abscísico (DEZAR et al., 2005). Na mandioca uma análise de transcriptoma utilizando três diferentes genótipos revelou que muitos genes MebZIP foram ativados devido ao estresse hídrico, evidenciando o envolvimento desses genes na tolerância à seca (HU et al., 2016). Portanto, há uma grande probabilidade de que os SNPs próximos aos transcritos relacionados a essas proteínas, estejam associados a domínios proteicos com função na indução de tolerância à seca.

Alguns SNPs também foram relacionados a transcritos envolvidos em outros estresses, como as proteínas quinase que pertencem a uma importante família de proteínas com múltiplas funções, incluindo respostas à estresses bióticos e abióticos. Na cultura da mandioca, essas proteínas apresentam associação com repetições ricas em leucina, conferindo uma maior tolerância a doenças (LOUIS & REY, 2015). Já o domínio Gnk2, frequentemente encontrado em associação com os domínios quinase, está relacionado a respostas ao estresse salino (MIYAKAWA et al., 2009) e o domínio CRM (*splicing* de RNA de cloroplasto e maturação ribossômica), está envolvido com o crescimento e a resposta à estresses abióticos em plantas (LEE et al., 2014). Já a proteína LURP1, encontra-se associada a defesa contra patógenos, conforme relatado em *Arabidopsis* (KNOTH & EULGEM, 2008). Portanto, a identificação de transcritos envolvidos nos processos biológicos relacionados a

estresses bióticos e abióticos, especificamente aqueles que conferem uma maior tolerância à seca, traz uma importante contribuição, pois proporcionam um direcionamento em pesquisas futuras sobre a caracterização funcional dos genes de interesse.

CONCLUSÃO

Em resumo, este é o primeiro estudo sobre a utilização de GWAS visando a compreensão da tolerância à seca na cultura da mandioca. Foram identificados SNPs associados a produtividade de raízes, amido e da parte aérea, assim como para teor de matéria seca, índice de tolerância a seca e de estabilidade da tolerância à seca. Sendo que alguns desses SNPs, encontram-se próximos a regiões genômicas relacionadas a proteínas envolvidas na tolerância a seca, previamente relatadas em outras culturas.

Avanços foram dados na direção da compreensão sobre as variantes causais em genes candidatos que podem exercer influência nos níveis de tolerância ao déficit hídrico. No entanto, há a necessidade de estudos complementares direcionados para o re-sequenciamento de genes candidatos, no mapeamento de QTL, análise de um painel de germoplasma de mandioca com um maior número de genótipos e validação da expressão dos genes candidatos.

REFERÊNCIAS BIBLIOGRÁFICAS

ATWELL, S.; HUANG, Y.S.; VILHJÁLMSSON, B.J.; WILLEMS, G.; HORTON, M.; et al. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. **Nature**, v.465, n.7298, p.627-631, 2010.

BEYENE, Y.; SEMAGN, K.; MUGO, S.; TAREKEGNE, A.; BABU, R.; et al. Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. **Crop Science**, v.55, n.1, p.154-163, 2015.

BOUSLAMA, M.; SCHAPAUGH, W.T. Stress tolerance in soybeans. Part 1: evaluation of three screening techniques for heat and drought tolerance. **Crop Science**, v.24, n.5, p.933-937, 1984.

BRACHI, B.; MORRIS, G.P.; BOREVITZ, J.O. Genome-wide association studies in plants: the missing heritability is in the field. **Genome Biology**, v.12, n.10, p.232, 2011.

BRITO, A.C.; OLIVEIRA, S.A.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.L.; BROWNING, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. **The American Journal of Human Genetics,** v.84, n.2, p.210-223, 2009.

CABELLO, R.; MONNEVEUX, P.; BONIERBALE, M.; KHAN, M.A. Heritability of yield components under irrigated and drought conditions in andigenum potatoes. **American Journal of Potato Research**, v.91, n.5, p.492-499, 2014.

CATTIVELLI, L.; RIZZA, F.; BADECK, F.W.; MAZZUCOTELLI, E.; MASTRANGELO, A.M.; FRANCIA, E.; MARÈ, C.; TONDELLI, A.; STANCA, A. M. Drought tolerance improvement in crop plants: an integrated view from breeding to genomics. **Field Crops Research**, v.105, n.1, p.1-14, 2008.

CEBALLOS, H.; OKOGBENIN, E.; PÉREZ, J.C.; LÓPEZ-VALLE, L.A.B.; DEBOUCK, D. Cassava. in: root and tuber crops. **Springer**, p.53-96, 2010.

CHEN, H.; CHEN, W.; ZHOU, J.; HE, H.; CHEN, L.; CHEN, H.; DENG, X.W. Basic leucine zipper transcription factor osbzip16 positively regulates drought resistance in rice. **Plant Science**, v.193, p.8-17, 2012.

CIAT. **International Center for Tropical Agriculture**. 2017. Disponível em: <http://ciat.cgiar.org/what-we-do/breeding-better-crops/rooting-for-cassava/>. Acesso em: dez, 2017.

DEKKERS, J.C.M. Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. **Journal of Animal Science**, v.82, n.13, p.313-328, 2004.

DEZAR, C.A.; GAGO, G.M.; GONZÁLEZ, D.H.; CHAN, R.L. Hahb-4, a sunflower homeobox-leucine zipper gene, is a developmental regulator and confers drought tolerance to *Arabidopsis thaliana* plants. **Transgenic Research**, v.14, n.4, p.429-440, 2005.

DOYLE, J.J.; DOYLE, J.L. A rapid DNA isolation procedure for small amounts of fresh leaf tissue. **Phytochemistry**, v.19, p.11–15, 1987.

EL-SHARKAWY, M.A. Physiological characteristics of cassava tolerance to prolonged drought in the tropics: implications for breeding cultivars adapted to seasonally dry and semiarid environments. **Journal of Plant Physiology**, v.19, p.257-286, 2007.

EL-SHARKAWY, M.A. Stress-tolerant cassava: the role of integrative ecophysiology-breeding research in crop improvement. **Open Journal of Soil Science**, v.2, n.2, p.162-186, 2012.

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, n.5, p.19379, 2011.

EMBRAPA SEMIÁRIDO. **Centro de Pesquisa Agropecuária do Trópico Semiárido**. Dados meteorológicos de 2013. Disponível: http://www.cpatsa.embrapa.br:8080/servicos/dadosmet/ceb-anual.html. Acesso em: dez, 2017.

EMBRAPA SEMIÁRIDO. **Centro de Pesquisa Agropecuária do Trópico Semiárido**. Dados meteorológicos de 2014. Disponível: http://www.cpatsa.embrapa.br:8080/servicos/dadosmet/ceb-anual.html. Acesso em: dez, 2017.

ESUMA, W.; HERSELMAN, L.; LABUSCHAGNE, M.T.; RAMU, P.; LU, F.; et al. Genome-wide association mapping of provitamin A carotenoid content in cassava. **Euphytica**, v.212, n.1, p.97-110, 2016.

FARFAN, I.D.B.; LA FUENTE, G.N.; MURRAY, S.C.; ISAKEIT, T.; HUANG, P.C.; et al. Genome wide association study for drought, aflatoxin resistance, and important agronomic traits of maize hybrids in the sub-tropics. **Plos One**, v.10, n.2, p.0117737, 2015.

FERNANDEZ, G.C.J. Effective selection criteria for assessing plant stress tolerance. **Adaptation of Food Crops to Temperature and Water Stress**, p.257-270, 1992.

FLINT-GARCIA, S.A.; THUILLET, A.C.; YU, J.; PRESSOIR, G.; ROMERO, S.M.; MITCHELL, S.E.; DOEBLEY, J.; KRESOVICH, S.; GOODMAN, M.M.; BUCKLER, E.S. Maize association population: a high-resolution platform for quantitative trait locus dissection. **The Plant Journal**, v.44, n.6, p.1054-1064, 2005.

GE, P.; MA, C.; WANG, S.; GAO, L.; LI, X.; GUO, G.; MA, W.; YAN, Y. Comparative proteomic analysis of grain development in two spring wheat varieties under drought stress. **Analytical and Bioanalytical Chemistry**, v.402, n.3, p.1297-1313, 2012.

GRADY, B.J.; TORSTENSON, E.S.; RITCHIE, M.D. The effects of linkage disequilibrium in large scale SNP datasets for MDR. **Biodata Mining**, v.4, n.1, p.11, 2011.

GRENIER, C.; CAO, T.V.; OSPINA, Y.; QUINTERO, C.; CHÂTEL, M.H.; et al. Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. **PloS One**, v.10, n.8, p.0136594, 2015.

GUPTA, P.K.; RUSTGI, S.; KULWAL, P.L. Linkage disequilibrium and association studies in higher plants: present status and future prospects. **Plant Molecular Biology**, v.57, n.4, p.461-485, 2005.

GUTIÉRREZ, L.; GERMÁN, S.; PEREYRA, S.; HAYES, P.M.; PÉREZ, C.A.; et al. Multi-environment multi-QTL association mapping identifies disease resistance QTL in barley germplasm from Latin America. **Theoretical and Applied Genetics**, v.128, n.3, p.501-516, 2015.

HAO, Z.; LI, X.; LIU, X.; XIE, C.; LI, M.; ZHANG, D.; ZHANG, S. Meta-analysis of constitutive and adaptive QTL for drought tolerance in maize. **Euphytica**, v.174, n.2, p.165-177, 2010.

HINKOSSA, A.; GEBEYEHU, S.; ZELEKE, H. Generation mean analysis and heritability of drought resistance in common bean (*Phaseolus vulgaris* L.). **African Journal of Agricultural Research**, v.8, n.15, p.1319-1329, 2013.

HU, W.; YANG, H.; YAN, Y.; WEI, Y.; TIE, W.; et al. Genome-Wide Characterization and analysis of Bzip transcription factor gene family related to abiotic stress in cassava. **Scientific Reports**, v.6, p.1-12, 2016.

IBGE. INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Produção Agrícola**. Disponível em: <ftp://ftp.ibge.gov.br/Producao_Agricola/Fasciculo_Indicadores_IBGE/estProdAgr_201704.pdf > Acesso em: dez, 2017.

JACINTO, E.; FACCHINETTI, V.; LIU, D.; SOTO, N.; WEI, S.; JUNG, S.Y.; HUANG, Q.; QIN, J.; SU, B. SIN1/MIP1 maintains rictor-mTOR complex

integrity and regulates Akt phosphorylation and substrate specificity. **Cell**, v.127, n.1, p.125-137, 2006.

KANG, H.M.; ZAITLEN, N.A.; WADE, C.M; KIRBY, A.; HECKERMAN, D.; DALY, M.J.; ESKIN, E. Efficient control of population structure in model organism association mapping. **Genetics**, v.178, n.3, p.1709-1723, 2008.

KAWANO, K.; FUKUDA, W.M.G.; CENPUKDEE, U. Genetic and environmental effects on dry matter content of cassava root 1. **Crop Science**, v.27, n.1, p.69-74, 1987.

KNOTH, C.; EULGEM, T. The oomycete response gene lurp1 is required for defense against *Hyaloperonospora parasitica* in *Arabidopsis thaliana*. **The Plant Journal**, v.55, n.1, p.53-64, 2008.

KORTE, A.; FARLOW, A. The advantages and limitations of trait analysis with GWAS: a review. **Plant Methods**, v.9, n.1, p.29-37, 2013.

LEE, K.; LEE, H.J.; KIM, D.H.; JEON, Y.; PAI, H.S.; KANG, H. A nuclear-encoded chloroplast protein harboring a single CRM domain plays an important role in the Arabidopsis growth and stress response. **BMC Plant Biology**, v.14, n.1, p.98-109, 2014.

LICAUSI, F.; OHME-TAKAGI, M.; PERATA, P. APETALA 2/Ethylene Responsive Factor (AP 2/ERF) transcription factors: mediators of stress responses and developmental programs. **New Phytologist**, v.199, n.3, p.639-649, 2013.

LIPKA, A.E.; TIAN, F.; WANG, Q.; PEIFFER, J.; LI, M.; BRADBURY, P.J.; GORE, M.A.; BUCKLER, E.S.; ZHANG, Z. GAPIT: genome association and prediction integrated tool. **Bioinformatics**, v.28, n.18, p.2397-2399, 2012.

LIPKA, A.E.; KANDIANIS, C.B.; HUDSON, M.E.; YU, J.; DRNEVICH, J.; BRADBURY, P.J.; GORE, M.A. From association to prediction: statistical methods for the dissection and selection of complex traits in plants. **Current Opinion in Plant Biology**, v.24, p.110-118, 2015.

LOKKO, Y.; ANDERSON, J.V.; RUDD, S.; RAJI, A.; HORVATH, D.; et al. Characterization of an 18,166 EST dataset for cassava (*Manihot esculenta* Crantz) enriched for drought-responsive genes. **Plant Cell Reports**, v.26, n.9, p.1605-1618, 2007.

LOUIS, B.; REY, C. Resistance gene analogs involved in tolerant cassava–geminivirus interaction that shows a recovery phenotype. **Virus Genes**, v.51, n.3, p.393-407, 2015.

LIU, J.; ZHENG, Q.; MA, Q.; GADIDASU, K.K.; ZHANG, P. Cassava genetic transformation and its application in breeding. **Journal of Integrative Plant Biology**, v.53, n.7, p.552-569, 2011.

LIU, S.; WANG, X.; WANG, H.; XIN, H.; YANG, X.; et al. Genome-wide analysis of *ZmDREB* genes and their association with natural variation in drought tolerance at seedling stage of *Zea mays* L. **PLoS Genetics**, v.9, n.9, p.1003790, 2013.

LIU, X.; HUANG, M.; FAN, B.; BUCKLER, E.S.; ZHANG, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. **PLoS Genetics**, v.12, n.2, p.1005767:1-24, 2016.

LU, Y.; XU, J.; YUAN, Z.; HAO, Z.; XIE, C.; LI, X.; SHAH, T.; LAN, H.; ZHANG, S.; RONG, T.; XU, Y. Comparative LD mapping using single SNPs and haplotypes identifies QTL for plant height and biomass as secondary traits of drought tolerance in maize. **Molecular Breeding**, v.30, n.1, p.407-418, 2012.

MA, X.; FENG, F.; WEI, H.; MEI, H.; XU, K.; CHEN, S.; LI, T.; LIANG, X.; LIU, H.; LUO, L. Genome-wide association study for plant height and grain yield in rice under contrasting moisture regimes. **Frontiers in Plant Science**, v.7, p.1801-1814, 2016.

MALOSETTI, M.; RIBAUT, J-M.; VAN EEUWIJK, F.A. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. **Frontiers in Physiology**, v.4, p.44-61, 2013.

MANGIN, B.; SIBERCHICOT, A.; NICOLAS, S.; DOLIGEZ, A.; THIS, P.; CIERCO-AYROLLES, C. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. **Heredity**, v.108, n.3, p.285-291, 2011.

MASUMBA, E.A.; KAPINGA, F.; MKAMILO, G.; SALUM, K.; KULEMBEKA, H.; et al. QTL associated with resistance to cassava brown streak and cassava mosaic diseases in a bi-parental cross of two Tanzanian farmer varieties, Namikonga and Albert. **Theoretical and Applied Genetics**, v.130, n.10, p.2069-2090, 2017.

MATHER, A.; CAICEDO, A.L.; POLATO, N.R.; OLSEN, K.M.; MCCOUCH, S.; PURUGGANAN, M.D. The extent of linkage disequilibrium in rice (*Oryza sativa* L.). **Genetics**, v.177, n.4, p.2223-2232, 2007.

MATHEWS, K.L.; MALOSETTI, M.; CHAPMAN, S.; MCINTYRE, L.; REYNOLDS, M.; SHORTER, R.; VAN EEUWIJK, F. Multi-environment QTL mixed models for drought stress adaptation in wheat. **Theoretical and Applied Genetics**, v.117, n.7, p.1077-1091, 2008.

MILLET, E.; WELCKER, C.; KRUIJER, W.; NEGRO, S.; NICOLAS, S.; PRAUD, S.; DRAYE, X.; et al. Genome-wide analysis of yield in Europe: allelic effects as functions of drought and heat scenarios. **Plant Physiology**, v.172, p.749-764, 2016.

MIYAKAWA, T.; MIYAZONO, K.I.; SAWANO, Y.; HATANO, K.I.; TANOKURA, M. Crystal structure of ginkbilobin-2 with homology to the extracellular domain of plant cysteine-rich receptor-like kinases. **Proteins: Structure, Function, and Bioinformatics**, v.77, n.1, p.247-251, 2009.

MOROSINI, J.S.; MENDONÇA, L.F.; LYRA, D.H.; GALLI, G.; VIDOTTI, M.S.; FRITSCHE-NETO, R. Association mapping for traits related to nitrogen use efficiency in tropical maize lines under field conditions. **Plant and Soil**, v.421, n.1-2, p.453-463, 2017.

OKOGBENIN, E.; SETTER, T.L.; FERGUSON, M.; MUTEGI, R.; CEBALLOS, H.; OLASANMI, B.; FREGENE, M. Phenotypic approaches to drought in cassava: review. **Frontiers in Physiology**, v.4, p.1-15, 2013.

OLIVEIRA, E.J.; AIDAR, S.T.; MORGANTE, C.V.; CHAVES, A.R.M.; CRUZ, J.L.; COELHO FILHO, M.A. Genetic parameters for drought-tolerance in cassava. **Pesquisa Agropecuária Brasileira**, v.50, n.3, p.233-241, 2015.

OLIVEIRA, E.J.; MORGANTE, C.V.; AIDAR, S.T.; CHAVES, A.R.M.; ANTONIO, R.P.; CRUZ, J.L.; COELHO FILHO, M.A. Evaluation of cassava germplasm for drought tolerance under field conditions. **Euphytica**, v.213, n.8, p.188-208, 2017.

OGOLA, J.B.O.; MATHEWS, C. Adaptation of cassava (*Manihot esculenta*) to the dry environments of Limpopo, South Africa: growth, yield and yield components. **African Journal of Agricultural Research**, v.6, n.28, p.6082-6088, 2011.

PANTALIÃO, G.F.; NARCISO, M.; GUIMARÃES, C.; CASTRO, A.; COLOMBARI, J.M.; BRESEGHELLO, F.; RODRIGUES, L.; VIANELLO, R.P.; BORBA, T.O.; BRONDANI, C. Genome wide association study (GWAS) for

grain yield in rice cultivated under water deficit. **Genetica**, v.144, n.6, p.651664, 2016.

R CORE TEAM. R: a language and environment for statistical computing. **R Foundation for Statistical Computing**. Disponível em: <URL https://www.R-project.org/>, 2018.

RAJ, A.; STEPHENS, M.; PRITCHARD, J.K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. **Genetics**, v.197, n.2, p.573-589, 2014.

ROSENBERG, N.A.; HUANG L.; JEWETT E.M.; SZPIECH Z.A.; JANKOVIC I.; BOEHNKE M. Genome-wide association studies in diverse populations. **Nature Reviews Genetics**, v.11, p.356–366, 2010.

SEDANO, J.C.S.; MORENO, R.E.M.; MATHEW, B.; LÉON, J.; CANO, F.A.G.; BALLVORA, A.; CARRASCAL, C.E.L. Major novel QTL for resistance to cassava bacterial blight identified through a multi-environmental analysis. **Frontiers in Plant Science**, v.8, p.1169, 2017.

STICH, B.; URBANY, C.; HOFFMANN, P.; GEBHARDT, C. Population structure and linkage disequilibrium in diploid and tetraploid potato revealed by genome-wide high-density genotyping using the SolCAP SNP array. **Plant Breeding**, v.132, n.6, p.718-724, 2013.

UNO, Y.; FURIHATA, T.; ABE, H.; YOSHIDA, R.; SHINOZAKI, K.; YAMAGUCHI-SHINOZAKI, K. *Arabidopsis* basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. **Proceedings of the National Academy of Sciences**, v.97, n.21, p.11632-11637, 2000.

VANRADEN, P.M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v.91, n.11, p.4414-4423, 2008.

TRUNTZLER, M.; RANC, N.; SAWKINS, M. C.; NICOLAS, S.; MANICACCI, D.; et al. Diversity and linkage disequilibrium features in a composite public/private dent maize panel: consequences for association genetics as evaluated from a case study using flowering time. **Theoretical and Applied Genetics**, v.125, n.4, p.731-747, 2012.

TUBEROSA, R.; SALVI, S. Genomics-based approaches to improve drought tolerance of crops. **Trends in Plant Science**, v.11, n.8, p.405-412, 2006.

XIANG, Y.; TANG, N.; DU, H.; YE, H.; XIONG, L. Characterization of OsbZIP23 as a key player of the basic leucine zipper transcription factor family for conferring abscisic acid sensitivity and salinity and drought tolerance in rice. **Plant Physiology**, v.148, n.4, p.1938-1952, 2008.

XUE, Y.; WARBURTON, M.L.; SAWKINS, M.; ZHANG, X.; SETTER, T.; XU, Y.; GRULDOYMA, P.; GETHI, J.; RIBAUT, J.M.; LI, W.; ZHANG, X.; ZHENG, Y.; YAN, J. Genome-wide association analysis for nine agronomic traits in maize under well-watered and water-stressed conditions. **Theoretical and Applied Genetics**, v.126, n.10, p.2587-2596, 2013.

WEHNER, G.G.; BALKO, C.C.; ENDERS, M.M.; HUMBECK, K.K.; ORDON, F.F. Identification of genomic regions involved in tolerance to drought stress and drought stress induced leaf senescence in juvenile barley. **BMC Plant Biology**, v.15, n.1, p.1, 2015.

WEISBERG, S. **Applied Linear Regression**, Hoboken, NJ: John Wiley & Sons. v.528, 2005.

WÜRSCHUM, T.; LANGER, S.M.; LONGIN, C.F.H.; KORZUN, V.; AKHUNOV, E.; et al. Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. **Theoretical and Applied Genetics**, v.126, n.6, p.1477-1486, 2013.

YU, J.; PRESSOIR, G.; BRIGGS, W.H.; BI, I.V.; YAMASAKI, M.; et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**, v.38, n.2, p.203, 2006.

ZHANG, Z.; ERSOZ, E.; LAI, C.Q.; TODHUNTER, R.J.; TIWARI, H.K.; GORE, M.A.; BUCKLER, E.; et al. S Mixed linear model approach adapted for genome-wide association studies. **Nature Genetics**, v.42, n.4, p.355-362, 2010.

ZHU, C.; GORE, M.; BUCKLER, E.S.; YU, J. Status and prospects of association mapping in plants. **The Plant Genome**, v.1, n.1, p.5-20, 2008.

MATERIAL SUPLEMENTAR

Tabela S1 – Acessos de mandioca avaliados sob condição irrigada e não irrigada, nos anos agrícolas de 2013 e 2014 em Petrolina (PE, Brasil).

Genótipos	Tipo	Reação*	Razão de seleção	Origem País/Estado
9624-09	Melhorado	D	Alta retenção de folhas	Brasil/Bahia
BGM-0089	Variedade Local	D	Alta retenção de folhas	Colômbia/Valle
BGM-0096	Variedade Local	D	Coleção Semiárido	Brasil/-
BGM-0116	Variedade Local	T	Coleção Semiárido	Brasil/Bahia
BGM-0163	Variedade Local	D	Coleção Semiárido	Brasil/Bahia
BGM-0279	Variedade Local	D	Alta retenção de folhas	Brasil/Bahia
BGM-0331	Melhorado	D	Alta retenção de folhas	Colômbia/Valle
BGM-0360	Melhorado	D	Alta retenção de folhas	Colômbia/Valle
BGM-0541	Variedade Local	D	Alta retenção de folhas	Brasil/Bahia
BGM-0598	Variedade Local	T	Alta retenção de folhas	Brasil/Rio Grande do Sul
BGM-0785	Variedade Local	D	Alta retenção de folhas	Brasil/Bahia
BGM-0815	Variedade Local	D	Coleção Semiárido	Brasil/Alagoas
BGM-0818	Variedade Local	D	Coleção Semiárido	Brasil/ Sergipe
BGM-0856	Variedade Local	D	Coleção Semiárido	Brasil/Sergipe
BGM-0876	Variedade Local	S	Alta retenção de folhas	Brasil/Pará
BGM-0908	Variedade Local	S	Alta retenção de folhas	Colômbia/Valle
BGM-1171	Variedade Local	D	Alta retenção de folhas	Brasil/Pará
BGM-1195	Variedade Local	D	Alta retenção de folhas	Brasil/ -
BGM-1482	Variedade Local	D	Coleção Semiárido	Brasil/Bahia
BGM-2020	Variedade Local	D	Alta retenção de folhas	Brasil/Bahia
Branquinha	Variedade Local	D	Variedade produtiva	Brasil/Pernambuco
BRS A. Burro	Melhorado	T	Tolerante à seca	Brasil/Piauí
BRS Dourada	Melhorado	D	Variedade produtiva	Brasil/Bahia
BRS Formosa	Melhorado	T	Tolerante à seca	Brasil/Bahia
BRS G. Ovo	Melhorado	T	Tolerante à seca	Brasil/Amazonas
BRS Kiriris	Melhorado	T	Tolerante à seca	Brasil/Bahia
Cacau	Variedade Local	S	Alta retenção de folhas	Brasil/Pernambuco
Cachimbo	Variedade Local	S	Alta retenção de folhas	Brasil/Pernambuco
Do Céu	Variedade Local	T	Tolerante à seca	Brasil/Pernambuco
E. Ladrão	Variedade Local	T	Tolerante à seca	Brasil/Piauí
Eucalipto	Variedade Local	D	Alta retenção de folhas	Brasil/Paraná
GCP-001	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-009	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-014	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-020	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-025	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-043	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-046	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-095	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-128	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-179	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-190	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-194	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-227	Melhorado	T	Tolerante à seca	Colômbia/Valle
GCP-374	Melhorado	T	Tolerante à seca	Colômbia/Valle
Mani Branca	Melhorado	D	Alta retenção de folhas	Brasil/Paraná
NG-310	Melhorado	D	Alta retenção de folhas	Brasil/Distrito Federal
Paulo Rosa	Variedade Local	S	Alta retenção de folhas	Brasil/Bahia
Sacai	Variedade Local	T	Tolerante à seca	Brasil/Bahia

*D – Desconhecido; S – Susceptível; T – Tolerante.

Tabela S2 – Análise do desequilíbrio de ligação avaliado em 49 genótipos de mandioca**.**

Cromossomo	**Número de SNPs**	**Média do r^2***	**Pares de SNPs com r^2 > 0,20 (%)**
1	2260	0,045	8,23
2	1729	0,042	4,91
3	1789	0,048	7,2
4	1318	0,045	3,79
5	1505	0,044	4,43
6	1530	0,042	4,39
7	1120	0,046	3,64
8	1312	0,044	3,64
9	1357	0,049	5,00
10	1365	0,041	3,59
11	1495	0,046	4,74
12	1137	0,049	4,11
13	1211	0,051	4,86
14	1589	0,049	6,35
15	1597	0,046	5,27
16	1134	0,050	4,31
17	954	0,046	3,58
18	1195	0,055	3,64
Média	1422	0,047	4,76

*r^2= coeficiente de correlação

Tabela S3- Ranqueamento dos 49 genótipos de mandioca com base na produtividade das características avaliadas sob condição de irrigada e de sequeiro em 2013.

Genótipo	PTR (u+g)	Genótipo	PPA (u+g)	Genótipo	DMC (u+g)	Genótipo	PAMD (u+g)
Condição irrigada - 2013							
BRS Formosa	73,2	BGM-0541	34,5	Sacai	36,8	BRS Formosa	20,5
GCP-001	54,0	Sacai	23,4	GCP-374	35,6	GCP-001	15,7
BRS Dourada	52,7	BGM-0116	23,3	BRS A. Burro	35,2	BRS Kiriris	14,4
BRS Kiriris	49,7	BGM-0815	22,6	GCP-194	35,0	BRS Dourada	12,1
GCP-009	42,8	BRS Dourada	21,5	BGM-2020	34,1	GCP-009	11,8
BGM-0163	39,0	BGM-0360	21,0	BGM-1171	33,8	BGM-0163	11,2
BGM-0785	38,9	GCP-194	20,3	GCP-046	33,7	Mani Branca	11,0
Mani Branca	38,7	BRS Formosa	20,1	BGM-0876	33,6	BGM-0815	10,6
BGM-0815	38,0	BGM-1482	19,6	GCP-043	33,5	GCP-190	9,9
BGM-0360	36,4	BGM-0598	18,8	BRS Kiriris	33,5	BGM-0360	9,5
GCP-190	35,0	Mani Branca	18,8	GCP-095	33,5	BGM-0785	9,2
BGM-0598	33,8	9624-09	17,4	GCP-001	33,4	BGM-0598	9,0
BGM-1482	33,6	GCP-009	17,1	BGM-0116	33,4	9624-09	8,9
GCP-020	32,9	BGM-0908	16,2	BGM-0856	33,3	BGM-1482	8,8
BGM-1195	31,8	BGM-1195	16,1	Branquinha	33,1	GCP-043	8,8
9624-09	31,7	GCP-001	15,8	BGM-0163	33,1	GCP-020	8,6
BGM-0908	31,4	BGM-0163	15,8	Mani Branca	33,0	BGM-0908	8,6
GCP-043	30,1	BGM-0331	15,7	GCP-014	32,9	Branquinha	8,0
BGM-0096	29,3	BGM-0279	15,6	GCP-190	32,8	Sacai	7,7
Branquinha	28,4	GCP-046	15,5	9624-09	32,6	GCP-095	7,7
Condição irrigada - 2014							
GCP-001	39,1	Mani Branca	47,5	BGM-0163	33,8	GCP-001	10,5
Mani Branca	37,1	BGM-0598	43,9	GCP-194	33,6	Mani Branca	10,0
BRS Formosa	35,9	Cacau	41,0	GCP-374	33,4	GCP-043	9,6
GCP-043	35,3	BGM-0116	38,4	BRS A. Burro	33,3	BRS Formosa	9,5
GCP-009	33,3	9624-09	35,9	Branquinha	33,0	GCP-009	8,6
BRS Dourada	28,5	GCP-009	35,0	9624-09	32,9	BGM-0360	7,7
BGM-0360	27,7	GCP-194	34,5	BGM-0876	32,9	9624-09	7,2
GCP-020	26,0	Sacai	34,4	BGM-0116	32,8	GCP-020	6,6
E. Ladrão	25,3	BGM-0541	34,1	Cacau	32,7	BRS Dourada	6,5
BGM-0096	25,2	GCP-001	34,1	GCP-227	32,7	GCP-025	6,5
9624-09	24,9	BGM-0360	33,7	BGM-2020	32,6	BGM-0096	6,5
NG-310	24,1	GCP-025	33,5	GCP-095	32,6	GCP-194	6,4
BRS Kiriris	23,6	GCP-043	30,7	BGM-0908	32,5	E. Ladrão	6,3
GCP-025	23,2	BGM-1195	30,3	BGM-0360	32,5	BRS Kiriris	6,3
GCP-190	22,6	BGM-0908	30,2	Sacai	32,5	GCP-190	6,2
Cacau	22,0	BGM-0815	29,6	GCP-025	32,4	Cacau	6,1
BGM-0598	22,0	NG-310	29,4	GCP-046	32,4	BGM-0908	6,1
GCP-194	21,8	BGM-1482	28,2	BGM-0598	32,2	BGM-0598	6,1
BGM-0908	21,7	BRS G. Ovo	28,2	BGM-0815	32,1	NG-310	5,5
BGM-1482	19,3	BRS Dourada	27,7	BRS G. Ovo	32,0	BGM-1482	5,2

Tabela S4- Ranqueamento dos 49 genótipos de mandioca avaliados com base na produtividade das características avaliadas sob condição de irrigada e de sequeiro em 2013.

Genótipo	PTR (u+g)	Genótipo	PPA (u+g)	Genótipo	DMC (u+g)	Genótipo	PAMD (u+g)
Condição de sequeiro - 2013							
BRS Formosa	21,8	GCP-009	10,0	BGM-0116	29,7	BRS Formosa	5,7
BGM-0279	11,1	BGM-0815	9,9	BGM-0815	29,2	BGM-0815	2,7
BGM-0815	10,2	BGM-0541	9,8	BGM-0163	28,5	BGM-0163	2,5
GCP-020	10,1	BGM-0279	9,7	GCP-128	28,4	BGM-0116	2,5
BGM-0163	9,6	BGM-1195	9,7	GCP-374	28,3	GCP-020	2,4
E. Ladrão	9,6	BGM-0598	9,4	Sacai	28,2	BGM-0279	2,3
9624-09	9,3	BGM-0116	8,7	BRS Formosa	27,9	E. Ladrão	2,2
GCP-009	8,7	E. Ladrão	8,6	BGM-0876	26,0	GCP-374	2,0
Cacau	8,5	BGM-0360	8,5	Cacau	25,6	Sacai	2,0
BGM-0598	8,3	BGM-0163	8,3	GCP-194	25,5	Cacau	2,0
GCP-374	8,1	GCP-043	8,3	BRS G. Ovo	24,6	GCP-009	1,8
BGM-1482	8,0	Sacai	8,0	BGM-2020	24,4	9624-09	1,7
Sacai	7,8	Cacau	8,0	BGM-0360	24,1	BGM-0876	1,7
BGM-0876	7,6	NG-310	8,0	BGM-0279	23,9	GCP-025	1,6
GCP-190	7,6	BGM-0908	7,8	BGM-1171	23,8	BGM-1195	1,6
GCP-001	7,5	BRS G. Ovo	7,8	GCP-009	23,8	BGM-0598	1,6
BGM-1195	7,4	BGM-0876	7,8	E. Ladrão	23,8	GCP-001	1,4
GCP-025	7,4	Mani Branca	7,7	GCP-025	23,6	GCP-190	1,4
Branquinha	7,2	GCP-374	7,7	GCP-043	23,6	Branquinha	1,3
BGM-0116	6,8	BGM-0331	7,7	BGM-0598	23,2	BGM-1482	1,3
Condição de sequeiro - 2014							
BRS Dourada	13,0	Do Céu	26,5	BGM-1171	30,4	BRS Kiriris	2,2
9624-09	10,5	BGM-0541	18,5	Cacau	28,1	9624-09	2,2
BRS Kiriris	10,4	BGM-0360	18,1	BGM-2020	27,5	BRS Dourada	2,0
BGM-0818	10,3	BGM-0116	16,6	GCP-194	27,3	BGM-0815	2,0
Do Céu	9,9	Mani Branca	15,5	Sacai	27,2	Do Céu	2,0
BGM-0815	9,6	BGM-0818	15,5	BRS Kiriris	27,1	BGM-0818	1,9
BGM-0096	9,0	9624-09	15,3	GCP-374	27,0	BGM-0096	1,8
BGM-0360	8,0	GCP-025	14,6	GCP-025	26,9	GCP-025	1,7
Mani Branca	7,9	BGM-0598	14,1	Cachimbo	26,8	BGM-0163	1,7
BRS Formosa	7,8	BRS G. Ovo	12,7	GCP-043	26,7	Branquinha	1,6
Branquinha	7,7	GCP-194	12,5	BRS G. Ovo	26,3	BRS Formosa	1,6
GCP-025	7,7	Cachimbo	12,5	BGM-0815	26,2	BGM-0360	1,6
BGM-0876	7,6	Sacai	12,3	Branquinha	26,1	Mani Branca	1,6
BGM-0908	7,5	Cacau	12,2	GCP-046	26,0	BGM-0876	1,5
BRS A. Burro	7,3	BRS Dourada	12,0	GCP-179	25,8	BGM-0598	1,5
BGM-0598	7,1	BGM-0096	11,9	BGM-0163	25,8	BRS A. Burro	1,5
GCP-001	7,1	BGM-0908	11,6	GCP-227	25,5	BGM-0908	1,4
GCP-179	6,6	BGM-0331	11,2	GCP-128	25,4	GCP-001	1,4
BGM-0163	6,5	BRS A. Burro	11,1	GCP-014	25,3	GCP-179	1,4
E. Ladrão	6,5	BGM-0163	11,1	BGM-0598	25,2	E. Ladrão	1,3

Tabela S5 - Anotação funcional *in silico* dos SNPs (*Single Nucleotide Polymorphism*) associados as características teor de matéria seca em genótipos de mandioca para a condição irrigada avaliadas em 2013 e 2014.

SNP	Transcrito	D(pb)	Descrição
S1_28193620	Manes.01G182700.1	0	pectinesterase/pectinesterase inhibitor 34-related
	Manes.01G182800.1	3204	-
	Manes.01G182900.1	4544	-
S12_25684461	Manes.12G110800.1	6466	scarecrow-like protein 32
	Manes.12G110900.1	3724	transcriptional regulator of rna polii, saga, subunit (saga-tad1)
	Manes.12G111000.1	2437	-
	Manes.12G111100.1	1761	peroxidase 35-related
S17_3366403	Manes.17G012000.1	5514	-
	Manes.17G012100.1	0	ps-locus glycoprotein domain (s_locus_glycop); d-mannose binding lectin (b_lectin); protein tyrosine kinase (pkinase_tyr)
S9_21799548	Manes.09G098500.1	6771	-
	Manes.09G098600.1	4654	probable lipid transfer (ltp_2)
	Manes.09G098700.1	3575	rop guanine nucleotide exchange factor 7
S13_26038260	Manes.13G132200.1	6545	nuclear autoantigenic sperm protein nasp -related
	Manes.13G132300.1	3972	-
	Manes.13G132400.1	0	btb/poz domain (btb); pf03000- nph3 family (nph3)
	Manes.13G132500.1	4852	cold-regulated 413 plasma membrane protein 1-related
	Manes.13G132600.1	6897	-
	Manes.13G132700.1	9541	Photosystem II oxygen-evolving enhancer protein 1 (psbO)
S15_9075007	Manes.15G120100.1	4299	DNA replication licensing factor MCM3 (MCM3
	Manes.15G120200.1	1432	-
	Manes.15G120300.1	0	-
	Manes.15G120400.1	3158	Eukaryotic cytochrome b561 (Cytochrom_B561)

Tabela S6 - Anotação funcional *in silico* dos SNPs (*Single Nucleotide Polymorphism*) associados a característica produtividade de raízes, amido e parte aérea, bem como teor de matéria seca em genótipos de mandioca para a condição de sequeiro avaliadas em 2013, 2014 e pela análise multiambiente.

SNP	Transcrito	D(pb)	Descrição
S1_18594607	Manes.01G063400.1	7520	LURP-one-related (LOR)
	Manes.01G063500.1	4212	LURP-one-related (LOR)
	Manes.01G063600.1	651	E3 ubiquitin ligase involved in syntaxin degradation; Rab6 GTPase-interacting protein involved in endosome-to-TGN transport
S2_4054519	Manes.02G053100.1	6873	Glycogen phosphorylase / Polyphosphorylase
	Manes.02G053200.1	0	C2H2-like zinc finger protein-related
	Manes.02G053300.1	4780	-
	Manes.02G053400.1	7400	-
S3_5509517	Manes.03G058200.1	2816	-
	Manes.03G058300.1	1972	-
	Manes.03G058400.1	0	Exostosin family protein
S15_7659343	Manes.15G102700.1	13746	enoyl- (fabI)
	Manes.15G102800.1	0	AP2 domain (AP2); B3 DNA binding domain (B3)
	Manes.15G102900.1	7263	oligopeptide transporter-related
S16_5925755	Manes.16G042900.1	968	NAD(P)-binding rossmann-fold superfamily protein
	Manes.16G043000.1	622	-
S3_3056452	Manes.03G037700.1	5562	Uncharacterized protein (K06889)
	Manes.03G037800.1	2979	Phospholipid:diacylglycerol acyltransferase / PDAT // Diacylglycerol O-acyltransferase / Diglyceride acyltransferase
S4_3558714	Manes.04G031400.1	3895	Small subunit ribosomal protein S15Ae (RP-S15Ae, RPS15A)
	Manes.04G031500.1	8237	histone H2B (H2B)
S14_6078279	Manes.14G074500.1	10221	Homeobox domain (Homeobox) //PF02183 - Homeobox associated leucine zipper (HALZ)
	Manes.14G074600.1	8352	-
	Manes.14G074700.1	0	ATP-dependent CLP protease proteolytic subunit 6, chloroplastic
S4_5993266	Manes.04G043500.1	4498	Transferred entry: 2.3.1.234LinksB M
	Manes.04G043600.1	11301	CRS1 / YhbY (CRM) domain (CRS1_YhbY)
S7_20574171	Manes.07G087500.1	0	Diphosphomevalonate decarboxylase / Mevalonate pyrophosphate decarboxylase
	Manes.07G087600.1	649	-
	Manes.07G087700.1	5744	4,5-9,10-diseco-3-hydroxy-5,9,17-trioxoandrosta-1(10),2-diene-4-oate hydrolase
S11_4654140	Manes.11G048600.1	0	Transcriptional regulator
S13_23968102	Manes.13G112800.1	9968	N-terminal acetyltransferase
	Manes.13G112900.1	5905	Casein kinase ii subunit alpha, chloroplastic
S4_21615445	Manes.04G077900.1	0	Nuclear pore complex protein Nup188 (NUP188)

Tabela S7 - Anotação funcional *in silico* dos SNPs (*Single Nucleotide Polymorphism*) com associação significativa para o índice de tolerância a seca com base nas características produtividade de amido e da parte aérea, e teor de matéria seca em genótipos de mandioca avaliados em experimentos irrigado e sequeiro em 2013, 2014 e em análise multiambiente.

SNP	Transcrito	D(pb)	Descrição
S4_3015131	Manes.04G028000.1	103	-
	Manes.04G028100.1	4418	respiratory burst oxidase homolog protein h-related
S14_6039650	Manes.14G074100.1	8209	oligopeptide transporter-related
	Manes.14G074200.1	8005	Protein tyrosine kinase (Pkinase_Tyr); Leucine rich repeat N-terminal domain (LRRNT_2)
S18_2494095	Manes.18G026500.1	0	leucine-rich repeat protein kinase-like protein
	Manes.18G026600.1	927	50S ribosomal protein L27
	Manes.18G026700.1	4370	IMP-GMP specific 5-nucleotidase, putative-related
S4_21615445	Manes.04G077900.1	0	nuclear pore complex protein Nup188 (NUP188)
S5_1282566	Manes.05G017400.1	1605	RNA recognition motif-containing
	Manes.05G017500.1	0	-
	Manes.05G017600.1	753	3-KETOACYL-COA synthase 1-related
S11_4654140	Manes.11G048600.1	0	transcriptional regulator
S14_23407800	Manes.14G169700.1	0	genomic DNA, chromosome 3, P1 cloNE: MRC8
S15_3918053	Manes.15G052400.1	8436	-
	Manes.15G052500.1	7618	-
	Manes.15G052600.1	0	WD40 repeat protein mucilage-modified 1
S16_21272865	Manes.16G066600.1	5187	dehydration-responsive protein RD22
S7_24121759	Manes.07G112600.1	7437	PPR repeat (PPR) // PPR repeat family
	Manes.07G112700.1	0	-
	Manes.07G112800.1	4510	mitochondrial transcription
	Manes.07G112900.1	7444	F19K23.4 protein-related
S8_32094464	Manes.08G157800.1	8291	RING finger domain-containing
	Manes.08G157900.1	2261	non-specific phospholipase C3-related
	Manes.08G158000.1	1211	membrane lipoprotein-related
	Manes.08G158100.1	0	NPH3 family (NPH3)
	Manes.08G158200.1	2199	deoxyribonuclease TATDN3-related
	Manes.08G158300.1	12762	P-LOOP containing nucleoside triphosphate hydrolases superfamily protein
S9_10268397	Manes.09G073100.1	3284	-
	Manes.09G073200.1	2550	-
S12_25167807	Manes.12G109000.1	0	transducin (beta)-like 1 (TBL1)
	Manes.12G109100.1	4445	-

Tabela S8 - Anotação funcional *in silico* dos SNPs (*Single Nucleotide Polymorphism*) com associação significativa para o índice de estabilidade da tolerância à seca com base nas características produtividade de parte aérea e teor de matéria seca em genótipos de mandioca avaliados em experimentos irrigados e de sequeiro em 2014.

SNP	Transcrito	D(pb)	Descrição
S5_5857490	Manes.05G077300.1	13628	F-box and leucine-rich repeat protein 2/20 (FBXL2_20)
	Manes.05G077400.1	5851	-
	Manes.05G077500.1	0	Plant protein of unknown function (DUF868) (DUF868)
S5_7261808	Manes.05G088700.1	7197	-
	Manes.05G088800.1	0	Threonine-tRNA ligase / Threonyl-tRNA synthetase
S6_18511652	Manes.06G070500.1	0	serine/threonine-protein phosphatase 2A regulatory subunit A (PPP2R1)
	Manes.06G070600.1	1733	protein plant cadmium resistance 11-related
	Manes.06G070700.1	7805	-
	Manes.06G070800.1	9207	-
	Manes.06G070900.1	9262	bZIP transcription factor
S12_552437	Manes.12G005000.1	11815	protein phosphatase 2C 33-related
	Manes.12G005100.1	0	cytosine deaminase
	Manes.12G005200.1	5114	basic leucine zipper and W2 domain-containing protein
S3_27980858	Manes.03G197400.1	9230	3-hydroxybutyryl-CoA dehydratase / Crotonase
	Manes.03G197500.1	0	fanconi anemia group I protein (FANCI)
	Manes.03G197600.1	369	dynein light chain LC8-type (DYNLL)
	Manes.03G197700.1	13055	dynein light chain LC8-type (DYNLL)
S3_26997330	Manes.03G183000.1	6687	-
	Manes.03G183100.1	60	ATP-dependent NAD(P)H-hydrate dehydratase / ATP-dependent H(4)NAD(P)OH dehydratase
	Manes.03G183200.1	188	nuclear transcription factor Y, gamma (NFYC)
	Manes.03G183300.1	3280	aluminum induced protein with ygl and lrdr motifs
S6_24608698	Manes.06G141300.1	7799	Phosphoglucomutase (glucose-cofactor) / Glucose-1-phosphate phosphotransferase
	Manes.06G141400.1	11781	fructokinase 1
S7_3279854	Manes.07G034000.1	11613	polyadenylate-binding protein 2 (PABPN1, PABP2)
	Manes.07G034100.1	0	Salt stress response/antifungal (Stress-antifung); Protein tyrosine kinase (Pkinase_Tyr)
S11_4654140	Manes.11G048600.1	0	transcriptional regulator
S11_2699771	Manes.11G031600.1	668	DNA repair protein RadA/Sms (sms, radA)
	Manes.11G031700.1	0	6-phosphofructokinase 1-related
	Manes.11G031800.1	11250	mitochondrial-processing peptidase subunit alpha (PMPCA, MAS2)
S12_5681456	Manes.12G062600.1	0	F-box associated (FBA_1) // F-box-like (F-box-like)
S10_24780638	Manes.10G136500.1	0	RNA polymerase sigma factor SIGD, chloroplastic
	Manes.10G136600.1	4302	Leucine Rich Repeat (LRR_1 e LRR_8) and N-terminal domain (LRRNT_2)
S6_18122549	Manes.06G067200.1	0	pentatricopeptide repeat-containing protein
	Manes.06G067300.1	6788	uncharacterized DUF292
S10_24426293	Manes.10G132400.1	993	basic helix-loop-helix (BHLH) DNA-binding superfamily protein-related
	Manes.10G132500.1	1598	-
	Manes.10G132600.1	2390	mitogen-activated kinase
S16_21413646	Manes.16G067500.1	0	transcriptional corepressor leunig

CONSIDERAÇÕES FINAIS

A identificação das características mais importantes e o estabelecimento de modelos preditivos para produtividade de raízes na cultura da mandioca em condições de irrigação normal e sob déficit hídrico representam uma ferramenta importante para os programas de melhoramento genético da cultura, para avaliação do germoplasma e populações segregantes. Quatro características foram selecionadas como mais importantes para predição da produtividade total de raízes (PTR), sendo duas fisiológicas: Área abaixo da curva de progressão da expansão das folhas com base no índice de área foliar (AACP.IAF) e Número de folhas mensurado no oitavo mês (NF.8); e duas agronômicas: Número de raízes por planta (NRP) e Produtividade da parte aérea (PPA).

Na condição irrigada, a seleção das características mais importantes resultou no melhor ajuste dos modelos, sendo GLMSS; ELM; e PLS os modelos que apresentaram maior confiabilidade de predição de acordo com os valores de $r^2 > 0{,}75$ com *RMSE* variando entre 0,49 e 0,51 na condição irrigada. Já na condição sob déficit hídrico o grupo Fisio+Agro-Sel, mesmo grupo para a condição irrigada, apresenta as variáveis com melhor capacidade de predição da produtividade de raiz, no entanto os valores de r^2 (0,56 para os modelos ELM e PLS; 0,58 para os modelos GLMSS e SVM), são intermediários, indicando que os modelos estão ajustados mas que ainda podem ser melhorados. Vale ressaltar que os valores intermediários de r^2 em condição de déficit hídrico são principalmente, devido à grande variação climática em ambientes submetidos a estresses.

O mapeamento associativo propiciou a identificação de 54 SNPs significativamente associados às quatro características avaliadas, bem como para o índice de tolerância à seca e de estabilidade da tolerância, localizados em 48 loci em diferentes cromossomos, considerando ambientes específicos e multiambientes. Esses SNPs estão próximos a 120 transcritos, previamente identificados, dos quais 90 já foram previamente descritos e 24 possuem anotação funcional conhecida. Sendo que alguns desses SNPs estão próximos a regiões genômicas relacionadas a proteínas envolvidas com uma maior

tolerância a seca, já relatadas em outras culturas, à exemplo do domínio Apetala 2 (AP2); da proteína potenciadora de oxigênio do fotossistema II; da proteína zíper de leucina; da zíper de leucina associada ao homeodomínio e do fator de transcrição BZIP. Representando um avanço significativo para estudos de mapeamento associativo para déficit hídrico na cultura da mandioca. Portanto, o presente estudo apresentou informações relevantes que servirão de auxílio para futuros estudos visando a compreensão da tolerância à seca na cultura da mandioca, tanto para estudos de mapeamento associativo, quando para estudos visando predição da produtividade de raízes.