**UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA**
**CENTRO DE CIÊNCIAS AGRÁRIAS AMBIENTAIS E BIOLÓGICAS**
**EMBRAPA MANDIOCA E FRUTICULTURA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS GENÉTICOS VEGETAIS**
**CURSO DE MESTRADO**

# DIVERSIDADE GENÉTICA E IDENTIFICAÇÃO DE DUPLICATAS DE *Manihot esculenta* Crantz COM BASE EM MARCADORES *SINGLE-NUCLEOTIDE POLYMORPHISM* (SNP)

**Hilçana Ylka Gonçalves de Albuquerque**

**CRUZ DAS ALMAS – BAHIA**
**AGOSTO – 2017**

# DIVERSIDADE GENÉTICA E IDENTIFICAÇÃO DE DUPLICATAS DE *Manihot esculenta* Crantz COM BASE EM MARCADORES *SINGLE-NUCLEOTIDE POLYMORPHISM* (SNP)

**Hilçana Ylka Gonçalves de Albuquerque**

Licenciatura em Ciências Biológicas

Universidade de Pernambuco, 2014

Dissertação apresentada ao Colegiado do Programa de Pós-Graduação em Recursos Genéticos Vegetais da Universidade Federal do Recôncavo da Bahia e Embrapa Mandioca e Fruticultura, como requisito parcial para obtenção do Título de Mestre em Recursos Genéticos Vegetais.

**Orientador:** Prof. Dr. Eder Jorge de Oliveira
**Coorientadora:** Drª. Ana Carla Brito

**CRUZ DAS ALMAS – BAHIA**
**2017**

## FICHA CATALOGRÁFICA


A345d

Albuquerque, Hilçana Ylka Gonçalves de.
Diversidade genética e identificação de duplicatas de *Manihot esculenta* crantz com base em marcadores Single-Nucleotide Polymorphism (SNP) / Hilçana Ylka Gonçalves de Albuquerque._ Cruz das Almas, BA, 2017.
112f.; il.

Orientador: Eder Jorge de Oliveira.
Coorientadora: Ana Carla Brito.

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

1.Mandioca – Recursos genéticos vegetais. 2.Mandioca – Germoplasma vegetal. 3.Variabilidade genética – Análise. I.Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas. II.Título.

CDD: 633.68

**UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA**
**CENTRO DE CIÊNCIAS AGRÁRIAS AMBIENTAIS E BIOLÓGICAS**
**EMBRAPA MANDIOCA E FRUTICULTURA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS GENÉTICOS VEGETAIS**
**CURSO DE MESTRADO**

# DIVERSIDADE GENÉTICA E IDENTIFICAÇÃO DE DUPLICATAS DE *Manihot esculenta* Crantz COM BASE EM MARCADORES *SINGLE-NUCLEOTIDE POLYMORPHISM* (SNP)

**Comissão Examinadora da Defesa de Dissertação**
Hilçana Ylka Gonçalves de Albuquerque

Aprovada em: 29 de Agosto de 2017

_____
Prof. Dr. Eder Jorge de Oliveira
Embrapa Mandioca e Fruticultura
Orientador

_____
Drª. Cláudia Fortes Ferreira
Embrapa Mandioca e Fruticultura
Examinador Externo

_____
Dr. Onildo Nunes de Jesus
Embrapa Mandioca e Fruticultura
Examinador Externo

## DEDICATÓRIA

Ao Senhor da minha vida, dono do meu respirar e que me conheceu antes mesmo da minha própria existência.

À Deus eu dedico.

**AGRADECIMENTOS**

Que darei ao Senhor por todos os benefícios que me tem feito? Por ter me ajudado a chegar até aqui? Não existem palavras que descrevam a gratidão que tenho ao Senhor por ter me sustentado com sua mão fiel.

Sou grata a minha amada família por simplesmente ser a minha família, pelo grande amor e apoio em todos os momentos que precisei. Em especial, a minha melhor amiga, companheira, confidente e mãe Maria Edelza, por cada palavra sempre dita na hora certa e a meu pai Agnaldo por todo amor por mim.

Agradeço a minha irmã querida Thaise Camila e a meu sobrinho Gabriel pelo carinho, amor e por cada oração feita em meu favor, vocês são os meus tesouros.

Agradeço aos amigos, Kathia Lucas, Carla Valdeci e ao pequeno Asaf, Elisângela Mendes, Brísila Raquel, Bianca, Janaína, Vanessa, Rosa, Juliana, Deisiany e Ana Carina pelo apoio e companheirismo de sempre, como também pelas orações.

A minha família que construí em Cruz das Amas, Rafaella Roque, Jucieny Ferreira e Elizete, minha eterna gratidão por tudo.

Ao meu orientador Dr. Eder Jorge de Oliveira, por toda paciência que teve comigo durante o mestrado. Que Deus possa abençoa-lo grandemente em tudo na vida e renove suas forças dia após dia.

A minha linda Coorientadora Ana Carla Brito pelo auxílio na minha caminhada acadêmica e por ter se tornado amiga e parceira do “cafezinho”. Já estou com saudades dos seus bolos e tortas.

Em especial, agradeço a Cátia Dias, como costumava cantar “o que eu sou sem você? ” Obrigada por tudo que me ensinou, te levarei no coração sempre! Você não vai se livrar desta filha aqui facilmente.

Gilmara Fachardo como posso te agradecer por tudo? Desde seus maravilhosos “dotes” culinários a grande companheira que foi, principalmente quando precisei de ajuda na imputação dos meus dados, jamais esquecerei tudo que fez por mim. Que Deus te recompense grandiosamente e conceda o desejo do seu coração, você é um exemplo de quem pratica altruísmo sempre.

Agradeço aos amigos do PPRGV, Simone, Poliana, Gabriela, Deyse, Isabel, Edivânia, Bruna, Lima, meu irmão moçambicano Virgílio e Manassés por toda ajuda e força que me deram.

A todos os amigos de Laboratório de Biologia Molecular, meus sinceros agradecimentos e a toda família mandioca por toda ajuda e alegria que me proporcionaram convivendo e aprendendo com cada um.

À Embrapa Mandioca e Fruticultura pela estrutura física para desenvolvimento do trabalho.

À Universidade Federal do Recôncavo da Bahia por todo o programa.

Muito obrigada!

## EPÍGRAFE

Nunca me deixes esquecer que tudo o que tenho, tudo o que sou, o que vier a ser vem de Ti, Senhor.

Ana Paula Valadão.

**DIVERSIDADE GENÉTICA E IDENTIFICAÇÃO DE DUPLICATAS DE *Manihot esculenta* Crantz COM BASE EM MARCADORES *SINGLE-NUCLEOTIDE POLYMORPHISM* (SNP)**

**RESUMO:** A mandioca (*Manihot esculenta* Crantz) é uma das fontes alimentares mais importantes nos trópicos, e grande parte da sua variabilidade genética é conservada em Bancos Ativos de Germoplasma (BAG). Assim, este trabalho teve como objetivo analisar a diversidade genética e a estrutura populacional de 1.580 acessos de mandioca, bem com identicar genótipos redundantes a partir da análise de 2.371 acessos conservados em diferentes unidades da Embrapa. Todas estas análises foram realizadas com base em marcadores *Single-Nucleotide Polymorphism (SNP)*, obtidos pela técnica de *genotyping-by-sequencing* (GBS). A média dos parâmetros de diversidade genética, conteúdo de informação polimórfica (PIC), endogamia (*f*), heterozigosidade observada (Ho) e esperada (He) foram de 0,24; 0,21; 0,23; e 0,30; respectivamente, tidos como elevados, quando considerado a natureza (predominantemente bialélica) dos SNPs e sistema reprodutivo da espécie. Os valores destes parâmetros foram bastantes similares nos 18 cromossomos da espécie. Em nível de indivíduo, os valores de *f* variaram entre 0,49 a 0,97, com média de 0,69, sendo que três acessos de mandioca apresentaram f > 0,90. Os valores de desequilíbrio de ligação (LD) se estendeu entre 15 e 20 kb ($r^2$= 0,20). A análise discriminante de componentes principais (ADCP) indicou a formação de 22 grupos, com probabilidade média de alocação dos indivíduos >0,99, contudo, não foi possível observar associação entre os grupos formados pela ADCP e classificação com base em informações fenotípicas, de origem genética e geográfica. Para identificação de duplicatas foi realizado um estudo com base no agrupamento de perfis multilocos (MLGs) nos acessos conservados em diferentes unidades da Embrapa. Foi possível identificar 1.757 acessos únicos e 614 acessos duplicados, cerca de 25,89% do total de acessos analisados, com a redundância variando de 22,47% (Embrapa Amazônia Oriental) a 40,0% (Embrapa Semiárido). Estes resultados proporcionarão um melhor entendimento sobre a variabilidade genética conservada e a organização populacional do germoplasma, uma vez que a presença de acessos duplicados compromete o desenvolvimento do germoplasma, e aumenta os custos necessários para a adequada conservação, caracterização, avaliação e uso destes materiais.

**Palavras chave:** *genotyping-by-sequencing*; mandioca; *MLGs*; variabilidade

# GENETIC DIVERSITY AND IDENTIFICATION OF *Manihot esculenta* Crantz DUPLICATES BASED ON *SINGLE-NUCLEOTIDE POLYMORPHISM* (SNP)

**ABSTRACT:** Cassava *(Manihot esculenta* Crantz) is one of the most important food sources in the tropics and much of their genetic diversity is preserved at germplasm collections. Therefore, the objective of this work was to analyze the genetic diversity and population structure of 1,580 cassava accessions, as well as to identify redundant genotypes from the analysis of 2,371 accessions preserved at different Embrapa research units. All these analysis were performed based on *Single-Nucleotide Polymorphism* (SNP) markers identified by *genotyping-by-sequencing* (GBS). Mean genetic diversity parameters, polymorphic information content (PIC), inbreeding *(f),* observed heterozygosity (Ho) and expected heterozygosity (He) was 0.24; 0.21; 0.23; and 0.30 respectively, considered high, considering the predominant biallelic nature of the SNPs and the species reproductive system. The values of these parameters were quite similar throughout the 18 chromosomes of the species. At individual level, *f* values ranged from 0.49 to 0.97; with an average of 0.69, for three cassava accessions, being f > 0.90. The values of linkage disequilibrium (LD) extended between 15 and 20 kb ($r^2$ = 0.20). The discriminant analysis of principal components (DAPC) indicated the formation of 22 groups, with individual allocation probability > 0.99. However, no association between the DAPC groups and the classification based on the phenotypic and genetic information, as well as the geographical origin, was identified. In order to identify duplicates a study was conducted based on the grouping of accessions according to multilocus analysis (MLGs) from different Embrapa units. 1,757 single and 614 duplicate accessions, approximately 25.89% of the total, with redundancy ranging from 22.47% (Embrapa Amazônia Oriental) to 40.0% (Embrapa Semiarido), were identified. These results will provide a better understanding of the preserved genetic variability and the germplasm population organization, since the presence of duplicates compromises germplasm development and increases the costs of proper conservation, characterization, evaluation and use of these materials.

**Keywords:** cassava; *genotyping-by-sequencing*; *MLGs*; variability

**SUMÁRIO**

**Página**

## INTRODUÇÃO GERAL

A mandioca (*Manihot esculenta* Crantz) é uma das fontes alimentares mais importantes para cerca de 800 milhões de pessoas em todo o mundo devido ao seu valor nutricional (LEBOT, 2009). Nos trópicos, é considerada primordial na alimentação humana, principalmente em populações de baixa renda. O seu cultivo requer menor uso de insumos agrícolas e manejo técnico especializado em comparação com outras culturas, o que contribui para a sua expansão em países em desenvolvimento com elevado potencial para tornar-se uma cultura modelo em sistemas de agricultura sustentável (SILVA et al., 2011).

É uma cultura que apresenta ampla variabilidade genética natural, decorrente da seleção natural, evolução da espécie, domesticação, facilidade de polinização cruzada e alta heterozigosidade, o que origina uma infinidade de genótipos capazes de adaptar-se à condições edafoclimáticas extremas (áreas propensas a secas, inundações, solos de alta acidez e baixa fertilidade), que geralmente não são toleráveis a outras culturas, permitindo assim, o seu cultivo em diversos países (VENTURINI et al., 2016).

Parte desta diversidade é conservada em Bancos Ativos de Germoplasma (BAG) distribuídos em diversos países. Somente no Brasil, mais de 4.000 acessos são mantidos pela Empresa Brasileira de Pesquisa e Agropecuária (EMBRAPA), sendo que a Embrapa Mandioca e Fruticultura, localizada em Cruz das Almas-BA, possui atualmente mais de 1.600 acessos conservados em condições de campo (*ex situ*) e de laboratório (*in vitro*).

A principal razão para o estabelecimento e manutenção de um BAG é a conservação da máxima variabilidade genética possível da espécie para uso imediato e futuro (GOMES et al., 2007). Além disso, os BAGs precisam armazenar informações que caracterizem e diferenciem os acessos, a fim de, identificar genótipos que reúnam características de importância agrícola e econômica, destinados para uso em programas de melhoramento genético ou diretamente no sistema de produção da cultura. Portanto, a disponibilidade dos recursos genéticos em um banco necessita da caracterização morfoagronômica, fitopatológica, entomológica e molecular (ZUIN et al., 2009).

A caracterização descreve a variabilidade dos acessos de uma coleção,

por meio de características de interesse, contribuindo assim para o entendimento do relacionamento genético e com base nisso, ajuda na definição de possíveis cruzamentos entre acessos promissores (ABACA et al., 2013). A caracterização de um germoplasma também é importante para resolver um grave problema enfrentado por diversos bancos de germoplasma, que se refere à duplicação de acessos (ROBICHAUD et al., 2006). Como a constante troca de materiais genéticos ocorre livremente no Brasil, em muitas situações, materiais originários de uma região do país são comumente encontrados em outras regiões com outros nomes, assim como também é encontrado um mesmo nome dado a diferentes acessos (MOURA et al., 2013). Essas duplicatas oneram a manutenção dos BAGs e precisam ser identificadas para se garantir uma maior e mais representativa diversidade genética da espécie, além de otimizar a conservação e o manejo do germoplasma.

Dentre as ferramentas utilizadas para estudar a estrutura da diversidade genética existente nos BAGs e identificar acessos duplicados, destaca-se o uso das ferramentas moleculares, em virtude do maior grau de informação sobre a diversidade genética dos indivíduos e de não sofrerem influência ambiental e nem do estádio de desenvolvimento da planta (RABBI et al., 2015). Assim, os *Single-Nucleotide Polymorphism* (SNP) representam a classe de marcadores com elevada distribuição no genoma das espécies. Na cultura da mandioca, um SNP é encontrado a cada 121pb (POOTAKHAM et al., 2014), tendo sua aplicação destacada pela evolução nos métodos de genotipagem que permitem o processamento de um grande número de marcadores e amostras simultaneamente, com menor custo, maior rendimento e reprodutibilidade superior em relação aos outros tipos de marcadores (ELSHIRE et al., 2011).

O uso dos marcadores moleculares na cultura da mandioca tem ajudado a responder várias questões, a exemplo da origem da espécie (OLSEN, 2004); identificação de acessos duplicados (RABBI et al., 2015) e diversidade genética (GONÇALVES et al., 2017). Contudo, ainda é preciso investir esforços para obtenção do conhecimento detalhado sobre a variabilidade e estrutura genética conservada nos BAGs, que são de grande importância para definir prioridades de conservação, a fim de auxiliar na escolha de estratégias eficientes de melhoramento, bem como entender a história evolutiva da espécie ao longo do processo de domesticação, diminuindo assim a erosão genética da espécie e

explorando todo o seu potencial.

## REFERENCIAL TEÓRICO

### Aspectos gerais e econômicos da cultura da mandioca

A mandioca pertence à família Euphorbiaceae e ao gênero *Manihot*, que compreende aproximadamente 98 espécies. Dentre as espécies deste gênero, destaca-se a *Manihot esculenta* Crantz, por ser a única cultivada comercialmente (OLSEN, 2004).

É uma das culturas em que tudo pode ser aproveitado, tanto a parte aérea quanto as raízes. A parte aérea (principalmente as folhas) é utilizada tanto na alimentação animal quanto humana. Na alimentação animal, as folhas são utilizadas como feno, silagem ou até mesmo frescas. Já na alimentação humana, as folhas também são consumidas como vegetal ou acompanhamento de pratos principais. Principalmente na região nordeste do Brasil, as folhas são desidratadas e usadas na forma de farinha. As hastes são utilizadas como material propagativo em novos plantios (LATIF; MULLER, 2015).

As raízes podem ser utilizadas tanto na alimentação humana e animal, como na forma processada, gerando produtos dos quais o amido é o componente majoritário. As raízes podem ser classificadas como mansas ou bravas, de acordo com a quantidade de compostos cianogênicos (HCN) presentes, sendo este, responsável pela toxidade das raízes, que é um fator limitante para o consumo humano (PENTEADO; FLORES, 2001). Assim, a maioria das variedades bravas é direcionada para o processamento de amido ou farinha, enquanto as mansas, para o consumo humano na forma *in natura*.

De fato, o amido é o componente de maior valor agregado da cultura. Um carboidrato que se apresenta na forma de grânulos, com formato e tamanho dependentes da sua fonte botânica cujos teores podem variar de acordo com o genótipo, condições edafoclimáticas e época de colheita (VASCONCELOS et al., 2017). O amido de mandioca apresenta propriedades especiais, como clareza, viscosidade, gelatinização e sabor suave, que o torna mais adequado para determinados usos na indústria alimentícia, em comparação ao amido de cereais, que apresenta sabor característico de

cereais (SINGH et al., 2007). É utilizado na alimentação humana e como fonte de matéria-prima para diferentes produtos industriais como etanol, embalagens, têxtil, farmacêutica e alimentos embutidos, pois confere excelente textura, poder espessante e menor retrogradação durante o processo de congelamento/descongelamento em comparação com amido de outras fontes. Tais aspectos permitem uma maior estabilidade e contribui para melhorar características sensoriais dos produtos nos quais incorporam o amido (FRANCO et al., 2001).

Além disso, o amido é o único produto com mercado internacional, competindo com o amido de milho, trigo e o da batata, que possuem propriedades similares. A partir do amido de mandioca podem ser produzidos diversos derivados, ou amidos modificados, utilizados nas indústrias de papel, química, alimentícia, entre outras (VILPOUX, 2008).

Em termos econômicos, a mandioca vem sendo cultivada em regiões tropicais e subtropicais da África, Ásia e América Latina, ocupando atualmente o sexto lugar na produção mundial, atrás da cana-de-açúcar, soja, milho, arroz e trigo (FAO, 2014; NJOKU et al., 2015). No Brasil, a sua produção é de 23 milhões de toneladas produzida em cerca de 2,34 milhões de hectares, ocupando assim, o quarto lugar na produção mundial, após a Nigéria, Tailândia e Indonésia (FAO, 2014; GONÇALVES et al., 2017). Dentre as principais regiões produtoras no Brasil, as regiões Norte e Nordeste destacam-se na produção com 40,0% e 23,1%, respectivamente. Também merece destaque a região Sul, com 20,8% de participação na produção (IBGE, 2017). A estimativa da produção de mandioca para o ano de 2017 é mais de 20 milhões de toneladas, com uma redução de aproximadamente 11,8% em relação ao ano anterior. Entretanto, há uma expectativa de crescimento para a região Nordeste de 1,5%, em função, principalmente, de um aumento de 6,5% no rendimento médio, já que as áreas plantadas e colhidas aumentaram em 6,9% e 4,7%, respectivamente (IBGE, 2017).

Os maiores estados do Brasil produtores de mandioca são o Pará, Paraná e Bahia, com participação de 24,7%, 13,2% e 8,4%, respectivamente, sendo a Bahia o principal produtor da região Nordeste, com cerca de 1,74 milhões de toneladas (IBGE, 2017). Atualmente, a maior concentração de área cultivada encontra-se em regiões de menor industrialização, como é o caso das

regiões Norte e Nordeste, cujo destino das raízes está voltado, sobretudo, para a fabricação de farinha (VILPOUX, 2008). Toda produção brasileira é praticamente voltada para o mercado interno, com menos de 0,5% da produção voltada para o mercado externo (BILIERI et al., 2014).

Alguns fatores são limitantes para a produção da mandioca no Brasil, como a baixa fertilidade dos solos; uso de manivas de baixa qualidade e variedades pouco produtivas e/ou mal adaptadas às regiões de cultivo; a elevada sensibilidade na competição com ervas daninha; bem como o ataque de pragas e doenças e a ocorrência generalizada de deterioração fisiológica pós-colheita (CARDOSO et al., 2013; VENTURINI et al., 2016). Contudo, a mandioca é uma espécie de fundamental importância para o país, por ser uma das culturas de maior relevância para a agricultura de subsistência e segurança alimentar, devido a uma demanda crescente para o desenvolvimento agrícola (KUNKEAW et al., 2011).

**Origem e variabilidade genética da mandioca**

A mandioca é originária da América do Sul, sendo o Brasil seu provável centro de origem e diversidade (OLSEN, 2004). Foi transferida pelos portugueses para todo o mundo, particularmente para o continente africano, no século XVI. Na África e América Latina, o cultivo e seleção de plantas voluntárias oriundas de sementes durante muito tempo pelos pequenos agricultores resultou em inúmeras variedades locais (NWEKE et al., 2002).

A origem da mandioca sempre foi muito controversa. Olsen; Schaal (1999) e Olsen (2004) realizaram estudos visando esclarecer a origem desta espécie com uso de marcadores moleculares, onde avaliaram características morfológicas, geográficas e filogenéticas. Com base em marcadores *Single-Nucleotide Polymorphisms* (SNP) e *Simple-Sequence Repeats* (SSR), os autores concluíram que a mandioca foi domesticada a partir de uma subespécie, a *Manihot esculenta* ssp*. flabellifolia,* possivelmente originária do sudoeste da bacia Amazônica e ancestral da espécie cultivada (ALLEN, 2002; OLSEN, 2004). Recentemente, Léotard et al. (2009) avaliaram a variação no gene *G3pdh* com uma amostragem mais abrangente da atual distribuição da *M. esculenta* ssp*. flabellifolia*, e ainda com outras espécies do gênero *Manihot* potencialmente hibridizadas com a mandioca. Seus resultados reiteram os

resultados anteriores, sugeridos por Olsen (2004). Contudo, ainda existem muitas discussões sobre a origem botânica da mandioca, pois alguns pesquisadores questionam as evidências de domesticação a partir da *M. esculenta* ssp*. flabellifolia*, no sentido de que os estudos que deram suporte a esta argumentação foram temporariamente e espacialmente limitados. Porém, estudos atuais apoiam a hipótese de que a mandioca é derivada da *M. esculenta* ssp. *flabellifolia* a partir do sudoeste da Bacia Amazônica (OLSEN, 2004; LÉOTARD et al., 2009; DUPUTIÉ et al., 2011).

Devido à grande variabilidade existente na cultura, torna-se indispensável a conservação de todo recurso disponível para evitar uma erosão genética, que levaria à perda de genes ou de combinações gênicas favoráveis. Atualmente, a variabilidade da espécie tem sido utilizada na geração de cultivares produtivas e resistentes a fatores bióticos e abióticos, de forma a garantir ampla base genética para subsidiar programas de melhoramento genético da espécie (VIEIRA et al., 2010). No Brasil existe uma ampla diversidade genética, considerando atributos de resistência às principais pragas e doenças, adaptação a diferentes condições edafoclimáticas, produtividade e qualidade de raiz e amido (VENTURINI et al., 2015; FREITAS et al., 2016; VENTURINI et al., 2016; VILAS-BOAS et al., 2016).

Embora seja uma espécie alógama, a mandioca pode ser propagada por manivas (assexuadamente) e sua conservação é feita em campo, em laboratório (*in vitro*) ou por meio de estocagem de sementes botânicas (FUKUDA, 2005). Entretanto, a manutenção da variabilidade genética é feita predominantemente de forma *ex situ*, em campo, seja por produtores locais, ou em bancos ativos de germoplasma e coleções de trabalhos distribuídas em instituições de pesquisa em várias regiões do país, preservando assim, grande parte da variabilidade genética da espécie, que se constitui em atividade fundamental para garantir a competitividade e sustentabilidade da cultura (GOMES et al., 2007).

A conservação *in vitro* é uma forma promissora de preservação do germoplasma, pois permite manter um elevado número de indivíduos em menor espaço físico. Entretanto, é dependente de mão-de-obra especializada e ambiente controlado em laboratório (FUKUDA, 2005; OLIVEIRA et al., 2014b). Já a conservação de sementes é a menos comum, utilizada para espécies

silvestres que apresentam dificuldades de enraizamento por estacas (manivas) e que não toleram a propagação *in vitro* (SECOND; IGLESIAS, 2000).

Os principais Bancos Ativos de Germoplasma de mandioca estão localizados no CIAT (Centro Internacional de Agricultura Tropical, Colômbia), com cerca de 6.000 acessos (OKOGBENIN et al., 2007), IITA (Instituto Internacional de Agricultura Tropical, Nigéria) com aproximadamente 4.000 acessos (DUMET et al., 2011) e na Embrapa (Empresa Brasileira de Pesquisa Agropecuária, Brasil), com aproximadamente 4.000 acessos oriundos de diversas regiões do país e também do exterior.

Apesar da reconhecida variabilidade genética existente, a conservação dos recursos genéticos só é justificada se forem bem caracterizados e avaliados, pois o sucesso de qualquer programa de melhoramento genético depende do conhecimento da magnitude de variação presente na espécie de interesse, permitindo sua plena utilização pelos melhoristas e agricultores. Na mandioca, as características de maior interesse para os melhoristas são a produtividade e qualidade de raízes, teor de matéria seca, produção de amido e resistência a pragas e doenças (VENTURINI et al., 2015; FREITAS et al., 2016).

**Diversidade genética**

Estudos sobre a diversidade genética entre populações de uma espécie é uma atividade essencial que permite descrever o estoque genético conservado, além de subsidiar políticas de exploração e manejo dos recursos vegetais. A partir desta atividade, é possível traçar estratégias que permitem ampliar a base genética da espécie conservada por meio da definição das melhores combinações de genitores (ABACA et al., 2013).

A caracterização morfológica e agronômica proposta por Fukuda e Guevara (1998), foram essenciais para uma descrição fenotípica dos acessos de mandioca. Os descritores propostos por estes autores são divididos em qualitativos (morfológicos) e quantitativos (agronômicos), sendo que esse último possui grande interferência ambiental, havendo dificuldades intrínsecas para seu uso na categorização de acessos, em função da menor herdabilidade das características. Embora esses descritores exerçam papel importante na diferenciação entre variedades, genótipos que são morfologicamente

semelhantes podem ser classificados erroneamente. As condições ambientais e as diferentes fases de desenvolvimento fenológico das plantas exercem influência sobre a caracterização fenotípica, reduzindo assim o poder de distinção entre os acessos (RABBI et al., 2015).

Kawuki et al. (2011) avaliaram a variabilidade genética para descritores qualitativos e quantitativos em mandioca e confirmaram que os qualitativos possuem um poder relativamente limitado de discriminação entre acessos. Por outro lado, os descritores quantitativos mostraram ampla variação, sobretudo, para caracteres de maior interesse agronômico. No entanto, seu aproveitamento é restrito devido à capacidade limitada de discriminação entre os acessos da cultura e elevada influência ambiental. Portanto, um dos principais entraves enfrentados nos bancos ativos de germoplasma tem sido a capacidade de mensuração, registro e comparação de toda a variabilidade genética conservada. Estas atividades são essenciais para descrever o material genético armazenado e subsidiar políticas de exploração e manejo dos recursos, permitindo, assim, traçar estratégias de conservação em grande escala, o que é notório na caracterização fenotípica, na qual as avaliações agronômicas em campo são muito laboriosas, envolvendo um grande número de genótipos e necessitando de muita mão-de-obra, área disponível e recursos financeiros. Portanto, até o momento, a maioria dos trabalhos sobre diversidade genética em mandioca tem utilizado apenas uma pequena amostra de coleções de plantas (OLIVEIRA et al., 2014).

De modo geral, as informações fenotípicas dos trabalhos de caracterização e avaliação de germoplasma são bastante úteis para orientação de cruzamentos, pois as estimativas de diversidade e desempenho agronômico podem orientar as estratégias de conservação e uso do germoplasma de forma simultânea (ZUIN et al., 2009), como também permitem avaliar os padrões de diversidade genética que contribuem significativamente para geração de informações sobre: preservação da variabilidade genética existente; descoberta de novos alelos; constatação de possíveis duplicatas de acessos; identificação de combinações parentais, que ao serem cruzados, possibilitam maior efeito heterótico na progênie e introgressão de genes desejáveis.

A necessidade de fomentar uma base de conhecimento sobre a caracterização fenotípica de acessos pertencentes ao germoplasma é uma

atividade essencial para a identificação e diferenciação de indivíduos. No entanto, as limitações desta abordagem são agora reconhecidas, motivando a busca por estratégias alternativas que permitam caracterizar toda a variabilidade da cultura que afetam o fluxo gênico, responsável pelos padrões de diversidade observados.

Estudos sobre a variabilidade genética do germoplasma de mandioca têm sido realizados com foco na identificação de fontes tolerantes à deterioração fisiológica pós-colheita (VENTURINI et al., 2015); dados agronômicos quantitativos e qualitativos (OLIVEIRA et al., 2015); teor de matéria seca de raízes (OLIVEIRA et al., 2016); identificação de fontes de resistência a doenças, à exemplo da podridão radicular (VILAS-BOAS et al., 2016) e diversidade fenotípica de grânulos de amido (VASCONCELOS et al., 2017). Apesar da ampla diversidade genética relatada na mandioca, a tendência da variabilidade global vem decaindo nos cultivos comerciais em função da substituição de variedades locais por variedades melhoradas (WILLEMEN et al., 2007). Em razão disto, toda estratégia relacionada à avaliação e caracterização de germoplasma é uma poderosa ferramenta para promover o uso de recursos genéticos de mandioca objetivando o progresso genético da cultura.

**Utilização de marcadores moleculares na cultura da mandioca**

Além da caracterização fenotípica, tem-se constatado ampla variabilidade genética entre genótipos de uma mesma espécie com uso de marcadores moleculares, que proporcionam maior grau de informação pelo fato de não sofrerem influência ambiental. Isto permite aos programas de melhoramento genético estimar a diversidade existente nas espécies independente do ambiente de cultivo para elucidar questões sobre filogenia e identificação de duplicatas de acessos (MOURA et al., 2013).

Na mandioca, os marcadores moleculares tem sido empregados em estudos de diversidade genética, incluindo marcadores R*andom Amplified Polymorphic DNA* (RAPD) (VIEIRA et al., 2010), A*mplified Fragment Length Polymorphism* (AFLP) (RAJI et al., 2009) e *Simple-Sequence Repeats* (SSR) (GONÇALVES et al., 2017), sendo estes últimos os mais vantajosos em relação aos anteriores, devido à alta reprodutibilidade, além do alto poder

discriminativo; características importantes que justifica seu uso em diversos estudos de genética populacional.

Atualmente, os marcadores *Single-Nucleotide Polymorphism* (SNP) têm sido bastante utilizados em virtude da sua abundância no genoma, baixa taxa de mutação, facilidade de automação e localização específica no cromossomo, o que vem intensificando sua aplicação em estudos de ecologia, evolução, conservação (HELYAR et al., 2011) e diversidade genética em diferentes espécies (SPANIC et al., 2016), inclusive na cultura da mandioca (OLIVEIRA et al., 2014).

Apesar de não serem considerados marcadores multialélicos, como os microssatélites, os SNPs se destacam na sua aplicabilidade devido a recentes tecnologias desenvolvidas para genotipagem, envolvendo o processamento de um grande número de marcadores e amostras simultaneamente, agregando menor custo, maior rendimento e nível de reprodutibilidade superior em relação a outros tipos de marcadores moleculares (ELSHIRE et al., 2011).

Recentes técnicas têm sido desenvolvidas com o intuito de auxiliar o uso de marcadores moleculares. Luikart et al. (2003) afirmaram que uma abordagem molecular ideal para genômica populacional deve descobrir centenas de marcadores polimórficos que cubram todo o genoma em uma única, simples e confiável análise. Assim, a tecnologia *Next-Generation Sequencing* (NGS) disponibiliza várias abordagens, que são capazes de descobrir, sequenciar e genotipar milhares de marcadores em quase todos os genomas de interesse em um único passo, mesmo em populações em que pouca ou nenhuma informação genética esteja disponível (DAVEY et al., 2011). Esta tecnologia vem sendo aplicada em larga escala, abrindo oportunidades fascinantes na biologia, e tornando possível a genotipagem de grandes populações e projetos de ressequenciamento de genomas, a fim de identificar e mapear amplo número de SNPs.

O aumento na densidade de marcadores, além de produzir quantidades enormes de dados de sequenciamento da ordem de bilhões de bases, amplamente distribuídos ao longo do genoma, permite que muitas questões biológicas sejam respondidas. Alguns exemplos referem-se à exploração da diversidade das espécies; construção de mapas genéticos e realização de estudos sobre associação genômica *genome-wide association study* (GWAS);

identificação de pontos de interrupção de recombinação para o mapeamento de ligação ou o mapeamento de *quantitative trait locus* (QTL), localizando regiões genômicas diferenciadas entre populações para estudos de genética quantitativa; genotipagem de grande número de indivíduos para seleção assistida por marcadores ou estudos de filogenia de dezenas de populações silvestres de forma simples e robusta, alterando a relação custo/benefício por sequenciamento e oferecendo um preço final mais atraente e em menor tempo (DAVEY et al., 2011; ELSHIRE et al., 2011; RABBI et al., 2015).

Uma das plataformas de sequenciamento de próxima geração amplamente disponível é a *genotyping-by-sequencing* (GBS), que combina a descoberta de polimorfismo e a genotipagem em um único passo, oferecendo um processo de produção de biblioteca simplificada mais propícia para ser utilizada em grande número de indivíduos. Por meio desta técnica, ao invés de se definir a priori quais SNPs serão genotipados ao longo do genoma, a metodologia detecta e promove a genotipagem dos SNPs, para uma seleção a posteriori dos marcadores que serão empregados nas mais diversas análises genéticas (ELSHIRE et al., 2011), isto é, o método não demanda o desenvolvimento prévio de marcadores e se baseia em reagentes universais, assim como na técnica de RAPD, que permitiu, em sua época, um avanço notável na capacidade de realizar análises genéticas moleculares em qualquer espécie.

A GBS é adequada para estudos de populações, caracterização de germoplasma, mapeamento de diversas características em qualquer espécie, oferecendo baixo custo por amostra, com sequências de tamanho fracionado e, orientadas por enzimas de restrição (ELSHIRE et al., 2011). Pelo método GBS, é possível gerar um grande número de SNPs, e isso tem sido uma ferramenta valiosa para análises de estrutura populacional na mandioca (PEREA et al., 2016), pois permite analisar a variação genética presente entre genótipos de forma cada vez mais eficaz.

A técnica envolve a digestão do DNA genômico alvo utilizando uma enzima de restrição, gerando uma biblioteca de fragmentos com tamanhos entre 200 e 400 pb (pares de base). Esta metodologia tem como diferencial o fato de que, para cada indivíduo a ser genotipado, uma sequência indexadora conhecida como “bacorde” ou “código de barras” contendo de 4 a 8 pares de

base de DNA, é incluída no adaptador. Dessa forma, cada amostra a ser genotipada terá uma sequência “bacorde” única, permitindo sequenciar centenas de milhares de amostras simultaneamente de modo multiplex. As sequências contendo de 75 a 100 pares de base, geradas para cada amostra, podem ser identificadas pelo seu “bacorde” e, assim, recuperadas individualmente, permitindo a genotipagem dos marcadores SNPs em cada indivíduo (ELSHIRE et al., 2011). No entanto, a redução da complexidade do genoma com uso de enzimas de restrição deve ser empregada para assegurar suficiente sobreposição na cobertura de sequência para espécies com genomas grandes, como no caso da mandioca.

A GBS tem sido amplamente utilizada em diversos germoplasma, como trigo (POLAND et al., 2012), soja (SONAH et al., 2013), sorgo (THURBER et al., 2013), arroz (SPINDEL et al., 2013) e recentemente na cultura da mandioca (RABBI et al., 2015), onde os resultados significativos referentes a esses estudos demonstraram a eficácia da GBS para a compreensão da diversidade molecular e estudos de filogenia em diversas plantas cultivadas (HUANG et al., 2014). Assim, considerando a aplicação bem-sucedida do método GBS em análises genéticas de alto rendimento em outras espécies, as informações obtidas no germoplasma da mandioca, pelo método GBS, certamente permitirão: i) o entendimento da diversidade genética e das relações entre acessos, que são essenciais para programas de melhoramento genético, auxiliando pesquisadores nas tomadas de decisões para restaurar a variação genética perdida, por meio de expedições de coleta direcionadas; ii) a identificação de progenitores contrastantes para hibridações intraespecíficas; iii) de forma direta e indireta contribuirá para melhoria dos níveis de resistência às principais pragas e doenças que acometem a cultura, como também o rendimento médio de cultivo, indicando cada vez mais, indivíduos promissores que atendam a demanda de mercado de forma precisa e robusta (FERGUSON et al., 2012).

**Identificação de acessos duplicados no germoplasma da mandioca**

Segundo Robichaud et al. (2006), grande parte dos acessos de mandioca estão armazenados em bancos ativos de germoplasma, visando conservar e utilizar de modo racional todo recurso genético disponível.

Entretanto, o trabalho de coleta dos acessos é realizado em épocas e locais diferentes. Assim, é comum a utilização de nomes distintos, que na verdade tratam-se do mesmo genótipo, nesse processo, os genótipos perdem sua identidade original. Também é possível que uma pobre classificação do material, que normalmente é baseado em poucos descritores morfológicos possa dar origem à presença de material idêntico registrado com nomes diferentes (KILIAN; GRANER, et al., 2012). Ademais, as condições ambientais e os diferentes estádios de desenvolvimento da planta influenciam tais descritores e, finalmente, o número de descritores se torna cada vez mais limitado à medida que são desenvolvidas variedades para se adequarem aos ideótipos desejados (RABBI et al., 2015). Já no caso de materiais selvagens, as expedições de coletas podem ter sido organizadas em áreas idênticas ou o material coletado pode ter sido adquirido sem conhecimento prévio suficiente sobre a distribuição da variação genética nas áreas naturais de coleta (VAN TREUREN; VAN HINTUM, 2003).

Todas estas inconsistências têm resultado na ocorrência de acessos repetidos, que comprometem a conservação e o avanço nos estudos do germoplasma. Por este motivo, é de grande importância trabalhos que visam à identificação de duplicatas em BAGs, resultando em uma diminuição do espaço físico, tornando ágil e dinâmico o manejo e conservação, e consequentemente, gerando diminuição de custos na manutenção das plantas em campo ou *in vitro* (MOURA et al., 2013).

Na cultura da mandioca, diversos estudos mostram a existência de materiais duplicados nos germoplasma e o quanto essas duplicatas prejudicam os avanços no melhoramento da espécie, principalmente no que diz respeito à manutenção destes acessos em campo, que consequentemente levam à redução da representatividade de toda diversidade genética conservada. (MOURA et al., 2013; RABBI et al., 2015; MOURA et al., 2016).

As primeiras formas de identificação de duplicatas foram realizadas com base em dados morfoagronômicos, de acordo com Fukuda e Guevara (1998), em que os clones são diferenciados apenas fenotipicamente. Apesar da importância desse tipo de caracterização em bancos de germoplasma, as características quantitativas constituem uma alternativa viável na identificação de duplicatas para caracteres de interesse agronômico e as características

qualitativas possibilitam estabelecer graus de similaridade entre diferentes clones. Contudo, esse tipo de caracterização não permite uma diferenciação completa entre acessos (OLIVEIRA et al., 2015).

Diante disso, a forma mais acurada para identificação de duplicatas é pela genotipagem molecular, já que os marcadores moleculares mostram regiões aleatórias do genoma permitindo diferenciar genótipos fenotipicamente similares e não sofrem influência ambiental. A partir do estabelecimento da relação genética entre os acessos, é possível traçar estratégias de conservação e uso do germoplasma nos programas de melhoramento genético para geração de novas variedades com características diferenciais àquelas já utilizadas no sistema de produção (EGBADZOR et al., 2014).

O número de trabalhos na literatura mostrando a eficiência dos SNPs na detecção de duplicatas tem aumentado, principalmente em culturas de importância econômica, como feijão-caupi (EGBADZOR et al., 2014), soja (SONG et al., 2015), kiwi (MELO et al., 2017), e mandioca (RABBI et al., 2015). Segundo Horna et al. (2010), o custo da identificação de um acesso de mandioca duplicado utilizando a caracterização molecular, é 12 vezes menor do que o custo de se conservar e utilizar o material como um acesso diferente em um germoplasma. Ademais, a tecnologia NGS tem permitido a identificação de milhares de marcadores no genoma de interesse em um único passo, com custos mais acessíveis e aliados a modelos estatísticos rigorosos, permitem a identificação de materiais duplicados com maior confiabilidade, pois, tem sido mais vantajoso para curadores e melhoristas obter todo conhecimento possível acerca dos materiais conservado, em vez de ampliar o germoplasma propriamente dito.

A identificação de acessos duplicados é um caminho importante para otimização da conservação do germoplasma de mandioca e priorização das atividades de caracterização e avaliação fenotípica para descoberta de novos genes e características funcionais. Isto se faz necessário, pois a conservação em campo envolve elevados custos financeiros e mão-de-obra, sobretudo para a realização de plantio, colheita, tratos culturais, e controle de pragas e doenças (JOACHIM-KELLER et al., 2013).

De acordo com Chavarriaga-Aguirre et al. (1999), os custos anuais de manutenção de um BAG de mandioca giram em torno de U$17,09 (dólares

americanos) anual por acesso mantido em campo. Segundo estes mesmos autores, a caracterização de acessos a fim de ser evitar materiais duplicados podem reduzir os custos de manutenção em torno de 20 a 25%. Um recente estudo no germoplasma do alho (*Allium sativum* L.), os autores afirmaram que somente os custos com mão-de-obra representam cerca de 81% dos custos anuais. (JOACHIM-KELLER et al., 2013).

**REFERÊNCIAS**

ABACA, A.; KAWUKI, R.; TUKAMUHABWA, P.; BAGUMA, Y.; PARIYO, A.; ALICAI, T.; OMONGO, A. C. C.; ABIDRABO, P.; KATONO, K., BUA, A. Genetic relationships of cassava genotypes that are susceptible or tolerant to cassava brown streak disease in Uganda. **Journal of Agricultural Science,** v. 5, p. 107-115, 2013.

ALLEN, A. C. The origin and taxonomy of cassava. In: HILLOCKS, R.J.; THRESH, J. M.; BELLOTTI, A. Cassava: biology, production and utilization. **New York**: **Wallingford**, UK. v. 13, p. 1-16, 2002.

BILIERI, C. E.; SILVA, R. H.; GOMES, C. M.; CARDOZO, C. J.; YAMASHITA, O. M.; CARVALHO, M. A. C.; ROBOREDO, D.; GERVAZIO, W. Organoleptic quality in different varieties of cassava, implemented in consortium with the pineapple in the municipality of Alta Floresta – MT. **Cadernos de Agroecologia.** v. 9, p. 1-9. 2014.

CARDOSO, A. D.; VIANA, A. E. S.; BARBOSA, R. P.; TEIXEIRA, P. R. G.; CARDOSO JÚNIOR, N.; FOGAÇA, J. J. N. L. Levantamento fitossociológico de plantas daninhas na cultura da mandioca em Vitória da Conquista, Bahia. **Bioscience Journal**, v. 29, p. 1130-1140, 2013.

CHAVARRIAGA-AGUIRRE, P.; MAYA, M. M.; TOHME, J.; DUQUE, M. C.; IGLESIAS, C.; BONIERBALE, M. W.; KRESOVICH, S.; KOCHERT, G. Using microsatellites, isozymes and AFLPs 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.

DAVEY, J. W.; HOHENLOHE, P. A.; ETTER, P. D.; BOONE, J. Q.; CATCHEN, J. M.; BLAXTER, M. L. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nature**, v. 12, p. 499-510, 2011.

DUMET, D.; KORIE, S.; ADEYEMI, A. Cryobanking Cassava Germplasm at IITA. **Acta Horticulturae**, v. 908, p. 439-446, 2011.

DUPUTIÉ, A.; SALICK, J.; McKEY, D. Evolutionary biogeography of *Manihot* (Euphorbiaceae), a rapidly radiating Neotropical genus restricted to dry environments. **Journal of Biogeography**, v. 38, p. 1033-1043, 2011.

EGBADZOR, F. K.; OFORI, K.; YEBOAH, M.; LAWRENCE, M. A.; OPOKU-AGYEMAN, O. M.; DANQUAH, Y. E.; OFFEI, K. S. Diversity in 113 cowpea [*Vigna unguiculata* (L) Walp] accessions assessed with 458 SNP markers. **Springer Plus,** v. 3, p. 541-556, 2014.

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

FRANCO, C. M. L.; DAIUTO, E. R.; DEMIATE, I. M.; CARVALHO, L. J. C. B.; LEONEL, M.; CEREDA, M. P.; VILPOUX, O. F.; SARMENTO, S. B. S. Propriedades do Amido. In: Série Culturas de Tuberosas Amiláceas Latino Americanas, **Propriedades Gerais do Amido**, CEREDA, M.P. (coord.). São Paulo: Fundação Cargill, v. 1, p. 224, 2001.

FERGUSON, M. E.; HEARNE, S. J.; CLOSE, T. J.; WANAMAKER, S.; MOSKAL, W. A.; TOWN, C. D.; YOUNG, J.; MARRI, P. R.; RABBI, I. Y.; VILLIERS, E. P. Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. **Theoretical Applied Genetics**, v. 124, p. 685–695, 2012.

FREITAS, J. P. X.; SANTOS, V. S.; OLIVEIRA, E. J. Inbreeding depression in cassava for productive traits. **Euphytica**, v. 209, p.137–145, 2016.

FUKUDA, W. M. G.; COSTA, I. R. S.; SILVA, S. O. Manejo e conservação de recursos genéticos de mandioca (*Manihot esculenta* Crantz) na Embrapa Mandioca e Fruticultura. Cruz das Almas: **EMBRAPA-CNPMF**, v. 18, p. 4, 2005.

FUKUDA, W. M. G.; GUEVARA, C. L. Descritores morfológicos e agronômicos para a caracterização de mandioca (*Manihot esculenta* Crantz). Cruz das Almas: EMBRAPA-CNPMF, **Documento 78**, p. 38, 1998.

GOMES, C. N.; CARVALHO, S. P.; JESUS, A. M. S.; CUSTÓDIO, T. N. Caracterização morfoagronômica e coeficientes de trilha de caracteres componentes da produção em mandioca. **Pesquisa Agropecuária Brasileira**, v. 42, p. 1121-1130, 2007.

GONÇALVES, T. M.; VIDIGAL FILHO, P. S.; VIDIGAL, M. C. G.; FERREIRA, R. C. U.; ROCHA, V. P. C.; ORTIZ, A. H. T.; MOIANA, L. D.; KVITSCHAL, M. V. Genetic diversity and population structure of traditional sweet cassava accessions from Southern of Minas Gerais State, Brazil, using microsatellite markers. **African Journal of Biotechnology**, v. 16, p. 346-358, 2017.

GROSS, B. L.; VOLK, G. M.; RICHARDS, C. M. Identification of “duplicate” accessions within the USDA-ARS National plant germplasm system *Malus* collection. **Journal of the American Society for Horticultural Science**, v.137, p. 333–342, 2012.

HELYAR, S. J.; HEMMER-HANSEN-HEMMER, J.; BEKKEVOLD, D.; TAYLOR, M. L.; OGDEN, R.; LIMBORG, M. T.; CARIANI, A.; MAES, G. E.; DIOPERE, E.; CARVALHO, G. R.; NIELSEN, E. E. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. **Molecular Ecology Resources**, v. 11, p. 123-136, 2011.

HORNA, D.; DEBOUCK, D.; CIPRIAN, A. Conservation and management of genetic resources of beans, cassava and tropical forages in the CIAT genebank. In: HORNA, D.; DEBOUCK, D.; DUMET, D.; HANSON, J.; PAYNE, T.; SACKVILLE-HAMILTON, R.; SANCHEZ, I.; UPADHYAYA, H. D.; VAN DEN HOUWE I. Evaluating cost effectiveness of collection management: *ex-situ* conservation of plant genetic resources in the CG system. **Consultative Group on International Agricultural Research**, p. 24-35, 2010.

HUANG, B. E.; RAGHAVAN, C.; MAULEON, R.; BROMAN, K. W.; LEUNG, H. Efficient imputation of missing markers in low-coverage genotyping-by-sequencing data from multi-parental crosses. **Genetics**, v. 205, p. 1-16, 2014.

IBGE - INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Levantamento Sistemático da produção Agrícola-LSPA. Disponível em: <ftp://ftp.ibge.gov.br/Producao_Agricola/Levantamento_Sistematico_da_Producao_Agricola[mensal]/Fasciculo/lspa_201701.pdf. > Acesso em: 06 março 2017.

IRISH, B. M.; GOENAGA, R.; ZHANG, D.; SCHNELL, R.; BROWN, J. S.; MOTAMAYOR, J. C. Microsatellite fingerprinting of the USDA-ARS tropical agriculture research station cacao (*Theobroma cacao* L.) germplasm collection. **Crop Science**, v. 50, p. 656-667, 2010.

KAWUKI, R. S.; FERGUSON, M.; LABUSCHAGNE, M. T.; HERSELMAN, L.; ORONE, J.; RALIMANANA, I.; BIDIAKA, M.; LUKOMBO, S.; KANYANGE, M. C.; GASHAKA, G.; MKAMILO, G.; GETHIH, J.; OBIERO, H. Variation in qualitative and quantitative traits of cassava germplasm from selected national breeding programmers in sub-Saharan Africa**. Field Crops Research,** v. 2, p. 151–156, 2011.

JOACHIM-KELLER, E. R.; ZANKE, C. D.; SENULA, A.; BREUING, A.; HARDEWEG, B.; WINKELMANN, T. Comparing costs for different conservation strategies of garlic (*Allium sativum* L.) germplasm in genebanks. **Genetic Resources and Crop Evolution**. v. 60, p. 913–926, 2013.

KILIAN, B.; GRANER, A. NGS technologies for analyzing germplasm diversity in genebanks. **Briefings in Functional Genomics**, v. 11, p. 38-50, 2012.

KUNKEAW, S.; YOOCHA, T.; SRAPHET, S.; BOONCHANAWIWAT, A.; BOONSENG, O.; LIGHTFOOT, D. A.; TRIWITAYAKORN, K.; TANGPHATSORNRUANG, S. Construction of a genetic linkage map using simple sequence repeat markers from expressed sequence tags for cassava (*Manihot esculenta* Crantz). **Molecular Breeding**, v. 27, p. 67-75, 2011.

LATIF, S.; MULLER, J. Potential of cassava leaves in human nutrition: A review. **Trends in Food Science & Technology**, v. 44, p. 147-158, 2015.

LEBOT, V. Tropical root and tuber crops: cassava, sweet potato, yams and aroids. **Crop Production Science in Horticulture,** v. 2, p. 432, 2009.

LÉOTARD, G.; DUPUTIÉ, A.; KJELLBERG, F.; DOUZERY, E.J.P.; DEBAIN, C.; GRANVILLE, J.J.; MCKEY, D. Phylogeography and the origin of cassava: new insights from the Northern rim of the Amazonian basin. **Molecular Phylogenetics and Evolution**, v. 53, p. 329-334, 2009.

LUIKART, G.; ENGLAND, P.; TALLMON, D.; JORDAN, S.; TABERLET, P. The power and promise of population genomics: from genotyping to genome typing. **Nature Reviews Genetics**, v. 4, p. 981–994, 2003.

MELO, A. T. O.; GUTHRIE, R. S.; HALE, I. GBS-Based Deconvolution of the Surviving north American collection of cold-hardy kiwifruit (*Actinidia* spp.) germplasm. **Plos One**, v. 12, p. 1-21, 2017.

MOURA, F. E.; FARIAS NETO, T. J.; SAMPAIO, E. J.; SILVA, T. D.;

RAMALHO, F. G. Identification of duplicates of cassava accessions sampled on the North Region of Brazil using microsatellite markers. **Acta Amazonica**, v. 43, p. 461-468, 2013.

MOURA, E. F.; SOUSA, N. R.; MOURA, M. F.; DIAS, M. C.; SOUZA, E. D.; FARIAS NETO, J. T.; SAMPAIO, J. E. Molecular characterization of accessions of a rare genetic resource: sugary cassava (*Manihot esculenta* Crantz) from Brazilian Amazon. **Genetic Resources and Crop Evolution**, v. 63, p. 583-593, 2016.

NJOKU, N. D.; GRACEN, E. V.; OFFEI, K. S.; ASANTE, K. I.; EGESI, N. C.; KULAKOW, P.; CABALLOS, H. Parent-offspring regression analysis for total carotenoids and some agronomic traits in cassava. **Euphytica**, v. 205, p. 1-10, 2015.

NWEKE, F. I.; SPENCER, D. S. C.; LYNAM, J. K. The cassava transformation. Africa's best kept secret. **Michigan State University, East Lansing**. v. 6, p.12, 2002.

OKOGBENIN, E.; PORTO, M. C. M. EGESI, C.; MBA, C.; ESPINOSA, E.; SANTOS, L. G.; OSPINA, C.; MARÍN, J.; BARRERA, E.; GUTIÉRREZ, J.; EKANAYAKE, I.; IGLESIAS, C.; FREGENE, M. A. Marker-Assisted Introgression of resistance to cassava mosaic disease into Latin American germplasm for the genetic improvement of cassava in Africa. **Crop Science**, v. 47, p. 1895-1904, 2007.

OLIVEIRA, E. J.; FERREIRA, C. F.; SANTOS, V. S.; JESUS, O. N.; OLIVEIRA, G. A. F.; SILVA, M. S. Potential of SNP markers for the characterization of Brazilian cassava germplasm, **Theoretical and Applied Genetics**, v. 127, p. 1423-1440, 2014.

OLIVEIRA, E. J.; FERREIRA, C. F.; SANTOS, V. S.; OLIVEIRA, G. A. F. Development of a cassava core collection based on single nucleotide polymorphism markers. **Genetics and Molecular Research**, v. 13, p. 6472-

6485, 2014b.

OLIVEIRA, E. J.; OLIVEIRA FILHO, O. S.; SANTOS, V. S. Classification of cassava genotypes based on qualitative and quantitative data. **Genetics and Molecular Research**, v. 14, p. 906-924, 2015.

OLIVEIRA, E, J.; SANTOS, P. E. F.; PIRES, A. J. V.; TOLENTINO, D. C.; SANTOS, V. S. Selection of cassava varieties for biomass and protein production in semiarid areas from Bahia. Bioscience Journal, v. 32, p. 661-669, 2016.

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

OLSEN, K. M.; SCHAAL, B. A. Evidence on the origin of cassava: Phylogeography of *Manihot esculenta*. **Proceedings of the National Academy of Sciences of the United States of America**, v. 96, p. 5586-5591, 1999.

PENTEADO, M. V. C.; FLORES, C. I. O. Folhas de mandioca como fonte de nutrientes. In: CEREDA, M. P. (coord): **Manejo, Uso e Tratamento de Subprodutos da Industrialização da Mandioca**. São Paulo: Fundação CARGILL, vol. IV, p. 49-65, 2001.

PEREA, C.; HOZ, J. F. L.; CRUZ, D. F.; LOBATON, J. D.; IZQUIERDO, P.; QUINTERO, J. C.; RAATZ, B.; DUITAMA, J. Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. **BioMed Central Genomics**, v. 17, p. 540-569, 2016.

POLAND, J.; ENDELMAN, J.; DAWSON, J.; RUTKOSKI, J.; SHUANGYE, W.; MANES, Y.; DREISIGACKER, S.; CROSSA, J.; SÁNCHEZ-VILLEDA, H.; SORRELLS, M.; JANNINK, J-L. Genomic selection in wheat breeding using genotyping-by-sequencing. **The Plant Genome**, v. 5, p. 103-113, 2012.

POOTAKHAM, W.; SHEARMAN, J. R.; RUANG-AREERATE, P.; SONTHIROD,

C.; SANGSRAKRU, D.; JOMCHAI, N.; YOOCHA, T.; TRIWITAYAKORN, K.; TRAGOONRUNG, S.; TANGPHATSORNRUANG, S. Large-scale SNP discovery through RNA sequencing and SNP genotyping by targeted enrichment sequencing in cassava (*Manihot esculenta Crantz*). **Plos One**, v. 9, p. 1-19, 2014.

RABBI, I, Y.; KULAKOW, P. A.; MANU-ADUENING, J. A.; DANKYI, A. A.; ASIBUO, J. Y.; PARKES, E. Y.; ABDOULAYE, T.; GIRMA, G.; GEDIL, M. A.; RAMU, P.; REYES, B.; MAREDIA, M. K. Tracking crop varieties using genotyping-by-sequencing markers: a case study using cassava (*Manihot esculenta* Crantz). **BioMed Central Genetics**, v. 16, p. 1-11, 2015.

RAJI, A. A. J.; FAWOLE, I.; GEDIL, M.; DIXON, O. G. A. Genetic differentiation analysis of African cassava (*Manihot esculenta*) landraces and elite germplasm using amplified fragment length polymorphism and simple sequence repeat markers. **Annals of Applied Biology,** v. 155, p. 187-199, 2009.

ROBICHAUD, R.; GLAUBITZ, J. C.; RHODES J. R. O. E.; WOESTE, K. A robust set of black walnut microsatellites for parentage and clonal identification. **New Forests**, v. 32, p. 179-196, 2006.

SECOND, G.; IGLESIAS, C. The state of the use of cassava genetic diversity and a proposal to enhance it. In: COOPER, H. D.; SPILLANE, C.; HODGKIN, T. **Broadening the genetic base of crop production**, **CABI**. v. 11, p. 201-222, 2000.

SILVA, P. V. K.; ALVES, C. A. A.; MARTINS, G. I. M.; MELO, F. A. C.; CARVALHO, R. Genetic variation among accessions of the genus *Manihot* by ISSR markers. **Pesquisa Agropecuária Brasileira,** v. 46, p. 1082-1088, 2011.

SINGH, J.; KAUR, L.; MCCARTHY, O. J. Factors influencing the physicochemical, morphological, thermal and rheological properties of some chemically modified starches for food applications: a review. **Food Hydrocolloids**, v.21, p.1-22, 2007.

SONAH, H.; BASTIEN, M.; IQUIRA, E.; TARDIVEL, A.; LÉGARÉ, G.; BOYLE, B.; NORMANDEAU, E.; LAROCHE, J.; LAROSE, S.; JEAN, M.; BELZILE, F. An improved genotyping by sequencing (GBS) approach offering increased, versatility and efficiency of SNP discovery and genotyping. **Plos One**, v. 8, p. 1-9, 2013.

SONG, Q.; HYTEN, D. L.; JIA, G.; QUIGLEY, C. V.; FICKUS, E. W.; NELSON, R. L.; CREGAN, P. B. Fingerprinting soybean germplasm and its utility in genomic research. **G3: GENES, GENOMES, GENETICS**, v. 5, p. 1999-2006, 2015.

SPINDEL, J.; WRIGHT, M.; CHEN, C.; COBB, J.; GAGE, J.; HARRINGTON, S.; LORIEUX, M.; AHMADI, N.; McCOUCH, S. Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. **Theoretical and Applied Genetics**, v. 126, p. 2699-2716, 2013.

SPANIC, V.; KORZUN, V.; EBMEYER, E. Assessing genetic diversity of wheat genotypes from different origins by SNP markers. **Cereal Research Communications**, v. 43, p. 361-369, 2016.

TARDIVEL, A.; SONAH, H.; BELZILE, F.; O'DONOUGHUE, L. S. Rapid identification of alleles at the soybean maturity gene E3 using genotyping by sequencing and a haplotype-based approach. **The Plant Genome**, v. 7, p. 1-9, 2014.

THURBER, C. S.; MA, M. J.; HIGGINS, R. H.; BROWN, P. Retrospective genomic analysis of sorghum adaptation to temperate-zone grain production. **Genome Biology**, v. 14, p. 1-12, 2013.

VAN TREUREN, R.; VAN HINTUM, T. J. L. Marker-assisted reduction of redundancy in germplasm collections: genetic and economic aspects. **Acta Horticulturae**, v. 623, p. 139-149, 2003.

VASCONCELOS, L. M.; BRITO, A. C.; CARMO, C. D.; OLIVEIRA, P. H. G. A.; OLIVEIRA, E. J. Phenotypic diversity of starch granules in cassava germplasm. **Genetics and Molecular Research**, v. 16, p. 1-15, 2017.

VENTURINI, M. T.; ARAÚJO, T. S.; ABREU, E. F. M.; ANDRADE, E. C.; SANTOS, V. S.; SILVA, M. R.; OLIVEIRA, E. J. Crop losses in Brazilian cassava varieties induced by the cassava common mosaic virus. **Scientia Agricola**, v. 73, p. 520-524, 2016.

VENTURINI, M. T.; SANTOS, L. R.; VILDOSO, C. I. A.; SANTOS, V. S.; OLIVEIRA, E. J. Variation in cassava germplasm for tolerance to post-harvest physiological deterioration. **Genetics and Molecular Research**, v. 15, p.1-18, 2015.

VIEIRA, A. E.; FIALHO, J. F.; FALEIRO, F. G.; BELLON, G.; FONSECA, K. G. D.; CARVALHO, L. J. C. B.; SILVA, M. S. Caracterização molecular e variabilidade genética de acessos elite de mandioca para fins industriais**.** **Ciência Rural**, v. 40, p. 1-5, 2010.

VILAS-BOAS, S. A.; HOHENFELD, C. S.; OLIVEIRA, S. A. S.; SILVA S. V.; OLIVEIRA. E. J. Sources of resistance to cassava root rot caused by *Fusarium* spp.: a genotypic approach. **Euphytica**, v. 209, p. 237-251, 2016.

WILLEMEN, L.; SCHELDEMAN, X.; CABELLOS, V. S.; SALAZAR, S. R.; GUARINO, L. Spatial patterns of diversity and genetic erosion of traditional cassava (*Manihot esculenta* Crantz) in the Peruvian Amazon: An evaluation of socio-economic and environmental indicators. **Genetic Resources and Crop Evolution**, v. 54, p. 1599-1612, 2007.

VILPOUX, F.O**.** Competitividade da mandioca no Brasil, como matéria-prima para amido. **Informações Econômicas**, v. 38, p. 27-38, 2008.

ZUIN, C. G.; VIDIGAL FILHO, S. P.; KVITSCHAL, V. M.; GONÇALVES-VIDIGAL, C. M.; COIMBRA, K. G. Genetic divergence among accesses of

sweet-cassava collected from Cianorte county, northwestern region of Paraná State. **Semina: Ciências Agrárias,** v. 30, p. 21-30, 2009.

Capítulo 1

# DIVERSIDADE GENÉTICA DO GERMOPLASMA DE *Manihot esculenta* Crantz COM BASE EM MARCADORES SNP

# Diversidade genética do germoplasma de *Manihot esculenta* Crantz com base em marcadores SNP

**RESUMO:** A mandioca (*Manihot esculenta* Crantz) é uma espécie cuja diversidade natural lhe assegura um enorme potencial alimentar, industrial e energético. O objetivo deste estudo foi avaliar a diversidade genética e a estrutura populacional de 1.580 acessos pertencentes ao germoplasma internacional conservado no Brasil, com uso de 20.601 marcadores *Single-Nucleotide Polymorphism* (SNP). Os valores médios de conteúdo de informação polimórfica (PIC), endogamia (*f*), heterozigosidade observada (Ho) e esperada (He) foram de 0,24; 0,21; 0,23; e 0,30; respectivamente, apresentando níveis de diversidade genética compatíveis com a natureza do marcador (predominantemente bialélico) e com o sistema reprodutivo da espécie. Em nível cromossômico, os valores médios de He, Ho, PIC e *f* foram muito similares nos 18 cromossomos da espécie. Em nível de indivíduo, os valores de *f* situaram-se entre 0,49 a 0,97, com média de 0,69, sendo que três acessos de mandioca apresentaram f >0,90. Os valores de desequilíbrio de ligação (LD) se estenderam entre 15 e 20 kb ($r^2$= 0,20). Em relação à análise discriminante de componentes principais (ADCP), foram formados 22 grupos com probabilidade de alocação dos indivíduos acima de 0,99. Não foi observada associação entre o agrupamento da ADCP e agrupamentos com base em dados fenotípicos (teor de compostos cianogênicos e cor da polpa das raízes) e de origem genética e geográfica. Contudo, o agrupamento ADCP resultou em uma maior variação entre grupos (14,85%). Estas informações sugerem extensivo intercâmbio de germoplasma de mandioca no Brasil e ampla diversidade molecular e fenotípica. Esses resultados proporcionarão um melhor entendimento sobre a variabilidade genética conservada e, a organização populacional do germoplasma.

**Palavras chave:** AMOVA; estrutura populacional; mandioca; variabilidade

# Genetic diversity of *Manihot esculenta* Crantz germplasm based on SNP markers

**ABSTRACT:** Cassava *(Manihot esculenta* Crantz) is a species whose natural diversity has great food, industrial and energy potential. This study aimed to evaluate the genetic diversity and population structure of 1,580 accessions belonging to the international germplasm conserved in Brazil using 20,601 *Single-Nucleotide Polymorphism (*SNP) markers. The average values of polymorphic information content (PIC), inbreeding *(f),* observed heterozygosity (Ho) and expected heterozygosity (He) was 0.24; 0.21; 0.23; and 0.30; respectively, presenting levels of genetic diversity compatible with the nature of the marker (predominantly biallelic) and with the species reproductive system. The mean values of He, Ho, PIC and *f* were similar for the 18 chromosomes of species. At individual level, *f* values ranged from 0.49 to 0.97; with an average of 0.69, whereas some cassava accessions showed *f* > 0.90. The values of linkage disequilibrium (LD) extended from 15 to 20 kb ($r^2$=0.20). Regarding the discriminant analysis of principal components (DAPC), 22 groups were formed with probability of allocation of individuals above 0.99. No association was observed between DAPC grouping and clusters based on phenotypic data (cyanogenic compound content and color of the root pulp) and of genetic and geographic origin. However, the DAPC grouping resulted in a greater variation between groups (14.85%). This information suggests extensive exchange of cassava germplasm in Brazil and a wide molecular and phenotypic diversity. These results will provide a better understanding of the preserved genetic variability and the population organization of the germplasm.

**Keywords:** AMOVA; cassava; population structure; variability

## INTRODUÇÃO

O Brasil é considerado o provável centro de origem e diversidade do gênero *Manihot*, composto aproximadamente, por 98 espécies, as quais, possivelmente, foram originadas de cruzamentos naturais e hibridações interespecíficas (OLSEN, 2004; BREDESON et al., 2016). Este gênero pertence às Euphorbiaceae, uma família de angiospermas que inclui várias outras espécies de importância econômica como a mamona, pinhão-manso e seringueira, que divergiram da mandioca há aproximadamente 35 milhões de anos (BREDESON et al., 2016).

Atualmente, sabe-se que a mandioca (*Manihot esculenta* Crantz), popularmente conhecida como aipim, macaxeira, mandioca e yuca, é a única espécie comercial do gênero *Manihot*, domesticada no sudoeste da Bacia Amazônica a partir do progenitor selvagem *M. esculenta ssp. flabillifolia* e difundida para além da América do Sul nos últimos 500 anos, exportada pelos colonizadores europeus e pelos comerciantes de escravos (NASSAR, 2002; OLSEN, 2004).

É uma das culturas tropicais mais difundidas no país, resultado da ampla variabilidade genética cultivada em todo território brasileiro, sendo uma das culturas fundamentais na segurança alimentar, por ser excelente fonte de carboidratos e apresenta plasticidade fenotípica quando submetidas a diferentes condições edafoclimáticas e pela diversidade de usos na alimentação humana e animal (LEBOT, 2009; RABBI et al., 2012).

A manutenção da diversidade genética tem sido um grande desafio considerando a substituição de variedades locais por variedades melhoradas (WILLEMEN et al., 2007). Em razão disto, é notória a preocupação mundial com a redução da variabilidade genética, que pode comprometer a utilização e conservação dos recursos genéticos de diversas espécies, principalmente na obtenção de progressos genéticos constantes, no âmbito de ação dos programas de melhoramento (OLIVEIRA et al., 2015).

A maior parte da variabilidade genética da mandioca tem sido preservada em Bancos Ativos de Germoplasma (BAG) sob condições de campo, tendo como objetivo a conservação da variabilidade genética existente na natureza e assim, evitar a erosão genética causada pela ação antrópica (FERGUSON et al., 2012).

Atualmente, os principais bancos ativos de germoplasma de mandioca estão localizados no: i) Centro Internacional de Agricultura Tropical (CIAT) - Colômbia, com aproximadamente 6.000 acessos, sendo 93% compostos por variedades locais coletadas de regiões tropicais e subtropicais e outros 7% por híbridos (OKOGBENIN et al., 2007); ii) Instituto Internacional de Agricultura Tropical (IITA) - Nigéria, com aproximadamente 4.000 acessos dentre parentes silvestres a variedades melhoradas, mantidos em campo (*ex situ*) e banco *in vitro* (DUMET et al., 2011); e iii) Empresa Brasileira de Pesquisa Agropecuária (Embrapa) – Brasil, com cerca de 4.000 acessos, onde aproximadamente 80% são variedades locais e 20% compostas por variedades ou híbridos melhorados, oriundos de diversas regiões do país e do exterior.

Para explorar toda base genética conservada nos BAGs, é que estudos sobre diversidade genética e avaliação de estrutura populacional são fundamentais, por serem precursores em abordagens convencionais e biotecnológicas aplicadas ao melhoramento vegetal, a exemplo de estudos de mapeamento associativo (ATWELL et al., 2010; TIAN et al., 2011), evolução (VAN-HEERWAARDEN et al., 2011) e seleção genômica (AZEVEDO et al., 2016; WOLFE et al., 2016). Esses estudos também são úteis na seleção de combinações parentais favoráveis para o desenvolvimento de progênies, com variabilidade genética máxima e na estimação de índices de diversidade que dão suporte na compreensão de aspectos evolutivos (mutação, seleção natural, migração ou fluxo gênico e deriva genética) (SEMAGN et al., 2012; RAMU et al., 2017), contribuindo para o melhoramento genético da espécie.

Em mandioca, a caracterização do germoplasma tem sido utilizada para identificar a variabilidade genética existente, com uso de marcadores bioquímicos (LEFÈVRE; CHARRIER, 1993), descritores quantitativos e qualitativos (KAWUKI et al., 2011a), marcadores moleculares, a exemplo dos microssatélites (GONÇALVES et al., 2017) e, mais recentemente, os SNPs (MTUNGUJA et al., 2017). Embora o entendimento sobre a diversidade genética e estrutura populacional por meio de métodos convencionais, utilizando a caracterização morfoagronômica seja uma estratégia bastante utilizada na cultura da mandioca (KAWUKI et al., 2011a), existem algumas limitações inerentes a este tipo de caracterização, sobretudo descritores quantitativos, devido à forte influência ambiental, típica de características de

herança complexa (VIGOUROUX et al., 2008; AGRAMA et al., 2010; MORRIS et al., 2013). Por outro lado, os descritores qualitativos têm como principal desvantagem seu limitado número para ampla discriminação fenotípica. Portanto, esses descritores certamente continuarão a ser bastante úteis na caracterização do germoplasma, mas como informações complementares a outros métodos contemporâneos para detecção de variabilidade genética.

Devido à necessidade de se aumentar a eficiência, a precisão das análises, bem como a redução de tempo e recursos financeiros envolvidos na caracterização do germoplasma, novas estratégias têm sido desenvolvidas para estimar de forma ampla e padronizada, toda a diversidade genética conservada em BAGs, com a mínima interferência ambiental, a exemplo dos marcadores de DNA (RABBI et al., 2015; AZEVEDO et al., 2016; GONÇALVES et al., 2017; MELO et al., 2017). Dentre os marcadores moleculares, os *Single-Nucleotide Polymorphism* (SNP) tem atraído atenção da pesquisa em virtude da abundância no genoma, localização específica no cromossomo, baixa taxa de mutação e a facilidade de automação (MAMMADOV et al., 2012).

Uma das principais vantagens desses marcadores, é a possibilidade de automação quando associada à tecnologia *Next-Generation Sequencing* (NGS). Dentre as plataformas de sequenciamento da próxima geração disponíveis, o método *genotyping-by-sequencing* (GBS), relativamente simples e de menor custo, torna viável a genotipagem de grandes populações de indivíduos (ELSHIRE et al., 2011). Deste modo, tornou-se prestigiado, particularmente para pesquisas que envolvem espécies não-modelo e com recursos genômicos limitados (SEEB et al., 2011; GLAUBITZ et al., 2014).

Os SNPs têm sido amplamente utilizados em estudos de evolução, conservação, diversidade genética e estrutura populacional em diferentes espécies, como girassol (FILIPPI et al., 2015); soja (SONG et al., 2015); trigo (SPANIC et al., 2016); arroz (TANG et al., 2016); e em organismos não-modelo (SEEB et al., 2011). A aplicação desses marcadores na cultura da mandioca, tem acompanhado a evolução das outras culturas, tendo seu uso focado na seleção genômica (OLIVEIRA et al., 2012); construção de mapas de ligação; mapeamento de *quantitative trait locus* (QTL) (HAMBLIN; RABBI 2014; POOTAKHAM et al., 2014); desenvolvimento de coleção nuclear (OLIVEIRA et al., 2014b); identificação de fontes de resistência, por meio da seleção

assistida por marcadores (exemplo do vírus do mosaico africano) (CARMO et al., 2015); identificação de acessos duplicados (RABBI et al., 2015) e caracterização de germoplasma (ORTIZ et al., 2016; MTUNGUJA et al., 2017).

Recentemente, 402 SNPs foram utilizados em estudos de diversidade genética e estrutura populacional (OLIVEIRA et al., 2014a) no germoplasma de mandioca da América Latina. Contudo, um número reduzido de SNPs, ligados a regiões específicas do genoma da espécie, e apenas parte do germoplasma de mandioca disponível, foi utilizada. Assim, ainda é preciso investir na análise da diversidade e estrutura populacional do maior número de acessos de *M. esculenta* com uso de marcadores com ampla cobertura genômica. Desse modo, o objetivo deste estudo foi analisar a estrutura genética de acessos pertencentes ao Banco Ativo de Germoplasma de mandioca da Embrapa Mandioca e Fruticultura, com base em marcadores SNPs, a fim de quantificar a variabilidade genética existente, contribuindo assim na elaboração de estratégias eficientes para a conservação e melhoramento genético da espécie.

## MATERIAL E MÉTODOS

### *Material Vegetal*

Foram utilizados 1.580 acessos do Banco Ativo de Germoplasma de Mandioca (BAG-Mandioca), pertencente à Embrapa Mandioca e Fruticultura, localizada em Cruz das Almas – BA, Brasil (12°40'19"S, 39°06'22"W, 226 m altitude). A região é de clima tropical com temperatura média anual de 24,5°C, 80% de umidade relativa do ar, e 1.250 mm de precipitação anual. Os acessos são oriundos de diferentes ecossistemas do Brasil, Colômbia, Nigéria, Panamá, Venezuela e Uganda, dos quais 1.183 são variedades locais e 397 são variedades melhoradas, obtidas a partir de métodos convencionais de melhoramento, seleção massal e cruzamentos artificiais, bem como seleção de variedades locais com alto potencial produtivo identificado por produtores e/ou instituições de pesquisa.

### *Extração de DNA*

O DNA foi extraído a partir de folhas jovens, de acordo com o protocolo CTAB (brometo de cetiltrimetilamônio) conforme descrito por Doyle; Doyle (1987), com pequenas modificações, a exemplo da adição de polivilpirrolidona (PVP) e aumento da concentração de 2-mercaptoetanol a 0,4%. A qualidade do

DNA foi avaliada por quantificação em gel de agarose 1,0% (p/v) corado com brometo de etídio (1,0mg/L) em tampão TBE 0,5 x (45 mM Tris-borate, 1 mM EDTA e q.s.p de água destilada), visualizado em luz UV e registrado com o fotodocumentador Gel Logic 212 Pro (Carestream Molecular Imaging, New Haven, USA) por comparação visual com uma série de concentrações de DNA conhecido do fago Lambda (Invitrogen, Carlsbad, CA). O DNA foi diluído em tampão TE (Tris-HCl 10mM e EDTA 1mM) para uma concentração final de 60 ng/µL e a qualidade foi verificada pela digestão de 250 ng do DNA genômico a partir de 10 amostras aleatórias com a enzima de restrição *Eco*RI (New England Biolabs, Boston, EUA) a 65° C durante duas horas e posteriormente, visualizada em gel de agarose.

***Genotipagem por Sequenciamento***

As amostras de DNA foram genotipadas no *Genomic Diversity Facility* pertencente à *Cornell University* (http://www.biotech.cornell.edu/brc/genomic-diversity-facility). O protocolo básico da *genotyping-by-sequencing* (GBS), foi descrito por Elshire et al. (2011), no qual o DNA foi digerido pela enzima *Ape*KI recomendada por (HAMBLIN; RABBI 2014), uma endonuclease de restrição tipo II que reconhece uma sequência degenerada de 5 bases (GCWGC, onde W é A ou T) com comprimentos de 100 pb. A ligação entre os fragmentos com corte *Ape*KI e o adaptador, foi realizada após a digestão das amostras e implementação de sistema multiplex com 192 amostras para realização do sequenciamento. A GBS foi realizada utilizando o *Genome Analyzer 2000* (Illumina, Inc., San Diego, CA). Para análise das sequências e filtros de qualidade foram utilizados os softwares Tassel versão 5.2.37 (BRADBURY et al., 2007) visando remover alelos com uma frequência mínima (MAF) inferior a 5% e SNPs com mais de 20% de dados perdidos. Posteriormente, os demais dados perdidos foram imputados no software R versão 3.3.4 (R Development Core Team, 2017), utilizando o pacote computacional *missForest*.

***Análise de diversidade***

Com base nas estimativas das frequências alélicas dos SNPs, foram obtidas as seguintes informações: 1) Heterozigosidade observada (Ho), calculada a partir das frequências genotípicas; 2) Heterozigosidade esperada (He), calculada de acordo com Nei (1973): $\mathrm{He} = 1 - \sum p_{ij}^2$, onde $p_{ij}$ é a frequência do $j^{th}$ alelo para o $i^{th}$ loco; 3) Conteúdo de informação polimórfica

(PIC), de acordo com Botstein et al. (1980): $\mathrm{PIC} = \sum_{j=i+1}^{k} 2p_i^2 p_j^2$, em que *k* é o número de alelos e $p_i$ e $p_j$ são as frequências dos alelos *i* e *j*; 4) Coeficiente de endogamia por loco, $f = \frac{\delta_{ij} - p_i}{1 - p_i}$, em que $\delta_{ij}$=1 se o loco for homozigótico (alelo *i* = alelo *j*) e $p_i$ é a frequência do alelo *i* na população. 5) Índice de endogamia $f$ por indivíduo, $f = 1 - \left(\frac{Ho}{He}\right)$; 6) Equilíbrio de Hardy-Weinberg (HWE *p*-values) para cada loco (p<0,01), pelo teste clássico $\chi^2$; e 7) O desequilíbrio de ligação (LD) estimado pelo coeficiente de correlação ($r^2$) foi dado por $r^2 = \frac{(pab - papb)^2}{pa\,(1-pa)pb(1-pb)}$., em que $pab$ é a frequência de gametas com o par de alelos $a$ e $b$, e $papb$ é o produto das frequências dos dois alelos $pa$ e $pb$. O quadrado do coeficiente de correlação $r^2(pa, pb, pab)$ pode variar de 0 a 1, assim como $pa, pb, pab$.

Todas as estimativas foram obtidas com auxílio do software R versão 3.3.4 (R Development Core Team, 2017), utilizando os pacotes computacionais *adegenet*, *hierfstat*, *ggplot2*, *genetics*, *pegas* e *LDheatmap*.

***Análise da estrutura populacional***

Duas análises complementares foram utilizadas para avaliar a estrutura e diversidade genética entre os acessos: agrupamento com base na análise discriminante de componentes principais (ADCP) e análise de variância molecular (AMOVA).

A ADCP baseia-se na transformação preliminar dos dados, utilizando a análise de componentes principais (ACP) como passo prévio à análise discriminante (AD), garantindo que as variáveis submetidas à AD são perfeitamente descorrelacionadas e o seu número seja menor que os indivíduos analisados sem necessariamente implicar na perda de informação genética. Esta transformação permite a AD aplicar-se a qualquer dado genético, minimizando as diferenças entre indivíduos dentro de cada grupo e maximizando-as entre grupos, no qual os acessos são melhores discriminados em grupos pré-definidos.

A ADCP foi analisada no pacote *adegenet* do software R versão 3.3.4. (R Development Core Team, 2017) sem informação prévia sobre os indivíduos para definição de grupos (JOMBART et al., 2010), onde o número de grupos foi definido com base no agrupamento *K-means*, considerando o argumento

*smoothNgoesup* da função *find.cluster* que seleciona o *K* mais adequado com base no maior aumento do valor do critério de informação bayesiano (BIC). O argumento *smoothNgoesup* utiliza um método não parâmetro de regressão chamado lowess (*local-weighted regression and smootthing scatterplot*) que adapta uma curva de regressão linear localmente ponderada para promover a suavização de gráficos de dispersão. Após a definição do número de grupos, os eixos da análise de componentes principais que explicaram mais de 85% da variância total dos dados, foram mantidos na análise (JOMBART et al., 2010).

A análise de variância molecular (AMOVA) foi realizada apenas com os SNPs mais informativos, tendo como critério de seleção um PIC > 0,30 (total de 7.436 SNPs). A AMOVA foi realizada considerando diferentes níveis hierárquicos: a) origem geográfica dos acessos Brasil (regiões Centro-Oeste, Norte, Nordeste e Sul), Colômbia (região do Valle), Nigéria, Panamá, Venezuela e Uganda; b) origem dos acessos (variedade melhorada ou local); c) teor de compostos cianogênicos (classificação em mandioca mansa, intermediária ou brava); d) agrupamentos formados pela ADCP e; e) cor de polpa da raiz (branca, creme, amarela e rosada). A AMOVA foi realizada utilizando o pacote *poppr* implementado no software R versão 3.3.4. (R Development Core Team, 2017).

## RESULTADOS

### *Análise dos marcadores SNPs*

O método GBS, utilizado para identificar os marcadores SNPs, apresentou uma cobertura de 71,69% no genoma da mandioca. Inicialmente, as sequências permitiram a identificação de 444.821 SNPs mapeados no genoma de referência da mandioca versão 6 (http://phytozome.jgi.doe.gov). Foram aplicados filtros de qualidade visando à remoção de alelos com baixa frequência (MAF<0,05) e locos com dados perdidos que não atendiam aos pré-requisitos (>20%).

Assim, foram obtidos SNPs de alta qualidade, totalizando 20.610, que representam 4,63% do número original. Dentre os 20.610 SNPs de alta qualidade, 19.607 (95,13%) foram fisicamente mapeados nos 18 cromossomos de *M. esculenta*, com uma densidade média de 26,85 kb (Tabela 1). As maiores e menores densidades de marcadores foram observadas nos

cromossomos 14 (21,11 kb) e 17 (34,66 kb), respectivamente. Os demais 1.003 SNPs (4,87%) não foram fisicamente alocados nos cromossomos da mandioca, por isso, foram alocados em *scaffolds* denominados A e B (Tabela 1).

**Tabela 1**. Distribuição e densidade dos 20.610 marcadores *Single-Nucleotide Polymorphism* (SNP) identificados em *Manihot esculenta* Crantz.

| Cromossomos | Tamanho do cromossomo (Mb) | Número de SNP | Porcentagem de SNP (%) | Densidade média (Kb) |
|---|---|---|---|---|
| 1 | 34,58 | 1577 | 7,65 | 21,93 |
| 2 | 32,55 | 1288 | 6,25 | 25,27 |
| 3 | 29,44 | 1347 | 6,54 | 21,86 |
| 4 | 28,78 | 892 | 4,33 | 32,26 |
| 5 | 28,45 | 1230 | 5,97 | 23,13 |
| 6 | 27,96 | 1234 | 5,99 | 22,66 |
| 7 | 27,09 | 951 | 4,61 | 28,49 |
| 8 | 34,03 | 1019 | 4,94 | 33,40 |
| 9 | 29,44 | 1030 | 5,00 | 28,58 |
| 10 | 26,36 | 1039 | 5,04 | 25,37 |
| 11 | 27,40 | 1059 | 5,14 | 25,87 |
| 12 | 28,47 | 963 | 4,67 | 29,56 |
| 13 | 28,14 | 915 | 4,44 | 30,75 |
| 14 | 24,87 | 1178 | 5,72 | 21,11 |
| 15 | 26,21 | 1124 | 5,45 | 23,32 |
| 16 | 28,99 | 976 | 4,74 | 29,70 |
| 17 | 27,42 | 791 | 3,84 | 34,66 |
| 18 | 25,26 | 994 | 4,82 | 25,41 |
| *Scaffold* A | - | 420 | 2,04 | - |
| *Scaffold* B | - | 583 | 2,83 | - |
| **Média** | | 1.030,50 | | 26,85 |

Nove SNPs mostraram-se monomórficos nos acessos de mandioca avaliados e, portanto, foram removidos, restando 20.601 SNPs para uso nas análises subsequentes de diversidade genética e estrutura populacional. Em média, foram identificados 1.030 SNPs por cromossomo, apresentando variação de 420 (*Scaffold* A) a 1.577 SNPs (cromossomo 1) (Tabela 1). Deste modo, foram identificados SNPs em todos os cromossomos, de maneira relativamente equilibrada, o que certamente pode trazer maior precisão e confiabilidade nas estimativas de divergência genética e estrutura populacional.

A relação entre sítios de transição e transversão foi de 1,26; ou seja, foram identificados 11.486 SNPs (55,73%) com transição e 9.124 SNPs

(44,27%) em transversão (Tabela 2). Não havendo equilíbrio entre as transições (A/G = 31,38% e C/T = 24,35%) e transversões (variação de 8,05% para C/G a 14,11% para A/T) nos SNPs analisados. As transições A/G e transversões C/G ocorreram em maior e menor frequência, respectivamente. Em relação aos polimorfismos dos SNPs, os tipos mais frequentes foram A (29,16%), G (24,31%), T (23,94%) e C (22,60%).

**Tabela 2**. Transições e transversões identificadas em marcadores *Single-Nucleotide Polymorphism* (SNP) obtidos por genotipagem por sequenciamento (GBS).

| | **Transições** | | **Transversões** | | | |
|---|---|---|---|---|---|---|
| | **A/G** | **C/T** | **A/C** | **A/T** | **C/G** | **G/T** |
| Número de SNPs | 6.467 | 5.019 | 2.646 | 2.909 | 1.659 | 1.910 |
| Percentagem SNPs | 31,38% | 24,35% | 12,84% | 14,11% | 8,05% | 9,27% |
| Total | 11.486 (55,73%) | | 9.124 (44,27%) | | | |

***Diversidade genética***

Em relação ao número de alelos, a maioria dos SNPs (97,38%) foram caracterizados como bialélicos. Entretanto, aproximadamente 2,51% e 0,11% dos SNPs foram caracterizados como tri e tetra-alélicos, respectivamente (Tabela 3). Portanto, considerando esses locos mutialélicos, foram identificados 41.758 alelos (média de 2,03 alelos por loco). As frequências alélicas foram bastante diferentes para os SNPs com diferentes números de alelos (Tabela 3), pois nos locos tri e tetra-alélicos, 69,52% e 80,44% das respectivas frequências alélicas, estavam presentes em baixa frequência (<0,50), enquanto nos SNPs bialélicos as frequências alélicas foram de aproximadamente 50% para cada alelo, conforme esperado para este tipo de marcador (Tabela 3). Dentre os alelos presentes no germoplasma com frequência elevada (>0,75), os SNPs bialélicos foram os mais abundantes (33,01%), seguidos pelos SNPs trialélicos (16,34%) e tetra-alélicos (8,70%).

Considerando todos os SNPs, a média da heterozigosidade observada (Ho) foi de 0,235 tendo variação de 0,01 a 1,00. Em 96% dos casos, as estimativas de Ho foram inferiores a 0,50. Por outro lado, apenas 3,71% dos SNPs distribuídos nos 18 cromossomos e nos s*caffolds* A e B foram identificados com Ho acima de 0,50. Entretanto, a densidade média para este

parâmetro mostrou que a maioria dos SNPs (76,23%) concentrou-se em um intervalo de 0,01 a 0,30 (Figura 1).

**Tabela 3**. Diversidade genética no germoplasma de mandioca, em função do número de alelos dos marcadores *Single-Nucleotide Polymorphism* (SNP).

| Nº de alelos | Nº SNPs | Valores médios | | | Média das frequências alélicas | | | |
|---|---|---|---|---|---|---|---|---|
| | | Ho | He | PIC | 0,0-0,25 | 0,25-0,50 | 0,51-0,75 | 0,75-1,00 |
| 2 | 20.071 | 0,23 | 0,29 | 0,24 | 33,04 | 17,03 | 16,92 | 33,01 |
| 3 | 516 | 0,36 | 0,38 | 0,33 | 53,57 | 15,95 | 14,14 | 16,34 |
| 4 | 23 | 0,45 | 0,43 | 0,37 | 61,96 | 18,48 | 10,87 | 8,70 |

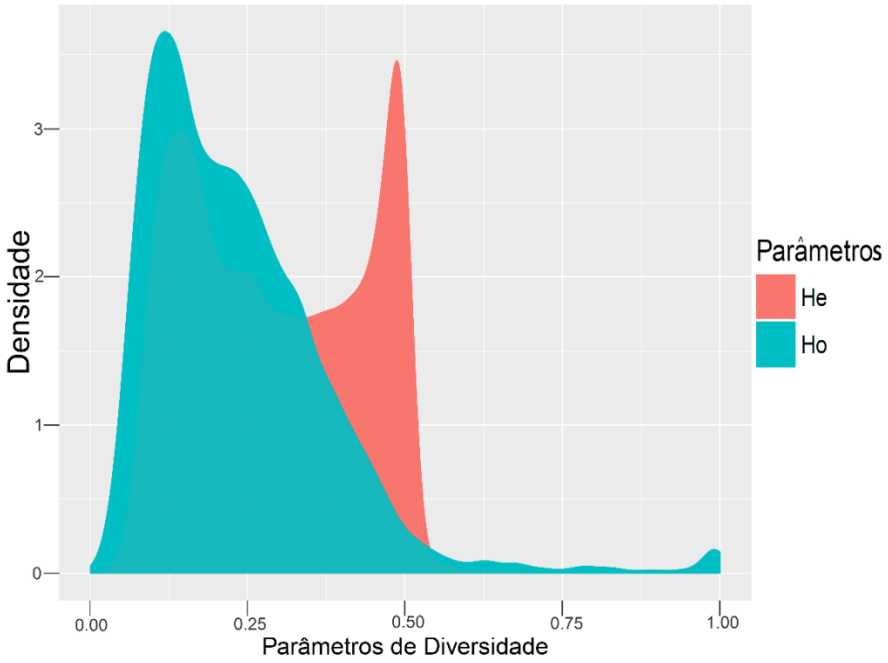


**Figura 1**. Distribuição dos valores de heterozigosidade esperada (He) e observada (Ho) com base na análise de 20.601 *Single-Nucleotide Polymorphism* (SNP).

Em relação à heterozigosidade esperada ou diversidade genética (He), foi observada variação entre 0,04 e 0,67, com média de 0,30. Entretanto, a maioria dos SNPs (99%) apresentaram He variando de 0,03 a 0,50 (Figura 1). Aproximadamente 70% dos SNPs apresentaram valores de Ho inferior a He, sugerindo um déficit de heterozigotos em relação ao esperado sob o equilíbrio de Hardy-Weinberg no caso de populações naturais. O nível de polimorfismo

avaliado pelo conteúdo de informação polimórfica (PIC), variou de 0,04 a 0,61, com média de 0,24 por SNP. Aproximadamente 49% dos SNPs apresentaram PIC superior a 0,25, sendo considerados como os marcadores mais informativos para os mais diversos estudos genéticos sobre variabilidade, associação com fenótipos e seleção assistida (Figura 2).

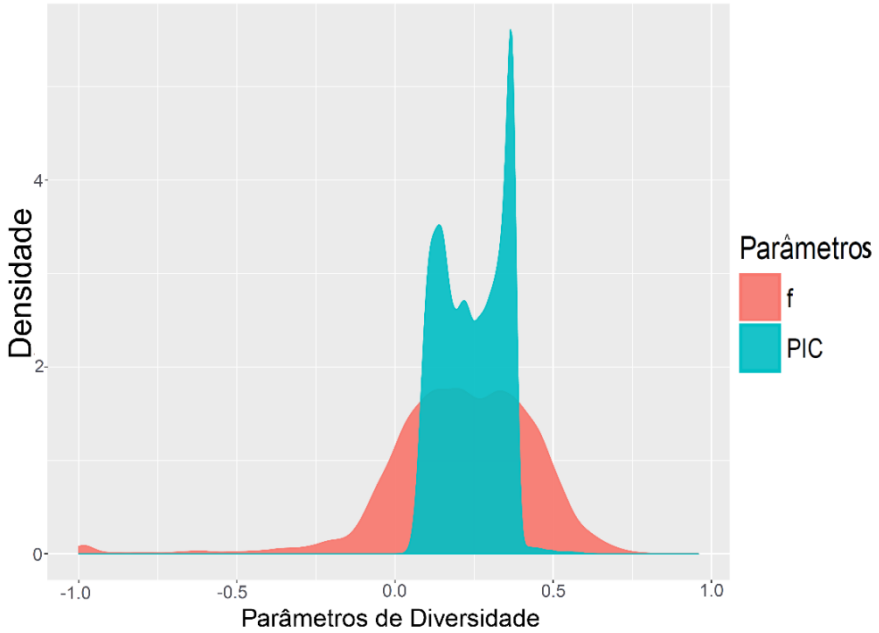


**Figura 2**. Distribuição dos valores de Conteúdo de Informação Polimórfica (PIC) e endogamia (f) com base na análise de 20.601 *Single-Nucleotide Polymorphism* (SNP).

Considerando os SNPs marcadores predominantemente bialélicos, o valor máximo de He seria de 0,50, porém a presença de SNPs tri e tetra-alélicos, certamente, contribuíram para que os SNPs excedessem este valor. Em termos médios, todas estas estimativas de Ho, He e PIC são maiores quando são identificados mais de dois alelos por loco. Por exemplo, o PIC médio dos marcadores bi, tri e tetra-alélicos foi de 0,24, 0,33 e 0,37, respectivamente (Tabela 3).

Em relação à endogamia em nível de marcadores, observou-se variação do *f* entre -1,00 a 0,96, com média de 0,21 (Figura 2). Foram identificados 2.632 locos (12,77%) com excesso de heterozigotos (-1,00 a -0,01), muito embora a maioria dos locos tenha apresentado níveis de homozigose (87,23%). A distribuição dos marcadores com endogamia negativa foi bastante

homogênea ao longo dos cromossomos da mandioca (Figura 3), indicando ausência de associação a regiões genômicas específicas, sujeitas a pressão de seleção ao longo do processo evolutivo.

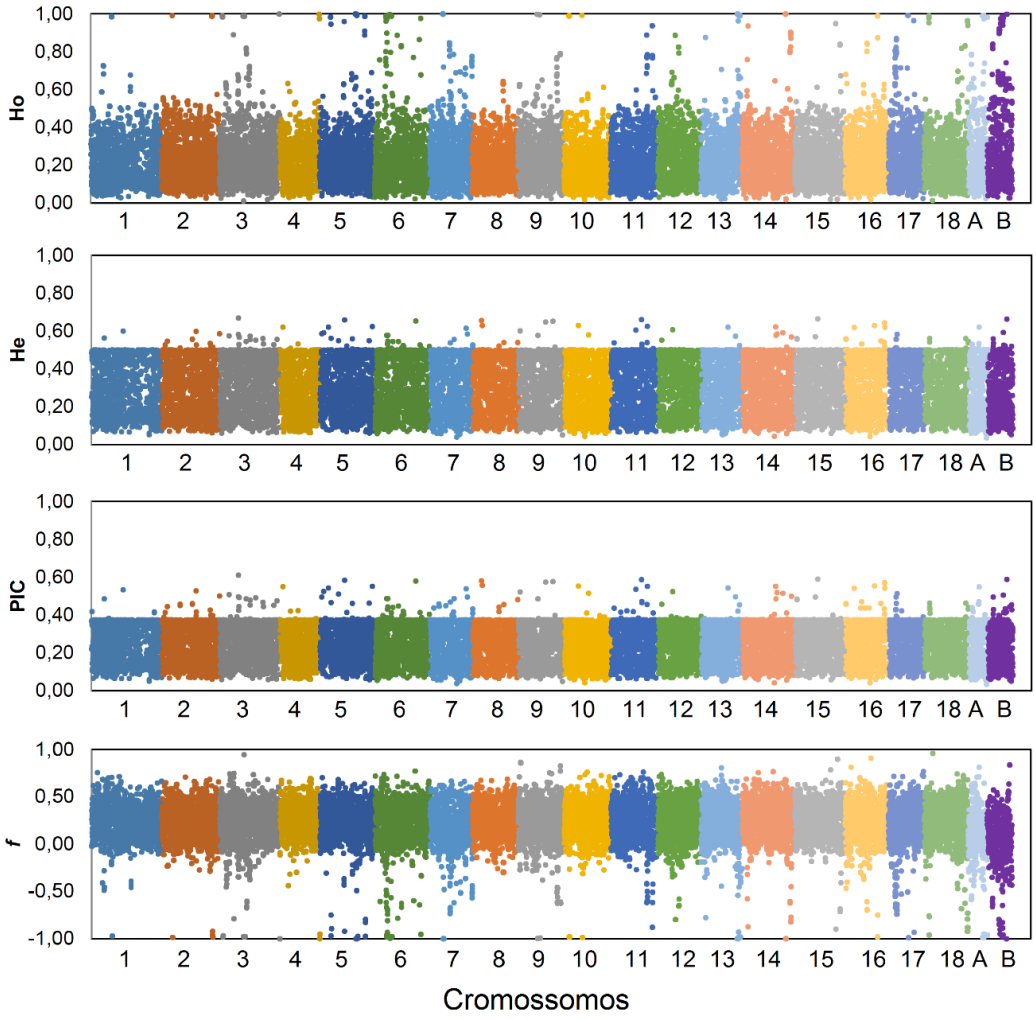


**Figura 3**. Distribuição das estimativas de heterozigosidade observada (Ho), heterozigosidade esperada (He), Conteúdo de Informação Polimórfica (PIC) e endogamia (*f*) nos diferentes cromossomos da mandioca, com base na análise de 20.601 *Single-Nucleotide Polymorphism* (SNP).

Em nível cromossômico, os valores médios de He, Ho, PIC e *f* foram muito similares nos 18 cromossomos de *M. esculenta*, à exceção dos *scaffolds* A e B que apresentaram valores mais baixos de endogamia (Figura 3). A variação nos valores de Ho foi de 0,21 (cromossomo 1) a 0,27 (*scaffold* B); o He variou de 0,26 (*scaffolds* A e B) a 0,31 (cromossomos 3, 9, 11, 12, 13, 14 e 17); o PIC variou de 0,21 (*scaffold* A) a 0,26 (cromossomo 14); enquanto os valores de endogamia variaram de 0,03 (*scaffold* B) a 0,26 (cromossomo 4).

Em nível de indivíduo, os valores de Ho variaram de 0,02 a 0,38, com média de 0,24, sendo que 90% dos indivíduos apresentaram Ho entre 0,18 e 0,38 (Figura 4). Em relação à endogamia observada nos acessos de mandioca, a variação foi de 0,49 a 0,97, com média de 0,69, onde 90% dos acessos de mandioca apresentaram endogamia variando de 0,49 a 0,75. Três acessos (BGM0104; BGM1226 e BGM1448) apresentaram endogamia acima de 0,90. A maioria dos 1.580 acessos de mandioca apresentaram baixos níveis de Ho e elevados níveis de endogamia.

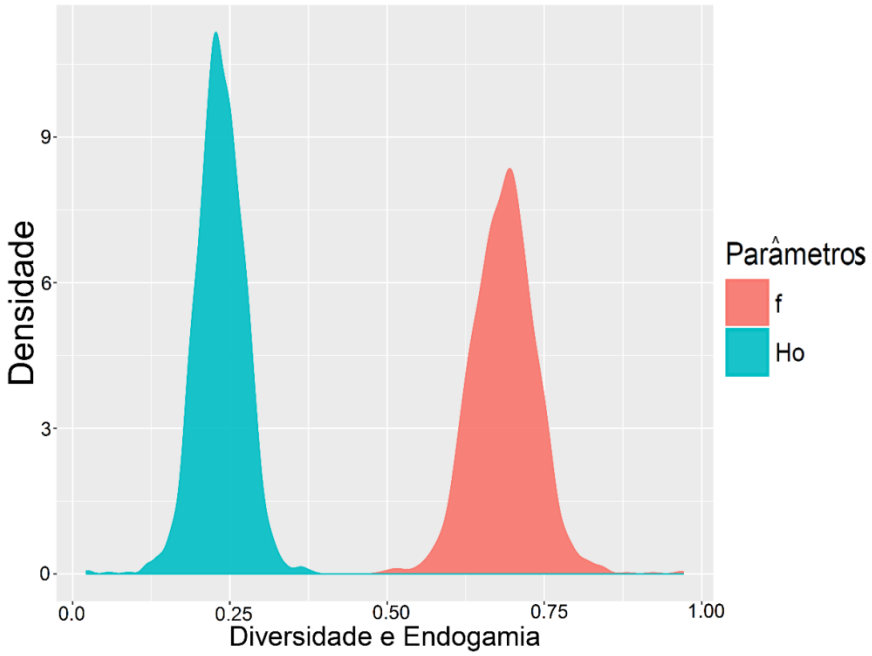


**Figura 4**. Distribuição dos valores de endogamia (*f*) e heterozigosidade observada (Ho) em 1580 acessos de mandioca, analisados com base em marcadores *Single-Nucleotide Polymorphism* (SNP).

Desvios do equilíbrio de Hardy-Weinberg (EHW) ($p<0,01$) foram identificados em 17.328 SNPs (84,11%), como resultado da menor heterozigosidade observada em relação à esperada e pela elevada endogamia dos acessos. SNPs com desvios do EHW foram observados em todos os cromossomos e nos dois *scaffolds,* indicando que este é um atributo intrínseco da espécie, e que sinais de seleção direcionada para características/regiões genômicas, parece não contribuir para os desvios. Portanto, os desvios das frequências alélicas esperadas sob equilíbrio, é uma das razões pelas quais os acessos de mandioca apresentam a maioria dos SNPs em desequilíbrio de

Hardy-Weinberg. Uma importante explicação para os desvios do EHW deve-se ao fato da mandioca não constituir uma população verdadeira do ponto de vista de cruzamentos ao acaso e manutenção de reprodução assexuada como forma de multiplicação da espécie. Na verdade, o principal tipo de propagação da mandioca é por via assexuada (clonal), fazendo com que apenas uma amostra dos indivíduos da população original seja amostrada, contribuindo para os desvios nas frequências alélicas observadas e esperadas. Outros fatores também podem estar associados a possíveis desvios do EHW são, erros de genotipagem e estratificação populacional (ZINTZARAS, 2010).

Os valores de desequilíbrio de ligação (LD) foram calculados para cada par de SNPs. A média geral $r^2$ para todas as comparações foi de 0,014 com variação de 0,00 a 1,00. Considerando apenas pares de SNPs com LD significativo, 0,93% apresentaram $r^2$ acima de 0,20; enquanto 0,30% e apenas 0,08% apresentaram $r^2$ acima de 0,50 e 0,80, respectivamente. O LD decaiu de forma mais brusca na região próxima de 8 kb, porém apresentou maior estabilidade no ajuste da regressão não linear entre 15 e 20 kb, cujos valores de $r^2$ correspondem a $r^2$= 0,2 (Figura 5).

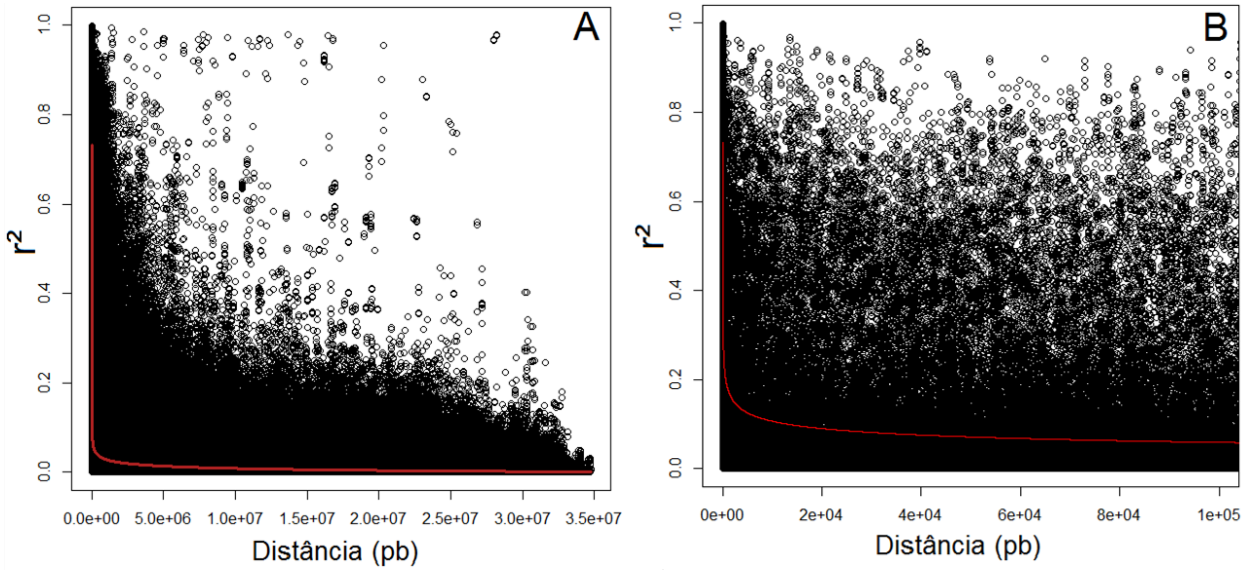


**Figura 5**. Desequilíbrio de ligação (LD – $r^2$) estimado entre os pares de *Single-Nucleotide Polymorphism* (SNP) em função da distância física em pares de base (pb) (A). Zoom da região de maior queda do $r^2$ até 100 Kb (B). A linha vermelha indica o ajuste da regressão não linear do LD ao longo do genoma em ambas as imagens.

***Estrutura Populacional***

Para obtenção do número ideal de clusters, 500 componentes principais (CP), obtidos por meio da análise de componentes principais (ACP), foram responsáveis por explicar mais de 85% da variação genética dos acessos e, portanto, mantidos na análise para transformação preliminar dos dados. De acordo com Oliveira et al. (2014a), a definição do número de CP a serem retidos na análise, é de suma importância para garantir o poder de redução na dimensionalidade dos dados, pois no contexto da ADCP é preciso definir um ponto de equilíbrio entre o poder de discriminação dos agrupamentos e a estabilidade de atribuições dos genótipos em cada grupo. Contudo, diversos trabalhos indicam o uso de componentes que retenham mais de 80% da variância genética. Assim, a análise com 500 CP garantiu alto poder estatístico para avaliar a estrutura genética do germoplasma de mandioca.

A definição do número de grupos (*K*) foi feita após a realização de 1.000 corridas independentes da função *find.cluster*, sendo que o valor *K* mais frequente para representar o número de grupos, foi 22. Assim, os 22 grupos foram plotados pela ADCP, na qual seis grupos foram claramente distintos dos demais (G10, G11, G14, G16, G18, G20) com base no *scatterplot* das funções discriminantes 1 e 2, sendo possível uma nítida separação em duas dimensões (Figura 6). Os demais grupos apresentaram sobreposição na plotagem em duas dimensões das funções discriminantes supracitadas. Por outro lado, a análise do *scatterplot* das funções discriminantes 2 e 3 possibilitou uma clara distinção dos grupos G2, G6, G7, G8, G9, G12, G13, G19 e G22, além dos grupos G10, G11, G14, G16, G18, G20 (Figura 7). Adicionalmente, as funções discriminantes 1 e 3 foram capazes de separar os grupos G1, G3, G4, G5, G15 e G17 (Figura 8). Portanto, as funções discriminantes da ADCP utilizadas para o agrupamento dos acessos de mandioca com base em marcadores SNPs foram bastante eficientes na discriminação dos acessos de mandioca nos diferentes grupos.

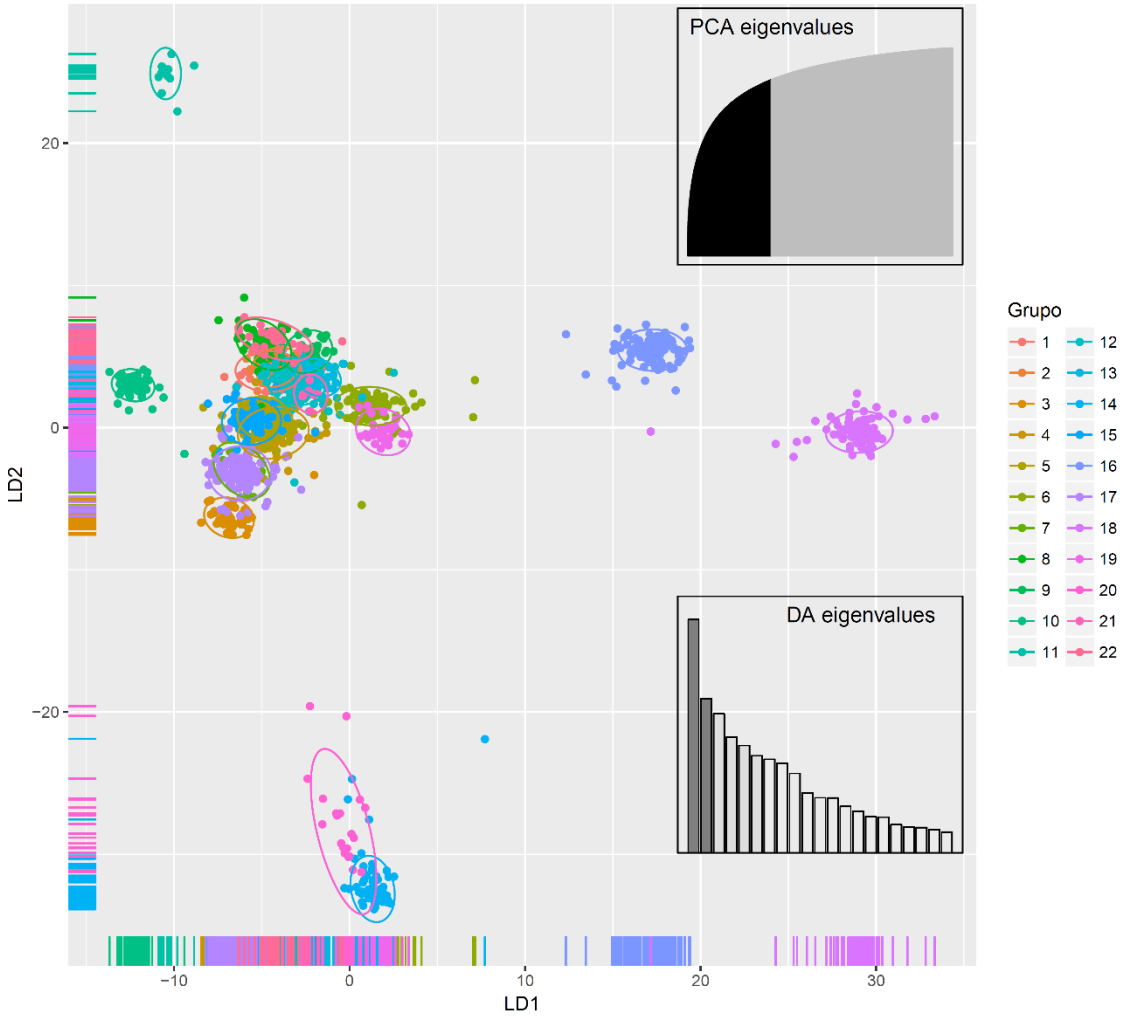


**Figura 6**. Gráfico de dispersão de 1.580 acessos de mandioca com base nas funções discriminantes 1 e 2 (LD1 e LD2, respectivamente) da análise de componentes principais, obtidas da análise de 20.601 marcadores *Single-Nucleotide Polymorphism* (SNP). Os grupos estão representados por cores de acordo com a legenda. O gráfico acima à direita representa a contribuição dos autovalores (*eigenvalues*) dos componentes principais selecionados para a ADCP, enquanto o gráfico abaixo à direita indica a variância explicada pelos autovalores das duas funções discriminantes utilizadas no *scatterplot*.

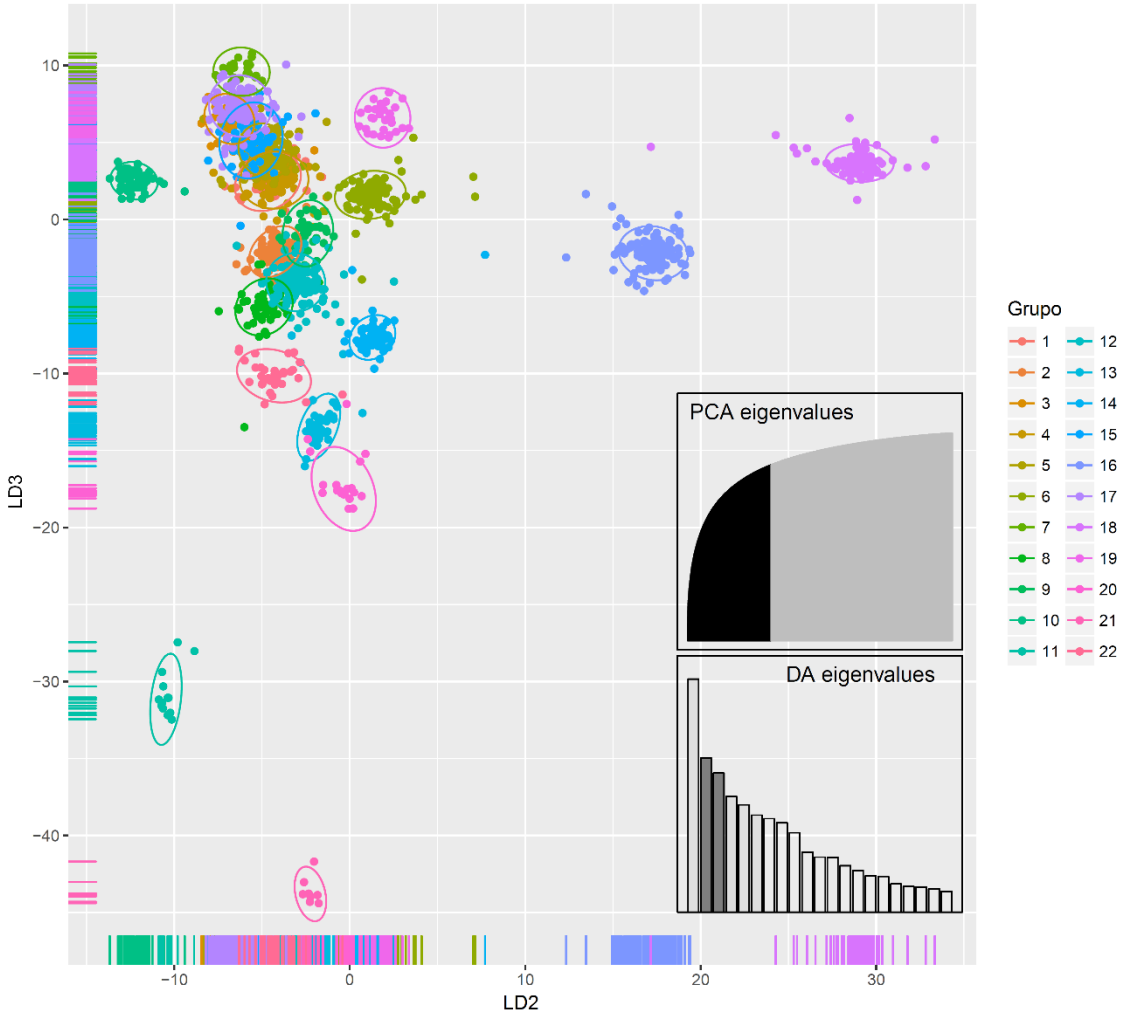


**Figura 7**. Gráfico de dispersão de 1.580 acessos de mandioca com base nas funções discriminantes 2 e 3 (LD2 e LD3, respectivamente) da análise de componentes principais, obtidas da análise de 20.601 marcadores *Single-Nucleotide Polymorphism* (SNP). Os grupos estão representados por cores de acordo com a legenda. O gráfico acima à direita representa a contribuição dos autovalores (*eigenvalues*) dos componentes principais selecionados para a ADCP, enquanto que o gráfico abaixo à direita indica a variância explicada pelos autovalores das duas funções discriminantes utilizadas no *scatterplot*.

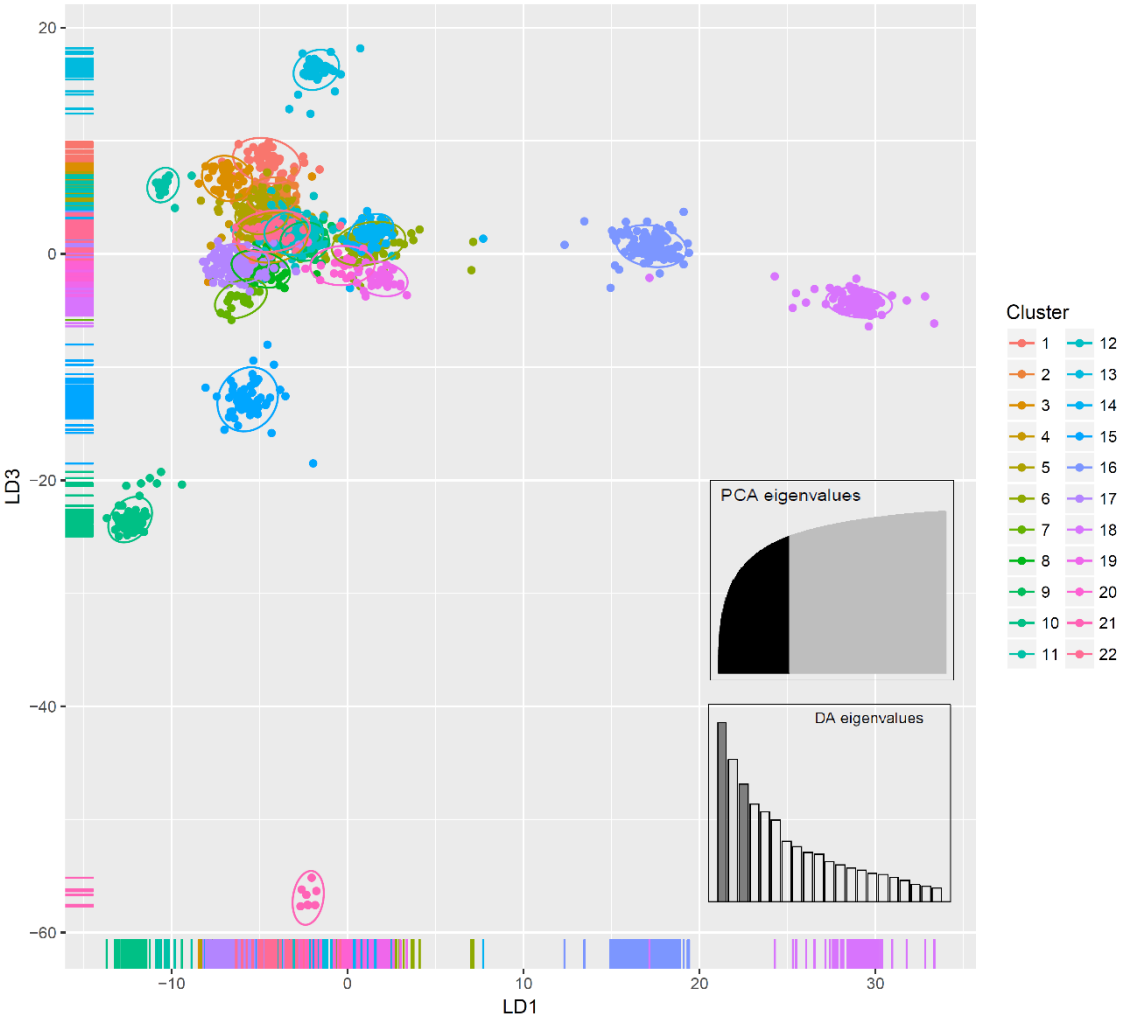


**Figura 8**. Gráfico de dispersão de 1.580 acessos de mandioca com base nas funções discriminantes 1 e 3 (LD1 e LD3, respectivamente) da análise de componentes principais, obtidas da análise de 20.601 marcadores *Single-Nucleotide Polymorphism* (SNP). Os grupos estão representados por cores de acordo com a legenda. O gráfico acima à direita representa a contribuição dos autovalores (*eigenvalues*) dos componentes principais selecionados para a ADCP, enquanto que o gráfico abaixo à direita indica a variância explicada pelos autovalores das duas funções discriminantes utilizadas no *scatterplot*.

Uma maneira de avaliar a qualidade dos agrupamentos da ADCP é através da probabilidade de alocação dos indivíduos em seus grupos originais baseando-se nas funções discriminantes, na qual altas probabilidades indicam elevado poder discriminatório dos grupos, enquanto valores baixos, sugerem grupos com alta proporção de misturas. A probabilidade de alocação dos acessos de mandioca considerando *K*=22 variou de 0,432 a 1,00. Entretanto,

para a maioria dos grupos (G1, G3, G6, G7, G10, G11, G13, G14, G16, G18, G19, G20 e G21) a probabilidade de alocação dos acessos de mandioca foi 1,00; indicando elevada acurácia na alocação dos acessos nestes grupos (Figura 9). Entretanto, os grupos G4, G5, G8, G9 e G17, apresentaram 7, 9, 1, 1 e 1 acessos, respectivamente, com probabilidade de alocação menor que 0,90; indicando a presença de acessos cuja ADCP pode ser atribuída a outros grupos (Figura 9). A porcentagem média de alocação dos acessos de mandioca a um determinado grupo foi de aproximadamente 0,99; embora alguns acessos tenham apresentado probabilidade de alocação menor que 0,70 nos grupos G4 (BGM0408 e BGM0573), G5 (BGM2081, BGM1196, BGM0208 e BGM2065) e G8 (BGM1026).

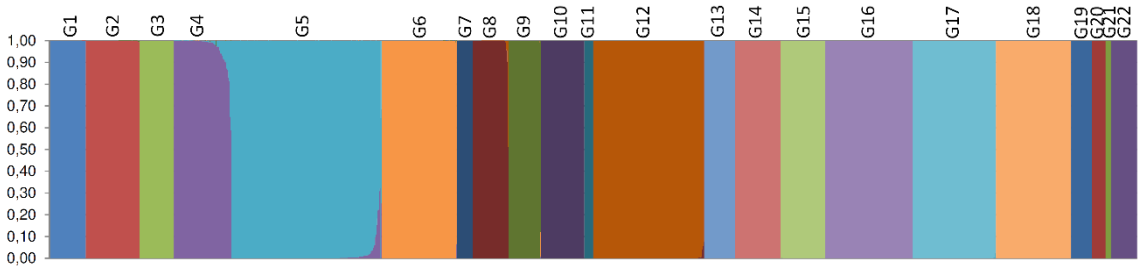


**Figura 9**. Probabilidade de alocação de acessos de mandioca com base na análise discriminante de componentes principais (ADCP) considerando *K*=22. Cada indivíduo é representado por uma linha vertical com as cores correspondendo à probabilidade de alocação em cada um dos 22 grupos.

Considerando algumas informações agronômicas importantes de raízes, como cor da polpa e teor de compostos cianogênicos, verificou-se a inexistência de associação entre os agrupamentos formados pela ADCP com estes atributos fenotípicos, pois a variação e a média dos teores de compostos cianogênicos, foi praticamente a mesma nos 22 grupos, não sendo possível identificar nenhum grupo caracteristicamente formado por acessos considerados mansos, bravos ou intermediários. Em relação à cor das raízes, na maioria dos grupos da ADCP foram alocados acessos com raízes brancas, cremes e amarelas, também não indicando nenhuma associação específica de grupo com a cor das raízes da mandioca. Exceção à esta tendência foi observada apenas nos grupos G6, G11, G19 e G22, que foram formados apenas por acessos com raízes de cor branca.

Os dados de origem geográfica dos acessos, sobretudo daqueles coletados no Brasil, também não apresentaram nenhuma associação específica com o agrupamento da ADCP, uma vez que acessos pertencentes a diferentes áreas de coleta (estados e regiões) foram distribuídos de forma bastante equilibrada entre os agrupamentos. A dificuldade de agrupamento de acessos de mandioca com base na sua origem geográfica, ocorre devido ao grande fluxo de material genético entre agricultores, sobretudo em momentos de catástrofes ambientais, a exemplo de seca ou inundações prolongadas, que levam à perda de material vegetal, e com isso, obrigam os agricultores a procurarem material propagativo em outras áreas. Quando coletadas as informações de passaporte destes acessos ficam, portanto, viesadas em relação ao verdadeiro local de origem. Por outro lado, a demonstração da baixa associação entre os grupos da ADCP com a origem geográfica poderia ser um argumento para questionar a confiabilidade da técnica. Porém, a ADCP, foi bastante confiável em alocar acessos com background genético similar em um único agrupamento (G10), a exemplo dos acessos BGM0298, BGM0303, BGM0310, BGM0312, BGM0314, BGM0316, BGM0323, BGM0327, BGM0338, BGM0360, BGM0425, BGM0601, e BGM0882, oriundos das mesmas populações segregantes, de origem colombiana.

De forma análoga à origem geográfica dos acessos de mandioca, a origem genética, também revelou pouca associação com os agrupamentos da ADCP, pois variedades locais e melhoradas, foram encontradas em todos os 22 grupos da ADCP. A classificação dos acessos de mandioca com base na origem genética, também é complexa, considerando que muitas variedades locais foram lançadas no Brasil como variedades melhoradas por apresentaram características agronômicas e de qualidade de raízes muito superiores às variedades utilizadas em determinadas regiões de cultivo. Portanto, para estes acessos, é provável que a definição da real origem genética, seja pouco acurada.

***Análise de variância molecular (AMOVA)***

A AMOVA dos marcadores SNPs com maior PIC (>0,30) em todos os acessos de mandioca foi realizada para analisar a distribuição da diversidade genética entre e dentro os grupos formados com diferentes critérios. A AMOVA

revelou que a maior parte da variação molecular (>98%) foi encontrada dentro dos grupos quando se utilizou os critérios: teor de compostos cianogênicos, cor de polpa das raízes, origem genética e geográfica dos acessos (Tabela 4). Menos de 1% da variância molecular foi encontrada entre os grupos de diversidade formados por critérios fenotípicos e de origem. Embora ainda pouco significativa, uma maior variância molecular foi observada entre grupos dentro da origem geográfica (1,5%). Porém, ainda é insuficiente para explicar a maior parte da variância genética identificada nos acessos de mandioca. Portanto, os resultados da AMOVA corroboraram para informações prévias referentes à baixa associação entre este padrão de agrupamento e o agrupamento formado pela ADCP.

**Tabela 4**. Resultados da análise de variância molecular (AMOVA) realizada em 1.580 indivíduos com base em marcadores SNPs.

| **Fonte de variação** | **Origem genética** | | | | **Teor de compostos cianogênicos** | | | |
|---|---|---|---|---|---|---|---|---|
| | GL | QM | % var | $F_{ST}$ | GL | QM | % var | $F_{ST}$ |
| Entre grupos | 1 | 151,67 | 0,25 | 0,00 | 7 | 95,55 | 0,29 | 0,00 |
| Dentro dos grupos | 1578 | 63,22 | 99,75 | | 1572 | 63,13 | 99,71 | |
| Total | 1579 | 63,27 | 100,00 | | 1579 | 63,27 | 100,00 | |
| **Fonte de variação** | **Cor da polpa das raízes** | | | | **Grupos ADCP** | | | |
| | GL | QM | % var | $F_{ST}$ | GL | QM | % var | $F_{ST}$ |
| Entre grupos | 4 | 132,74 | 0,72 | 0,01 | 21 | 719,78 | 14,85 | 0,47 |
| Dentro dos grupos | 1575 | 63,09 | 99,28 | | 1558 | 54,42 | 85,15 | |
| Total | 1580 | 63,27 | 100,00 | | 1579 | 63,27 | 100,00 | |
| **Fonte de variação** | **Origem geográfica** | | | | | | | |
| | GL | QM | % var | $F_{ST}$ | | | | |
| Entre grupos | 7 | 75,30 | 0,09 | 0,01 | | | | |
| Entre grupos dentro de origem geográfica | 27 | 109,64 | 1,50 | | | | | |
| Dentro dos grupos | 1545 | 62,41 | 98,41 | | | | | |
| Total | 1579 | 63,27 | 100,00 | | | | | |

O agrupamento da ADCP como fator hierárquico na AMOVA foi capaz de alocar maior variância molecular entre grupos (14,85%) sendo, portanto,

mais efetiva que os critérios fenotípicos e de origem para explicar a diversidade genética do germoplasma de mandioca. Contudo, ainda a maior parte da variância molecular foi alocada dentro dos grupos da ADCP (85,15%). Isto faz com que as atividades rotineiras de manutenção deste germoplasma, bem como no seu uso como parentais em cruzamentos, ainda devam levar em consideração a ampla diversidade genética dentro dos grupos.

A proporção da diferenciação entre os grupos avaliada pelo índice de diferenciação genética ($F_{ST}$), variou de 0,00 a 0,47, portanto, parte dos agrupamentos está fixado em diferentes frequências alélicas (grupos da ADCP) indicando diferenciação moderada entre os grupos, e parte deles (classificados quanto à origem genética, teor de compostos cianogênicos, cor da polpa das raízes e origem geográfica), possuem alelos em frequências alélicas bastante similares, sem nenhuma diferenciação aparente.

## DISCUSSÃO

### *Polimorfismo dos SNPs e diversidade genética no germoplasma de mandioca*

Dos 444.821 SNPs identificados no genoma de *Manihot esculenta* Crantz, 20.610 (4,63%) SNPs passaram pelo filtro de qualidade para uso nas análises genéticas. Destes SNPs, apenas nove foram caracterizados como monomórficos, e por isso, excluídos das análises. O número de SNPs de alta qualidade identificados pela GBS pode ser afetado pelo tamanho do genoma, cobertura do sequenciamento e os objetivos do estudo para os quais os marcadores serão utilizados.

Os primeiros estudos relacionados à identificação de SNPs em maior número em mandioca foram realizados em bibliotecas de EST (*expressed sequence tags*) e de modo geral, o número de SNPs não redundantes, variou entre 2.000 e 3.000 (FERGUSON et al., 2012; POOTAKHAM et al., 2014). Com o surgimento de técnicas de genotipagem em larga escala, a exemplo da GBS, o número de SNPs identificados em mandioca aumentou consideravelmente para mais de 56 mil (RABBI et al., 2015). A despeito do uso de técnicas semelhantes (GBS), a diferença no número de SNPs relatos por Rabbi et al. (2015), em comparação com o presente trabalho, possivelmente deve-se ao uso de menor MAF e tolerância de maior número de dados perdidos.

Considerando a ampla cobertura genômica que o método GBS apresenta, assim como a diversidade do germoplasma de mandioca, foi possível identificar diversos SNPs trialélicos (2,51%) e tetra-alélicos (0,11%). Embora SNPs tri e tetra-alélicos sejam relativamente raros, seu relato tem sido feito em diversas espécies. Ao utilizarem a GBS para detectar SNPs no germoplasma da espécie *Ziziphus jujuba*, Chen et al. (2017) relataram que 0,80% dos SNPs identificados (38 de 4680) foram caracterizados como trialélicos. Relatos desta natureza também foram feitos na análise do germoplasma de três espécies de cafeeiro (*Coffea arabica*, *C. canephora* e *C. racemosa*), cujos SNPs tri e tetra-alélico foram observados em 3,59% e 1,51% dos SNPs identificados (ZHOU et al., 2016). De acordo com JI et al. (2013) a presença de SNPs em regiões com elevada taxa de mutação resulta no aumento do polimorfismo, levando esse marcador a apresentar comportamento multialélico.

De modo geral, a distribuição genômica dos SNPs de mandioca foi relativamente uniforme, considerando uma densidade média de 26,85±4,22 SNPs/Kb. Esuma et al. (2016), também relataram uma densidade de SNPs muito próximo ao presente trabalho (23 SNPs/Kb). Alguns autores mencionam que a existência de *hotspots* em algumas regiões genômicas, e diversas pressões de seleção, podem resultar em variações na densidade dos SNPs ao longo do genoma de diferentes espécies (ROGOZIN; PAVLOV, 2003).

Em relação aos parâmetros de diversidade genética, a variação do PIC de 0,04 a 0,61 (média de 0,24) foi similar a outros estudos em mandioca com marcadores SNPs, cujo PIC médio foi de 0,28 (FERGUSON et al., 2012) e 0,26 (OLIVEIRA et al., 2014a). Em arroz, valores de PIC similares entre marcadores SNPs (0,23) e SSR (0,25), demonstraram que dependendo do sistema reprodutivo, ambos os tipos de marcadores podem fornecer o mesmo padrão de informação (SINGH et al., 2013). Assim, diversos fatores como, o tipo de reprodução da espécie, a diversidade genética e o tamanho da coleção, a sensibilidade do método de genotipagem e a localização genômica dos marcadores, provavelmente, exercem forte influência sobre os valores de PIC.

Especificamente em mandioca, os valores de PIC reportados na literatura, com base em marcadores SSR, variaram de 0,09 a 0,99 (FREGENE et al., 2003; RAGHU et al., 2007). Esta diferença no polimorfismo dos SSR em

comparação com os SNPs, deve-se à natureza multialélica dos primeiros marcadores (JONES et al., 2007; REN et al., 2013) e ao fato da mandioca ser uma espécie alógama, com possibilidade de identificação de vários alelos por loco. Alguns autores têm demonstrado que tanto SNPs quanto SSRs, são igualmente apropriados para se estimar a diversidade genética em germoplasma (INGHELANDT et al., 2010). No entanto, se o objetivo é estudar a variação intraespecífica do germoplasma, os SNPs são mais apropriados do que os SSRs devido à ampla distribuição nos genomas, alto grau de automação no processo de genotipagem, de forma a gerar milhares de marcas ao longo do genoma com baixo custo, além do maior potencial de identificação de forte associação com características de interesse agronômico (SYVANEN, 2001; APPLEBY et al., 2009; ANITHAKUMARI et al., 2010; HAYWARD et al., 2012).

Os valores de PIC acima de 0,50 para os marcadores SNPs, ocorreram devido à presença de loco tri (0,17%) e tetra-alélicos (0,02%). Situação semelhante também foi reportada em outras espécies. De acordo com Chen et al. (2017), a presença de SNPs trialélicos no germoplasma de *Ziziphus jujuba*, também favoreceu um aumento de 11% do valor de PIC (~0,38), onde a média desse parâmetro para os locos bialélicos foi de 0,27. Contudo, estes valores ainda mantiveram-se dentro das condições teóricas para o comportamento mais frequente (bialélico) destes marcadores.

Os valores médios de He (0,30) e Ho (0,24) obtidos pelos SNPs do presente estudo, estão próximos dos valores relatos por outros autores, a exemplo, 0,35 e 0,36, respectivamente (FERGUSON et al., 2012) e 0,32 e 0,32; respectivamente (OLIVEIRA et al., 2014a). A endogamia dos SNPs variou de -1,00 a 0,96, dos quais, 12,77% mostraram excesso de heterozigotos (-1,00 a -0,001). Por outro lado, a endogamia da maioria dos SNPs (79,75%) variou de 0,00 a 0,49, indicando a presença de muitos locos com excesso de homozigose. Embora fatores como dispersão gênica e sistema reprodutivo influenciam os padrões de diversidade genética dentro e entre populações (MELONI et al., 2013), o processo histórico de domesticação e seleção exerce um grande efeito na redução da diversidade genética em determinadas regiões genômicas. De acordo com Vigouroux et al. 2005, evidências de domesticação e seleção artificial em diferentes características morfológicas e agronômicas no

milho, apresentaram um efeito na redução da diversidade genética calculada com marcadores microssatélites. Lopes et al. (2015) também observaram diversos locos microssatélites homozigóticos em algumas populações de milho, associados à baixa heterozigosidade média. De acordo com Bilska e Szczecinska (2016), outros fatores como mecanismos evolutivos em diferentes populações e a taxa de evolução, contribuem para mudanças que podem ocorrer de forma diferenciada ao longo do genoma, o que pode ajudar a explicar as diferenças nos parâmetros de diversidade em diferentes marcadores SNPs distribuídos no genoma da mandioca.

Em nível de indivíduo, observou-se elevada endogamia (variação de 0,49 a 0,97, média de 0,69), sendo que três acessos (BGM0104, BGM1226 e BGM1448) apresentaram $f > 0,90$. De modo geral, em pequenas populações alógamas, é esperado elevado grau de parentesco e, consequentemente maior endogamia entre os indivíduos (WILLI et al., 2013). Por outro lado, mesmo sendo alógama, a arquitetura da inflorescência e os mecanismos de dispersão das sementes de mandioca, limitam o fluxo genético de longa distância; favorecem o acasalamento entre plantas próximas (KAWUKI et al., 2013); e até mesmo a autofecundação por cruzamento entre diferentes plantas de um mesmo clone. Neste último caso, é possível que sejam formadas subpopulações compostas por indivíduos aparentados e com elevado grau de endogamia, que pode ser mantida no germoplasma pela propagação vegetativa.

Embora o aumento da endogamia esteja associado com a existência de forte depressão por endogamia (ROJAS et al., 2009; KAWUKI et al., 2011b; FREITAS et al.; 2016), recentemente, a busca por linhagens endogâmicas tem despertado o interesse dos melhoristas de mandioca porque possibilita reduzir a carga genética dos genótipos; descobrir características recessivas úteis; facilitar o intercâmbio e conservação de germoplasma; e contribuir para o desenvolvimento de híbridos superiores (CEBALLOS et al., 2015). Os acessos mais endogâmicos podem ser direcionados para a formação de linhagens puras divergentes e complementares entre si que, posteriormente, podem ser cruzadas para exploração do vigor híbrido e lançamento de novos clones com características desejáveis (FREITAS et al., 2016). Portanto, os marcadores SNPs podem ser utilizados para descobrir parentais e selecionar indivíduos

segregantes mais homozigóticos via seleção assistida para acelerar o desenvolvimento de linhagens endogâmicas.

***Desequilíbrio de Ligação***

Estudos sobre LD vêm sendo aplicados à cultura da mandioca como abordagem complementar para estudos de localização de QTLs (RABBI et al., 2014); associação genômica (WOLFE et al., 2016); seleção genômica e construção de mapas de ligação (RAMU et al., 2017); bem como para estudos de diversidade genética e estrutura de populacional em bancos de germoplasma (OLIVEIRA et al., 2014a).

No germoplasma de mandioca analisado, o LD decaiu mais bruscamente próximo de 8 kb, porém com melhor ajuste da regressão não linear entre 15 e 20 kb, com valores de $r^2$= 0,20. Ao analisarem um painel com 6.128 acessos de mandioca da África, Wolfe et al. (2016) relataram que o LD se estendeu a uma distância de 10 a 50 kb ($r^2$ >0,20), enquanto Ramu et al. (2017) reportaram que o LD decaiu a uma menor distância ($r^2$= 0,10 próximo de 3 kb), ao avaliaram um painel composto por 241 acessos de mandioca predominantemente de origem africana via *whole-genome sequencing* (WGS). Estas diferenças no LD estão relacionadas, sobretudo à complexidade e tamanho do genoma da espécie, sistema reprodutivo e a quantidade de marcadores utilizados para capturar com precisão toda a variação genética existente. Outros fatores como a deriva genética, estrutura da população e seleção, também podem afetar a extensão do LD. Em estudos de associação genômica, estudos prévios sobre o LD podem melhorar o poder estatístico e diminuir a taxa de falsos positivos na descoberta de genes candidatos ligados a QTLs (CARPUTO et al., 2003).

O entendimento da extensão do desequilíbrio de ligação (LD) é de fundamental importância para a clonagem posicional de genes de interesse, pois os SNPs mapeados e ordenados nos cromossomos podem ser agrupados em grupos distintos de haplotipos (GABRIEL et al., 2002). Quando os SNPs estão em forte LD, alelos de alguns SNPs em determinado haplótipo podem predizer os alelos dos outros SNPs, por isso é possível reduzir o número de SNPs a serem utilizados nas análises genéticas ao considerar as informações do LD. Além disso, uma das principais limitações das análises de associação é

a exigência de número suficiente de marcadores para fornecer uma alta probabilidade de identificação de marcadores em LD com todos os QTLs da característica de interesse. Portanto, a densidade de marcadores necessária para uma análise de associação e subsequente implementação da seleção assistida por marcadores depende da extensão do LD em todo o genoma (KHATKAR et al., 2008).

A compreensão dos padrões de LD no genoma da mandioca poderá contribuir para os estudos de mapeamento associativo para localizar variantes genéticas que influenciam características de interesse como teor de matéria seca, resistência a pragas e doenças, produtividade da parte aérea, raiz e amido. Portanto, o efeito da extensão do LD no germoplasma de mandioca e sua aplicação na seleção assistida será refinado em outros estudos de GWAS.

***Estrutura Populacional com uso da ADCP***

A análise dos SNPs com base na ADCP indicou a formação de vinte dois (22) grupos. A formação do agrupamento do germoplasma de mandioca pelo método ADCP foi realizada sem critérios definidos *à priori* para classificação e ordenamento. Em espécies de propagação clonal como a mandioca, as variedades são muitas vezes derivadas de complexos cruzamentos, e embora este tipo de estrutura populacional possa influenciar a precisão na definição do número de grupos (RABBI et al., 2015), o número de grupos sugeridos pela ADCP, normalmente é muito próximo do valor real (JOMBART et al., 2010).

De fato, Pometti et al. (2014) compararam o número de grupos sugeridos por diferentes técnicas como ADCP e análise bayesiana baseada em equilíbrio de ligação e de Hardy-Weinberg a exemplo dos modelos implementados nos programas STRUCTURE (PRITCHARD et al., 2000) e GENELAND (GUILLOT et al., 2005). Os resultados demonstraram que a ADCP se mostrou tão precisa quanto às demais abordagens para inferir o número ideal de agrupamentos para representar a organização da estrutura populacional na espécie *Acacia caven*. Ademais, a ADCP também forneceu elevadas probabilidades de alocação dos indivíduos aos diferentes grupos (>0,85), indicando ser uma técnica bastante confiável para definição da estrutura genética. Além disso, possui uma baixa demanda computacional,

sobretudo por não levar em consideração as suposições de panmixia, equilíbrio de Hardy-Weinberg e desequilíbrio de ligação (JOMBART et al., 2010).

A probabilidade de alocação dos acessos de mandioca do presente trabalho variou de 0,432 a 1,00, porém em média, a alocação dos acessos de mandioca a um determinado grupo, foi de aproximadamente 0,99, indicando que a maioria dos acessos de mandioca do Brasil foram alocados com bastante confiabilidade ao seu respectivo agrupamento. Contudo, não foi possível estabelecer nenhuma associação específica entre os agrupamentos formados pela ADCP com: i) dados de origem geográfica dos acessos; ii) informações fenotípicas dos acessos de mandioca para cor da polpa das raízes e teor de compostos cianogênicos; e iii) origem genética em termos de classificação em variedades locais e melhoradas. A ausência de um padrão claro de agrupamento com base em informações moleculares e fenotípicas também foi relatada em outras culturas como milho (SEMAGN et al., 2012) e girassol (FILIPPI et al., 2015). Na mandioca a falta de um padrão claro de agrupamento com base nos dados geográficos e de origem genética é justificável considerando o intenso intercâmbio entre germoplasma no Brasil, e ao fato de que muitas variedades recomendadas pelas instituições de pesquisa são variedades locais com alta estabilidade e produtividade de raízes. Com isso, a classificação entre variedade melhorada e local é bastante complexa. Por outro lado, a classificação dos acessos de mandioca com base na cor de polpa das raízes e teor de compostos cianogênicos parece ser bastante simplória para agrupar a diversidade do germoplasma de mandioca, considerando que poucos genes podem ser responsáveis pela expressão destas características. Por outro lado, a caracterização molecular é a melhor maneira de descobrir a constituição genética dos acessos, e os marcadores SNPs têm se mostrado uma poderosa alternativa para determinar a diversidade genética da espécie (RABBI et al., 2015), além de identificar lacunas no germoplasma que podem ser preenchidas e, entender melhor a complexidade das relações entre cultivares e variedades locais, apesar das informações limitadas de passaportes.

***Análise de variância molecular (AMOVA)***

A maior parte da diversidade genética (>98%) no germoplasma de mandioca está presente dentro dos agrupamentos hierárquicos (origem genética, teor de compostos cianogênicos, cor da polpa das raízes e origem geográfica). Isto demonstra que a grande diversidade genética existente neste germoplasma traz uma enorme complexidade na categorização dos acessos. Por outro lado, a variância molecular obtida pelos grupos formados pela ADCP resultou em uma maior variação entre grupos (14,85%), indicando que o agrupamento com os marcadores SNPs foi mais efetivo em relação aos critérios fenotípicos, origem geográfica e genética na discriminação da variação genética entre grupos. Resultados semelhantes foram reportados por Costa et al. (2013), onde 77% da variação genética também foi observada dentro dos agrupamentos teóricos formados a partir da análise de estrutura populacional. Outros estudos em mandioca também demostraram pouca diferenciação genética entre agrupamentos baseados em informações moleculares comparadas com informações fenotípicas e/ou dados de passaporte (MONTEIRO-ROJAS et al., 2011; TURYAGYENDA et al., 2012; BEOVIDES et al., 2015; ORTIZ et al., 2016; GONÇALVES et al., 2017).

Os baixos valores do índice de fixação alélica ($F_{ST}$) para a maioria dos agrupamentos hierárquicos no germoplasma de mandioca corroboram com as informações da AMOVA, quanto à maior distribuição da variação dentro dos grupos. Ao avaliarem diversos acessos de mandioca de origem africana, Kawuki et al. (2013) também observaram uma diferenciação moderada entre os acessos baseados em marcadores microssatélites levando-se em consideração a origem geográfica ($F_{ST}$=0,104) e genética das variedades locais ($F_{ST}$=0,104) e melhoradas ($F_{ST}$= 0,084). Por outro lado, o $F_{ST}$ obtido com base em marcadores AFLP (0,656) e SSR (0,746) por RAJI et al. (2009) foi bastante superior aos observados no presente trabalho. De acordo com estes mesmos autores, a diferenciação genética foi menor nos acessos elites do que entre as variedades locais originarias da África, como resultado da alta pressão de seleção durante o desenvolvimento de cultivares elites, juntamente com a deriva genética. Essa alta diferenciação genética entre variedades locais e cultivares elite, sugere que variações geográficas ou regionais podem ser responsáveis pela maior parte da diferenciação genética observada.

Estudos anteriores com o germoplasma de mandioca do Brasil também relataram maior parte da diversidade dentro dos agrupamentos baseados em critérios fenotípicos, moleculares e dados de passaporte (OLIVEIRA et al., 2014a). De acordo com Ortiz et al. (2016), a maior parte da variação genética observada em variedades de mandioca de mesa coletadas em diferentes regiões do sul do Brasil detectada por microssatélites, foi observada dentro dos agrupamentos ($F_{ST}$ = 0,107). Por outro lado, a análise de variedades de mandioca da Amazônia cultivada em diferentes tipos de solo com marcadores microssatélites, revelou uma diferenciação das variedades cultivadas nas planícies inundadas em comparação com as cultivadas em oxisolos ($F_{ST}$ = 0,093) e terra preta da Amazônia ($F_{ST}$ = 0,108), sugerindo importante estruturação genética em função dos diferentes tipos de solo (ALVES-PEREIRA et al., 2011).

## CONCLUSÃO

Um painel com cerca de 20 K SNPs distribuídos de forma bastante equilibrada nos diferentes cromossomos da mandioca foram utilizados para o entendimento do desequilíbrio de ligação, diversidade genética e estrutura populacional no germoplasma de mandioca da América Latina. A He (0,30) e Ho (0,24) média foram consistentes com o esperado de marcadores majoritariamente bialélicos como os SNPs e com a natureza genética do polimorfismo de espécies alógamas, porém de propagação vegetativa, como é o caso da mandioca. Cerca de 49% dos SNPs com elevado conteúdo informativo (PIC > 0,25) poderão ser utilizados na elaboração de chips para análise de diversidade genética de forma rotineira nos bancos ativos de germoplasma de mandioca. A ADCP indicou a formação de 22 grupos de diversidade genética, com elevada probabilidade de alocação dos indivíduos dentro de cada grupo. Embora o agrupamento formado pela ADCP não tenha possibilitado um perfil claro de distinção dos acessos de mandioca com base em dados agronômicos, de origem geográfica e genética, a variância molecular explicada pelo agrupamento da ADCP foi capaz de maximizar a variação entre grupos (14,85%), sendo, portanto, uma técnica bastante útil para categorização de germoplasma vegetal. Este trabalho contribui para o entendimento da diversidade molecular do germoplasma de mandioca para otimização do

processo de conservação, caracterização e uso do germoplasma, em estudos de mapeamento genômico, GWAS, grupos contrastantes de diversidade para geração de populações segregantes e desenvolvimento de novas variedades.

**REFERÊNCIAS**

AGRAMA, H. A.; YAN, W.; JIA, M.; FJELLSTROM, R.; McCLUNG, M. A. Genetic structure associated with diversity and geographic distribution in the USDA rice world collection. **Natural Science**, v. 2, p. 247-291, 2010.

ALVES-PEREIRA, A.; PERONI, N.; ABREU, A. G.; GRIBEL, R.; CLEMENT, C. R. Genetic structure of traditional varieties of bitter manioc in three soils in Central Amazonia. **Genetics**, v. 139, p. 1259-1271, 2011.

ANITHAKUMARI, A. M.; TANG, J.; VAN ECK, H. J.; VISSER, R. G.; LEUNISSEN, J. A.; VOSMAN, B.; VAN DER LINDEN, C. G. A pipeline for high throughput detection and mapping of SNPs from EST databases. **Molecular Breeding**, v. 26, p. 65–75, 2010.

APPLEBY, N.; EDWARDS, D.; BATLEY, J. New technologies for ultra-high throughput genotyping in plants. **Methods in Molecular Biology**, v. 513, p. 19-39, 2009.

ATWELL, S.; HUANG, Y. U. S.; VILHJÁLMSSON, B. J.; WILLEMS, G.; HORTON, M.; LI, Y.; MENG, D.; PLATT, A.; TARONE, A. M.; HU, T. T.; JIANG, R.; MULIYATI, N. W.; ZHANG, X.; AMER, M. A.; BAXTER, I.; BRACHI, B.; CHORY, J.; DEAN, C.; DEBIEU, M.; MEAUX, J.; ECKER, J. R.; FAURE, N.; KNISKERN, J. M. JONES, J. D. G.; MICHAEL, T.; NEMRI, A.; ROUX, F.; SALT, D. E.; TANG, C.; TODESCO, M.; TRAW, M. B.; WEIGEL, D.; MARJORAM, P.; BOREVITZ, J. O.; BERGELSON, J.; NORDBORG, M. Genome-wide association study of 107 phenotypes in a common set of *Arabidopsis thaliana* inbred lines. **Nature**, v. 3, p. 627-631, 2010.

AZEVEDO, C. F.; RESENDE, M. D. V.; SILVA, F. F.; VIANA, J. M. S.; VALENTE, M. S. F.; RESENDE JUNIOR, M. F. R.; OLIVEIRA, E. J. New

accuracy estimators for genomic selection with application in a cassava (*Manihot esculenta*) breeding program. **Genetics and Molecular Research**, v. 15, p. 1-14, 2016.

BEOVIDES, Y.; FREGENE, M.; GUTIÉRREZ, J. P.; MILIÁN, M. D.; COTO, O.; BUITRAGO, C.; CRUZ, J. A.; RUIZ, E.; BASAIL, M.; RAYAS, A.; RODRÍGUEZ, D.; SANTOS, A.; LÓPEZ, J.; MEDERO, V. Molecular diversity of Cuban cassava (*Manihot esculenta* Crantz) cultivars assessed by simple sequences repeats (SSR). **Agronomy, Society and Environment**, v. 19, p. 364-377, 2015.

BILSKA, K.; SZCZECINSKA, M. Comparison of the effectiveness of ISJ and SSR markers and detection of outlier locos in conservation genetics of *Pulsatilla patens* populations. **PeerJ**, v. 4, p. 1-17, 2016.

BOTSTEIN, D.; WHITE, R. L.; SKOLNICK, M.; DAVIS, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **American Journal of Human Genetics**, v. 32, p. 314-331, 1980.

BRADBURY, P. J.; ZHANG, Z.; KROON, D. E.; CASSTEVENS, T. M.; RAMDOSS, Y.; BUCKLER, E. S. TASSEL: software for association mapping of complex traits in diverse samples. **Bioinformatics Applications Note**, v. 23, p. 2633–2635, 2007.

BREDESON, J. V.; LYONS, J. B.; PROCHNIK, S. E.; WU, G. A.; HA, C. M.; EDSINGER-GONZALES, E.; GRIMWOOD, J.; SCHMUTZ, J.; RABBI, I. Y.; EGESI, C.; NAULUVULA, P.; LEBOT, V.; NDUNGURU, J.; MKAMILO, G.; BART, R. S.; SETTER, T. L.; GLEADOW, R. M.; KULAKOW, P.; FERGUSON, M. E.; ROUNSLEY, S.; ROKHSAR, D. S. Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. **Nature**, v. 34, p. 562-571, 2016.

CARMO, C. D.; SILVA, M. S.; OLIVEIRA, G. A. F.; OLIVEIRA, E. J. Molecular-assisted selection for resistance to cassava mosaic disease in *Manihot*

*esculenta* Crantz. **Scientia Agricola**, v. 72, p. 520-527, 2015.

CARPUTO, D.; FRUSCIANTE, L.; PELOQUIN, S. J. The role of 2n gametes and endosperm balance number in the origin and evolution of polyploids in the tuber-bearing Solanums. **Genetics**, v. 163, p. 287-294, 2003.

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

CHEN, W.; HOU, L.; ZHANG, Z.; PANG, X.; LI, Y. Genetic Diversity, population structure, and linkage disequilibrium of a core collection of *Ziziphus jujuba* assessed with genome-wide SNPs developed by genotyping-by-sequencing and SSR Markers. **Frontiers in Plant Science**, v. 8, p. 1-14, 2017.

COSTA, T. R.; VIDIGAL FILHO, P. S.; GONÇALVES-VIDIGAL, M. C.; GALVÁN, M. Z.; LACANALLO, G. F.; SILVA, L. I.; KVITSCHAL, M. V. Genetic diversity and population structure of sweet cassava using simple sequence repeat (SSR) molecular markers. **African Journal of Biotechnology**, v. 12, p. 1040-1048, 2013.

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, v.19, p.11-15, 1987.

DUMET, D.; KORIE, S.; ADEYEMI, A. Cryobanking cassava germplasm at IITA. **Acta Horticulturae**, v. 908, p. 439-446, 2011.

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

ESUMA, W.; HERSELMAN, L.; LABUSCHAGNE, M. T.; RAMU, P.; LU, F.; BAGUMA, Y.; BUCKLER, E. S.; KAWUKI, R. S. Genome-wide association

mapping of provitamin A carotenoid content in cassava. **Euphytica**, v. 212, p. 97-110, 2016.

FERGUSON, M. E.; HEARNE, S. J.; CLOSE, T. J.; WANAMAKER, S.; MOSKAL, W. A.; TOWN, C. D.; YOUNG, J.; MARRI, P. R.; RABBI, I. Y.; VILLIERS, E. P. Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. **Theoretical Applied Genetics**, v. 124, p. 685–695, 2012.

FILIPPI, C. V.; AGUIRRE, N.; RIVAS, J. G.; ZUBRZYCKI, J.; PUEBLA, A.; CORDES, D.; MORENO, M. V.; FUSARI, C. M.; ALVAREZ, D.; HEINZ, R. A.; HOPP, H. E.; PANIEGO, N. B.; LIA, V. V. Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. **BioMed Central Plant Biology**, v. 15, p. 1-12, 2015.

FREGENE, M. A.; SUAREZ, M.; MKUMBIRA, J.; KULEMBEKA, H.; NDEDYA, E.; KULAYA, A.; MITCHEL, S.; GULLBERG, U.; ROSLING, H.; DIXON, A. G.; DEAN, R.; KRESOVICH, S. Simple sequence repeat marker diversity in cassava landraces: genetic diversity and differentiation in an asexually propagated crop. **Theoretical Applied Genetics**, v. 107, p. 1083–1093, 2003.

FREITAS, J. P. X.; SANTOS, V. S.; OLIVEIRA, E. J. Inbreeding depression in cassava for productive traits. **Euphytica**, v. 209, p.137–145, 2016.

GABRIEL, S.B.; SCHAFFNER, S.F.; NGUYEN, H.; MOORE, J.M.; ROY, J.; BLUMENSTIEL, B.; HIGGINS, J.; DEFELICE, M.; LOCHNER, A.; FAGGART, M.; LIU-CORDERO, S.N.; ROTIMI, C.; ADEYEMO, A.; COOPER, R.; WARD, R.; LANDER, E.S.; DALY, MJ.; ALTSHULER, D. The structure of haplotype blocks in the human genome. **Science**, v. 296, p.2225-2229, 2002.

GLAUBITZ, J. C.; CASSTEVENS, T. M.; LU, F.; HARRIMAN, J.; ELSHIRE, R. J.; SUN, Q.; BUCKLER, E. S. TASSEL GBS: A high capacity genotyping-by-sequencing analysis pipeline. **Plos One**, v. 9, p. 1-11, 2014.

GONÇALVES, T. M.; VIDIGAL FILHO, P. S.; VIDIGAL, M. C. G.; FERREIRA, R. C. U.; ROCHA, V. P. C.; ORTIZ, A. H. T.; MOIANA, L. D.; KVITSCHAL, M. V. Genetic diversity and population structure of traditional sweet cassava accessions from Southern of Minas Gerais State, Brazil, using microsatellite markers. **African Journal of Biotechnology**, v. 16, p. 346-358, 2017.

GUILLOT, G.; ESTOUP, A.; MORTIER, F.; COSSON, J. F. A spatial statistical model for landscape genetics. **Genetics**, v. 170, p. 1261-1280, 2005.

HAMBLIN, M. T.; RABBI, I. Y. The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in cassava (*Manihot esculenta*). **Crop Science**, v. 54, p. 1–6, 2014.

HAYWARD, A.; MASON, A. S.; MORGAN, J. D.; ZANDER, M.; EDWARDS, D.; BATLEY, J. SNP discovery and applications in *Brassica napus*. **Plant Biotechnology Journal**, v. 39, p. 49-61, 2012.

INGHELANDT, D. V.; MELCHINGER, A. E.; LEBRETON, C.; STICH, B. Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. **Theoretical Applied Genetics**, v. 120, p. 1289–1299, 2010.

JI, K.; ZHANG, D.; MOTILAL, A. L.; BOCCARA, M.; LACHENAUD, P.; MEINHARDT, W. L. Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. **Genetic Resources and Crop Evolution**, v. 60, p. 441-453, 2013.

JOMBART, T. adegenet: a R package for the multivariate analysis of genetic markers. **Bioinformatics**, v. 24, p. 1403-1405, 2008.

JOMBART, T.; DEVILLARD, S.; BALLOUX, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. **BioMed Central Genetics**, v.11, p. 1-15, 2010.

JONES, E. S.; SULLIVAN, H.; BHATTRAMAKKI, D.; SMITH, J. S. C. A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays* L.). **Theoretical Applied Genetics**, v.115, p. 361–371, 2007.

KAWUKI, R. S.; FERGUSON, M. E.; LABUSCHAGNE, M. T.; HERSELMAN, L.; ORONE, J.; RALIMANANA, I.; BIDIAKA, M.; LUKOMBO, S.; KANYANGE, M. C.; GASHAKA, G.; MKAMILO, G.; GETHIH, J.; OBIERO, H. Variation in qualitative and quantitative traits of cassava germplasm from selected national breeding programmers in sub-Saharan Africa. **Field Crops Research**, v. 2, p. 151-156, 2011a.

KAWUKI, R. S.; HERSELMAN, L.; LABUSCHAGNE, M. T.; NZUKI, I. Genetic diversity of cassava (*Manihot esculenta* Crantz) landraces and cultivars from southern, eastern and central Africa. **Plant Genetic Resources**, v. 11, p. 170-181, 2013.

KAWUKI, R. S.; NUWAMANYA, E.; LABUSCHAGNE, M. T.; HERSELMAN, L.; FERGUSON, M. E. Segregation of selected agronomic traits in six $S_1$ cassava families. **Journal of Plant Breeding and Crop Science**, v. 3, p. 154-160, 2011b.

KHATKAR, M.; NICHOLAS, F.; COLLINS, A.; ZENGER, K.; CAVANAGH, J.; BARRIS, W.; SCHNABEL, R.; TAYLOR, J.; RAADSMA, H. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics, v.9, p.187, 2008.

LEBOT, V. Tropical root and tuber crops: cassava, sweet potato, yams and aroids. **Crop Production Science in Horticultur**e, v. 2, p. 432, 2009.

LEFÈVRE, F.; CHARRIER, A. Isozyme diversity within African *Manihot* germplasm. **Euphytica** v. 66, p. 73-80, 1993.

LOPES, A.D.; SCAPIM, C.A.; MACHADO, M.F.P.S.; MANGOLIN, C.A.; SILVA, T.A.; CANTAGALI, L.B.; TEIXEIRA, F.F.; MORA, F. Genetic diversity assessed by microsatellite markers in sweet corn cultivars. **Scientia Agricola**, v.72, p.513-519, 2015.

MAMMADOV, J. A.; AGGARWAL, R.; BUYYARAPU, R.; KUMPATLA, S. SNP Markers and their impact on plant breeding. **International Journal of Plant Genomics**, v. 2012, p. 1-12, 2012.

MELO, A. T. O.; GUTHRIE, R. S.; HALE, I. GBS-Based deconvolution of the surviving north American collection of cold-hardy kiwifruit (*Actinidia* spp.) germplasm. **Plos One**, v. 12, p. 1-21, 2017.

MELONI, M.; REID, A.; CAUJAPÉ-CASTELLS, J.; MARRERO, A.; FERNANDEZ-PALACIOS, J. M.; MESA-COELO, R. A.; CONTI, E. Effects of clonality on the genetic variability of rare, insular species: the case of *Ruta microcarpa* from the Canary Islands. **Ecology Evolution**, v. 3, p. 1569-1579, 2013.

MONTERO-ROJAS, M.; CORREA, A. M.; SIRITUNG D. Molecular differentiation and diversity of cassava (*Manihot esculenta* Crantz) taken from 162 locations across Puerto Rico and assessed with microsatellite markers. **AoB Plants**, v. 10, p. 1-13, 2011.

MORRIS, G. P.; RAMU, P.; DESHPANDE, S. P.; HASH, C. T.; SHAH, T.; UPADHYAYA, H. D.; RIERA-LIZARAZU, O.; BROWN, P. J.; ACHARYA, C. B.; MITCHELL, S. E.; HARRIMAN, J.; GLAUBITZ, J. C.; BUCKLER, E. S.; KRESOVICH, S. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. **Proceedings of the National Academy of Science of the United States of America (PNAS)**, v. 110, p. 453-458, 2013.

MTUNGUJA, M. K.; RANJAN, A.; LASWAI, H. S.; MUZANILA, Y.; NDUNGURU, J.; SINHA, N. R. Genetic diversity of farmer-preferred cassava landraces in Tanzania based on morphological descriptors and single nucleotide

polymorphisms. **Plant Genetic Resources: Characterization and Utilization**, v. 15, p. 138-146, 2017.

NASSAR, N. M. Cassava, *Manihot esculenta* Crantz, genetic resources: origin of the crop, its evolution and relationships with wild relatives. **Genetics and Molecular Research**, v. 1, p. 298-305, 2002.

NEI, M. Analysis of gene diversity in subdivided populations. **Proceedings of the National Academy of Science of the United States of America (PNAS)**, v. 70, p. 3321-3323, 1973.

OKOGBENIN, E.; PORTO, M. C. M. EGESI, C.; MBA, C.; ESPINOSA, E.; SANTOS, L. G.; OSPINA, C.; MARÍN, J.; BARRERA, E.; GUTIÉRREZ, J.; EKANAYAKE, I.; IGLESIAS, C.; FREGENE, M. A. Marker-assisted introgression of resistance to cassava mosaic disease into Latin American germplasm for the genetic improvement of cassava in Africa. **Crop Science**, v. 47, p. 1895-1904, 2007.

OLIVEIRA, E. J.; FERREIRA, F. C.; SANTOS, V. S.; JESUS, N. O.; OLIVEIRA, F. A. G.; SILVA, S. M. Potential of SNP markers for the characterization of Brazilian cassava germplasm. **Theoretical and Applied Genetics**, v. 127, p. 1423-1440, 2014a.

OLIVEIRA, E. J.; FERREIRA, F. C.; SANTOS, V. S.; OLIVEIRA, G. A. F. Development of a cassava core collection based on single nucleotide polymorphism markers. **Genetics and Molecular Research**, v. 13, p. 6472-6485, 2014b.

OLIVEIRA, E. J.; OLIVEIRA FILHO, O. S.; SANTOS, V. S. Classification of cassava genotypes based on qualitative and quantitative data. **Genetics and Molecular Research**, v. 14, p. 906-924, 2015.

OLIVEIRA, E. J.; RESENDE, M. D. V.; SANTOS, V. S.; FEREIRA, C. F.; OLIVEIRA, G. A. F.; SILVA, M. S.; OLIVEIRA, L. A.; AGUILAR-VILDOSO, C. I. Genome-wide selection in cassava. **Euphytica**, v. 187, p. 263-276, 2012.

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

ORTIZ, A. H. T.; ROCHA, V. P. C.; MOIANA, L. D.; GONÇALVES-VIDIGAL, M. C.; GALVÁN, M. Z.; VIDIGAL FILHO, P. S. Population structure and genetic diversity in sweet cassava cultivars from Paraná, Brazil. **Plant Molecular Biology Reporter**, v. 34, p. 1153-1166, 2016.

POMETTI, C. L.; BESSEGA, C. F.; SAIDMAN, B. O.; VILARDI, J. C. Analysis of genetic population structure in *Acacia caven* (*Leguminosae*, *Mimosoideae*), comparing one exploratory and two Bayesian-model-based methods. **Genetics and Molecular Biology**, v. 37, p. 64-72, 2014.

POOTAKHAM, W.; SHEARMAN, J. R.; RUANG-AREERATE, P.; SONTHIROD, C.; SANGSRAKRU, D.; JOMCHAI, N.; YOOCHA, T.; TRIWITAYAKORN, K.; TRAGOONRUNG, S.; TANGPHATSORNRUANG, S. Large-scale SNP discovery through RNA sequencing and SNP genotyping by targeted enrichment sequencing in cassava (*Manihot esculenta* Crantz). **Plos One,** v. 9, p. 1-19, 2014.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics,** v. 155, p. 945-959, 2000.

R Development Core Team (2017). R: A language and environment for statistical computing, reference index version 3.3.4. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0 http://www.R-project.org.

RABBI, I, Y.; KULAKOW, P. A.; MANU-ADUENING, J. A.; DANKYI, A. A.; ASIBUO, J. Y.; PARKES, E. Y.; ABDOULAYE, T.; GIRMA, G.; GEDIL, M. A.; RAMU, P.; REYES, B.; MAREDIA, M. K. Tracking crop varieties using

genotyping-by-sequencing markers: a case study using cassava (*Manihot esculenta* Crantz). **BioMedCentral Genetics**, v. 16, p. 1-11, 2015.

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, p. 329-342, 2012.

RABBI, I.; HAMBLIN, M.; GEDIL, M.; KULAKOW, P.; FERGUNSON, M.; IKPAN, A. S.; LY, D.; JANNINK, J-L. Genetic mapping using genotyping-by-sequencing in the clonally propagated cassava. **Crop Science**, v. 54, p. 1384-1396, 2014.

RAGHU, D.; SARASWATHI, T.; RAVEENDRAN, M.; GNANAM, R.; VENKATACHALAM, R.; SHANMUGASUNDARAM, P.; MOHAN, C. Morphological and simple sequence repeats (SSR) based finger printing of South Indian cassava germplasm. **International Journal of Integrative Biology**, v. 1, p. 141-148, 2007.

RAJI, A. A.; FAWOLE, I.; GEDIL, M.; DIXON, A. G. O. Genetic differentiation analysis of African cassava (*Manihot esculenta)* landraces and elite germplasm using amplified fragment length polymorphism and simple sequence repeat markers. **Annals Applied Biology**, v.155, p. 187–199, 2009.

RAMU, P.; ESUMA, W.; KAWUKI, R.; RABBI, I. Y.; EGESI, C.; BREDESON, J. V.; BART, R. S.; VERMA, J.; BUCKLER, E. S.; LU. F. Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. **Nature Genetics**, v. 49, p. 1-7, 2017.

REN, J.; SUN, D.; CHEN, L.; YOU, F. M.; WANG, J.; PENG, Y.; NEVO, E.; SUN, D.; LUO, M-C.; PENG, J. Genetic diversity revealed by single nucleotide polymorphism markers in a worldwide germplasm collection of durum wheat. **International Journal of Molecular Sciences**, v. 14, p. 7061-7088, 2013.

ROGOZIN, I. B.; PAVLOV, Y. I. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. **Mutation Research/Reviews in Mutation Research**, v. 544, p. 65-85, 2003.

ROJAS, C. M.; PÉREZ, J. C.; CEBALLOS, H.; BAENA, D.; MORANTE, N.; CALLE, F. Introduction of inbreeding and analysis of inbreeding depression in eight $S_1$ cassava families. **Crop Science**, v. 49, p. 543-548, 2009.

SEEB, J. E.; CARVALHO, G.; HAUSER, L.; NAISH, K.; ROBERTS, S.; SEEB, L. W. Single-Nucleotide Polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. **Molecular Ecology Resources**, v. 11, p. 1-8, 2011.

SEMAGN, K.; MAGOROKOSHO, C.; VIVEK, B. S.; MAKUMBI, D.; BEYENE, Y.; MUGO, S.; PRASANNA, B. M; WARBURTON, M. L. Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using Single-Nucleotide Polymorphism markers. **BioMedCentral Genomics**, v. 13, p. 1-13, 2012.

SINGH, N.; CHOUDHURY, D. R.; SINGH, A. K.; KUMAR, S.; SRINIVASAN, K.; TYAGI, R. K.; SINGH, N. K.; SINGH, R. Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. **Plos One**, v. 8, p. 1-14, 2013.

SONG, Q.; HYTEN, D. L.; JIA, G.; QUIGLEY, C. V.; FICKUS, E. W.; NELSON, R. L.; CREGAN, P. B. Fingerprinting soybean germplasm and its utility in genomic research. **G3: Genes, Genomes, Genetics**, v. 5, p. 1999-2006, 2015.

SPANIC, V.; KORZUN, V.; EBMEYER, E. Assessing genetic diversity of wheat genotypes from different origins by SNP markers. **Cereal Research Communications**, v. 43, p. 361-369, 2016.

SYVANEN, A. C. Accessing genetic variation: genotyping Single-Nucleotide Polymorphisms. **Nature Reviews Genetics**, v. 2, p.930–942, 2001.

TANG, W.; WU, T.; YE, J.; SUN, J.; JIANG, Y.; YU, J.; TANG, J.; CHEN, G.; WANG, C.; WAN, J. SNP based analysis of genetic diversity reveals important alleles associated with seed size in rice. **BioMedCentral Plant Biology Journal**, v. 16, p. 1-11, 2016.

TIAN, F.; BRADBURY, P. J.; BROWN, P. J.; HUNG, H.; SUN, Q.; FLINT-GARCIA, S.; ROCHEFORD, T. R.; McMULLEN, M. D.; OLLAND, J. B.; BUCKLER, E. S. Genome-wide association study of leaf architecture in the maize nested association mapping population. **Nature**, v. 43, p. 159-164, 2011.

TURYAGYENDA, L. F.; KIZITO, E. B.; FERGUSON, M. E.; BAGUMA, Y.; HARVEY, J. W.; GIBSON, P.; WANJALA, B. W.; OSIRU, D. S. O. Genetic diversity among farmer-preferred cassava landraces in Uganda. **African Crop Science Journal**, v. 20, p. 15-30, 2012.

VAN-HEERWAARDEN, J.; DOEBLEY, J.; BRIGGS, W. H.; GLAUBITZ, J. C.; GOODMAN, M. M.; GONZALEZ, J. J. ROSS-IBARRA, J. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. **Proceedings of the National Academy of Science of the United States of America (PNAS)**, v. 108, p. 1088-1092, 2011.

VIGOUROUX, Y.; GLAUBITZ, J. C.; MATSUOKA, Y.; GOODMAN, M. M.; SÁNCHEZ, J. G.; DOEBLEY, J. Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. **American Journal of Botany**, v. 95, p. 1240-1253, 2008.

VIGOUROUX, Y.; MITCHELL, S.; MATSUOKA, Y.; HAMBLIN, M.; KRESOVICH, S.; SMITH, J.S.; JAQUETH, J.; DOEBLEY, J. An analysis of genetic diversity across the maize genome using microsatellites. **Genetics**, v.169, p.1617-1630, 2005.

WILLEMEN, L.; SCHELDEMAN, X.; CABELLOS, V. S.; SALAZAR, S. R.; GUARINO, L. Spatial patterns of diversity and genetic erosion of traditional cassava (*Manihot esculenta* Crantz) in the Peruvian Amazon: An evaluation of socio-economic and environmental indicators. **Genetic Resources and Crop Evolution**, v. 54, p. 1599-1612, 2007.

WILLI, Y.; GRIFFIN, P.; BUSKIRK, J. V. Drift load in populations of small size and low density. **Heredity**, v. 110, p. 296-302, 2013.

WOLFE, M. D.; RABBI, I. Y., RABBI; EGESI, C.; HAMBLIN, M.; KAWUKI, R.; KULAKOW, P.; LOZANO, R.; CARPIO, D. P. D.; RAMU, P.; JANNINK, J-L. Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. **The Plant Genome**, v. 9, p. 1-13, 2016.

ZHOU, L.; VEGA, F. E.; TAN, H.; LIUCH, A. E. R.; MEINHARDT, L. W.; FANG, W.; MISCHKE, S.; IRISH, B.; ZHANG, D. Developing Single-Nucleotide Polymorphism (SNP) markers for the identification of coffee germplasm. **Tropical Plant Biology**, v. 9, p. 82-95, 2016.

ZINTZARAS, E. Impact of Hardy–Weinberg equilibrium deviation on allele-based risk effect of genetic association studies and meta-analysis. **European Journal of Epidemiology**, v. 25, p. 553-560, 2010.

Capítulo 2

# IDENTIFICAÇÃO DE DUPLICATAS DE *Manihot esculenta* Crantz COM BASE EM MARCADORES *SINGLE-NUCLEOTIDE POLYMORPHISM* (SNP)

## Identificação de duplicatas de *Manihot esculenta* Crantz com base em marcadores *Single-Nucleotide Polymorphism* (SNP)

**RESUMO:** A redundância genética presente no germoplasma de mandioca (*Manihot esculenta* Crantz) é um desafio para uma gestão eficiente dos recursos genéticos da espécie. Este trabalho teve como objetivo identificar e definir a estrutura genética de acessos duplicados no germoplasma de mandioca oriundos de diversas unidades de pesquisa da EMBRAPA – Brasil, com uso de marcadores *Single-Nucleotide Polymorphism* (SNP). Foram avaliados 2.371 acessos com 20.712 marcadores SNPs. A identificação de duplicatas foi realizada com base na identificação de genótipos multilocos (MLG), adotando limiar máximo de distância genética de 0,05. A estrutura populacional foi definida com base na análise discriminante de componentes principais (ADCP). Foram identificados 1.757 acessos únicos e 614 acessos duplicados. A redundância das coleções variou de 22,47% (Embrapa Amazônia Oriental) a 40,0% (Embrapa Semiárido), com média geral de 25,89%. Esta redundância entre diferentes unidades de pesquisa possivelmente deve-se ao compartilhamento histórico de acessos, bem como coletas realizadas em uma mesma região, ou mesmo ao intenso intercâmbio de germoplasma entre agricultores com troca de nomes dos genótipos. Em termos de estrutura genética, a manutenção de 250 componentes principais (CP) explicou 88% da variação genética dos marcadores SNPs e definiu a estrutura hierárquica do germoplasma duplicado de mandioca em 12 grupos. Todos os MLGs foram alocados dentro do mesmo grupo da ADCP, corroborando as análises de duplicatas e ainda revelando elevada variabilidade entre grupos, que foram bastantes distintos com base nas duas primeiras funções discriminantes. Nossos resultados trazem contribuições importantes para otimização da conservação dos recursos genéticos para o entendimento da diversidade e seu uso no melhoramento da cultura.

**Palavras chave**: *genotyping-by-sequencing*; mandioca; multilocos

## Identification of duplicates *Manihot esculenta* Crantz based on *Single-Nucleotide Polymorphism* (SNP)

**ABSTRACT:** The genetic redundancy present in cassava germplasm *(Manihot esculenta* Crantz) is a challenge for efficient management of the genetic resources of the species. This study aimed to identify and define the genetic structure of duplicates in cassava germplasm from various EMBRAPA research units, using Single-Nucleotide Polymorphism markers (SNP). We evaluated 2,371 accessions with 20,712 SNPs. The identification of duplicates was performed based on the identification of *multilocus genotypes* (MLG), adopting maximum genetic distance threshold of 0.05. Population structure was defined based on discriminant analysis of principal components (DAPC). A total of 1,757 unique and 614 duplicate accessions were identified. The redundancy of the collections ranged from 22.47% (Embrapa Amazônia Oriental) to 40.0% (Embrapa Semiárido), with average of 25.89%. This redundancy between different research units is probably due to the historical sharing of accessions, as well as collections carried out in the same region, or even to the intense germplasm exchange between farmers with different genotype names. In terms of genetic structure, the maintenance of 250 principal components explained 88% of the genetic variation of the SNP markers and defined the hierarchical structure of the duplicate cassava germplasm in 12 groups. All MLGs were allocated within the same DAPC group, corroborating duplicate analyzes and still revealing high variability between groups that were quite distinct based on the first two discriminant functions. Our results contribute to optimize the conservation of genetic resources, understanding diversity and its use in crop improvement.

**Keywords:** genotyping-by-sequencing; cassava; multilocus

## INTRODUÇÃO

A mandioca (*Manihot esculenta* Crantz) é nativa da América do Sul, tendo o Brasil como seu provável centro de origem e diversidade (OLSEN, 2004). Após muitos anos de cultivo e seleção de tipos de plantas, milhares de variedades locais foram selecionadas pelo homem com adaptações a diferentes regiões de cultivo, diferentes temperaturas e tipos de solo (EL-SHARKAWY, 2004). Por isso, a importância da conservação e uso destes recursos genéticos advém de uma série de atributos potencialmente úteis, adquiridos ao longo do processo evolutivo e de domesticação da espécie para qualidade de raízes (OLIVEIRA et al., 2015), resistência a doenças (VILAS-BOAS et al., 2016) e para o desenvolvimento de variedades comerciais, como características diferenciais de amido (VASCONCELOS et al., 2017).

Os Bancos Ativos de Germoplasma de Mandioca (BAG-Mandioca) da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) conservam uma enorme variedade de acessos desta espécie, com extrema relevância para uso futuro em diversos estudos genéticos e nos programas de melhoramento da cultura. Entretanto, ao mesmo tempo em que muitos acessos são conservados, outros são perdidos no campo por problemas climáticos e ataques de pragas e doenças. O problema disso, é que não se conhece precisamente o que se armazena e o que se perde, em função da inexistência de dados completos referentes à caracterização e avaliação agronômica destes acessos. Além disso, a falta de informações agronômicas completas, dificulta a identificação de acessos redundantes no BAG-Mandioca, que em muitos casos são observados pelos curadores e melhoristas por meio de dados de passaporte e observações de campo. Contudo, na maioria dos casos, quando se trabalha com grandes coleções, apenas uma pequena fração de acessos pode ser apontada como duplicada.

Para a mandioca, os custos de manutenção dos bancos de germoplasma são relativamente altos porque as plantas são mantidas no campo e ainda é necessário que haja uma cópia de segurança *in vitro*. De acordo com Shands (2010), o custo anual para manutenção e distribuição de acessos de mandioca (capital e custos variáveis) do CIAT (Centro Internacional de Agricultura Tropical) é da ordem de US$92,30 por acesso/ano. Portanto, medidas que visem a racionalização das coleções de germoplasma, no sentido

de reduzir o número de cópias, são de suma importância para otimizar o espaço físico de armazenamento dos acessos, seja em laboratório ou campo, bem como reduzir o custo de manutenção das coleções (TREUREN; HINTUM, 2003).

Atualmente em nível mundial, menos de 30% dos recursos vegetais mantidos em bancos de germoplasma são considerados distintos (FAO, 2010). Em muitos desses estudos, a identificação de duplicatas de acessos tem sido realizada com base em informações de passaporte, descritores morfoagronômicos ou mesmo com base em dados históricos (VAN HINTUM; KNÜPFFER, 1995; GREENE; PEDERSON, 1996; VAN HINTUM et al., 1996). Por outro lado, com os avanços da biologia molecular na cultura da mandioca, inúmeras oportunidades surgiram para aplicação dessas informações no aumento da eficiência do uso dos recursos genéticos e no melhoramento convencional. Neste momento, é preciso utilizar todas essas informações para a integração do conhecimento atual da genômica da mandioca, de forma a consolidar o uso de técnicas e estratégias avançadas de genética e biologia molecular no dia-a-dia do programa de melhoramento genético da cultura.

Diversos tipos de marcadores têm sido utilizados para os estudos moleculares na cultura da mandioca, a exemplo do *Restriction Fragment Length Polymorphism* (RFLP) e *Random Amplification of Polymorphic DNA* (RAPD) (FREGENE et al., 1997); microssatélites genômicos, derivados de EST (*Expressed Sequence Tag*) e genes candidatos a resistência a doenças (MBA et al., 2001; LOPEZ et al., 2007); *Simple-Sequence Repeats* (SSR) e *Amplified Fragment Length Polymorphism* (AFLP) (KUNKEAW et al., 2010); microssatélites genômicos, derivados de EST e SRAPs (*Sequence-Related Amplified Polymorphisms*) (CHEN et al., 2010) e marcadores do tipo Single-Nucleotide Polymorphism (SNP) (ORTIZ et al., 2016; MTUNGUJA et al., 2017; WANG et al., 2017). Os marcadores SNPs, tem sido atualmente os mais recomendados para os mais diversos tipos de estudos moleculares, bem como ferramentas auxiliares para análise de duplicatas em bancos de germoplasma, por apresentarem polimorfismo abundante e serem passíveis de automação resultando em alto rendimento analítico (MAMMADOV et al., 2012). Juntamente com os dados fenotípicos e de passaporte, os marcadores SNPs

podem contribuir em muito para uma distinção clara e eficiente do germoplasma de mandioca.

Recentemente, com a otimização dos processos de genotipagem de marcadores SNPs, a geração de informações sobre a variação genética em coleções de germoplasma, de forma a direcionar as estratégias de conservação e uso deste germoplasma, cresceu enormemente. Porém, estudos sobre o uso de marcadores moleculares SNPs na identificação de duplicatas de acessos são relativamente raros na literatura, embora existam relatos de casos de sucesso, a exemplo dos relatados no germoplasma de algumas culturas como, feijão caupi (EGBADZOR et al., 2014), soja (BANDILLO et al., 2015; SONG et al., 2015), kiwi (MELO et al., 2017) e em mandioca (RABBI et al., 2015).

O custo da identificação de um acesso de mandioca duplicado, utilizando a caracterização molecular é 12 vezes menor do que o custo de conservar e utilizar o material como um acesso diferente na coleção de germoplasma (HORNA et al., 2010). Portanto, estas informações moleculares devem ser utilizadas como informações auxiliares para maximizar a eficiência da conservação do germoplasma de mandioca, de modo que os curadores possam precisamente identificar acessos com *backgrounds* genéticos redundantes. Assim, o objetivo deste trabalho foi identificar a presença de duplicatas de acessos no BAG-Mandioca da EMBRAPA, bem como avaliar a estrutura genética das duplicatas de acessos com o uso de marcadores SNPs.

## MATERIAL E MÉTODOS

### Material Vegetal

Foram analisados 2.371 acessos pertencentes aos Bancos Ativos de Germoplasma (BAGs) conservados na EMBRAPA, onde 1.553 acessos pertencem à Embrapa Mandioca e Fruticultura, Cruz das Almas – BA (CNPMF); 356 acessos à Embrapa Amazônia Oriental, Belém – PA (CPATU); 327 acessos à Embrapa Cerrados, Brasília – DF (CPAC) e 135 acessos à Embrapa Semiárido, Petrolina - PE (CPATSA).

### Extração de DNA

O DNA foi extraído a partir de folhas jovens, de acordo com o protocolo CTAB (brometo de cetiltrimetilamônio) conforme descrito por Doyle e Doyle

(1987), com adição de polivilpirrolidona (PVP) e aumento da concentração de 2-mercaptoetanol a 0,4%. A qualidade do DNA foi avaliada por quantificação em gel de agarose 1,0% (p/v) corado com brometo de etídio (1,0mg/L) em tampão TBE 0,5 x (45 mM Tris-borate, 1 mM EDTA e q.s.p de água destilada), visualizado em luz UV e registrado com o fotodocumentador Gel Logic 212 Pro (Carestream Molecular Imaging, New Haven, USA) por comparação visual com uma série de concentrações de DNA conhecido do fago Lambda (Invitrogen, Carlsbad, CA). O DNA foi diluído em tampão TE (Tris-HCl 10mM e EDTA 1mM) para uma concentração final de 60 ng/µL e a qualidade verificada pela digestão de 250 ng do DNA genômico a partir de 10 amostras aleatórias com a enzima de restrição *Eco*RI (New England Biolabs, Boston, EUA) a 65° C durante duas horas e posteriormente, visualizada em gel de agarose.

**Genotipagem por Sequenciamento**

As amostras de DNA foram genotipadas no *Genomic Diversity Facility* pertencente à *Cornell University* (http://www.biotech.cornell.edu/brc/genomic-diversity-facility). O protocolo básico da *genotyping-by-sequencing* (GBS), foi descrito por Elshire et al. (2011), no qual o DNA foi digerido pela enzima *Ape*KI recomendada por (HAMBLIN; RABBI, 2014), uma endonuclease de restrição tipo II que reconhece uma sequência degenerada de 5 bases (GCWGC, onde W é A ou T) com comprimentos de 100 pb. A ligação entre os fragmentos com corte *Ape*KI e o adaptador, foi realizada após a digestão das amostras e implementação de sistema multiplex com 192 amostras para realização do sequenciamento. A GBS foi realizada utilizando o *Genome Analyzer 2000* (Illumina, Inc., San Diego, CA). Para análise das sequências e filtros de qualidade, foi utilizado o software Tassel versão 5.2.37 (BRADBURY et al., 2007), visando remover alelos com uma frequência alélica mínima (MAF) inferior a 0,05, e SNPs com mais de 20% de dados perdidos, totalizando, ao final 20.712 SNPs.

**Identificação de acessos duplicados**

A matriz de distância genética de Hamming foi calculada pela função *bitwise.dist* do pacote *poppr* do software R versão 3.3.4 (R Development Core Team, 2017), bem como o número de diferenças alélicas entre dois acessos, contando os dados ausentes como equivalentes por comparação. A identificação de duplicatas foi feita com base na detecção de genótipos

multilocos (MLGs = *multilocus genotypes*). Este algoritmo de agrupamento foi implementado sobre a matriz de distância genética com base no método do "*vizinho mais próximo",* que agrupa acessos que estejam a uma distância mínima definida (*threshold* = 0,05). O limiar de distância crítico para declarar que dois acessos são idênticos foi determinado empiricamente a partir da distribuição da distância genética calculada entre os acessos, levando em consideração a existência de possíveis erros de genotipagem nos dados genéticos em estudo. Assumindo que o conjunto de dados possuem distâncias variáveis entre acessos, o limiar foi definido como o ponto máximo de dissimilaridade genética entre indivíduos para declará-los idênticos, isto é, foram considerados semelhantes os indivíduos que apresentarem uma distância genética inferior ao limite dado de 0,05.

Os MLGs foram identificados com base na função *mlg.filter*, disponível no pacote computacional *poppr* versão 2.3.0 do software *R* versão 3.3.4 (R Development Core Team, 2017). Além de definir os MLGs, a função *mlg.filter*, indica quais genótipos compartilham do mesmo MLG, os quais apresentam perfil genético semelhante.

O método utilizado foi especificamente desenvolvido para análise de populações clonais, implementado com o objetivo de visualizar relacionamentos entre MLGs desconhecidos (KAMVAR et al., 2014). No entanto, estudos em populações clonais, baseados somente na identificação e comparação de MLGs entre indivíduos, pode não ser preciso, pois a existência de mutações somáticas e possíveis erros na identificação dos marcadores genéticos devem ser levados em consideração. Com a geração de grande quantidade de dados via *next-generation sequencing* (NGS), a resolução genética tem aumentado bastante, embora a possibilidade de erros de genotipagem e elevado número de dados perdidos também sejam frequentes (ELSHIRE et al., 2011). Entretanto, a função *mlg.filter* considera estes possíveis viés, e permite a escolha da distância genética e a abordagem para agrupar os MLGs comuns identificados em mais de um indivíduo que, biologicamente, seja mais relevante para a população em estudo, considerando nível de ploidia e natureza dos marcadores de DNA utilizados (KAMVAR et al., 2015).

**Agrupamentos dos acessos duplicados**

A análise discriminante de componentes principais (ADCP) disponível no pacote computacional *adegenet*, do programa R versão 3.3.4 (R Development Core Team 2017), foi utilizada para definição dos agrupamentos dos acessos duplicados de mandioca, pois esta técnica não requer uma definição a priori de grupos genéticos (JOMBART et al., 2010).

A ADCP baseia-se na transformação preliminar dos dados, utilizando a análise de componentes principais (ACP) como passo prévio à análise discriminante (AD), garantindo que as variáveis submetidas à AD são perfeitamente não correlacionadas e seu número seja menor que os indivíduos analisados, sem necessariamente implicar na perda de informação genética. Essa transformação permite que a AD seja aplicada a qualquer dado genético, ao passo que, a análise minimiza as diferenças entre indivíduos dentro dos grupos e as maximiza entre grupos, no qual os acessos são melhores discriminados em grupos pré-definidos (JOMBART et al., 2010).

Foram utilizados sucessivos agrupamentos com o método *K-means* e o Critério de Informação Bayesiano (BIC) para definição do número ideal de grupos, em que atribui-se *K* com menor valor BIC para representar o mais provável número de grupos para o conjunto de dados em análise. Contudo, na presença de estruturação genética de forma hierárquica, os valores de BIC podem ser reduzidos após a identificação do verdadeiro valor de *K*. Assim, a redução dos valores BIC foi analisada empiricamente para identificar o valor de *K* na qual os valores BIC reduziram apenas sutilmente (JOMBART et al., 2010). Foram testados valores de *K* de 1 a 80. Após definido o número de grupos, foram retidos os eixos da análise de componentes principais que explicam mais de 80% da variância total dos dados.

**RESULTADOS**

**Identificação de acessos duplicados com base na análise multilocos**

Dos 2.371 acessos de mandioca analisados, 1.757 apresentaram MLGs únicos, enquanto os outros 614 acessos apresentaram um perfil não exclusivo de marcadores SNPs, sendo assumidos como duplicatas, pois cada MLG corresponde a um único genótipo (ARNAUD-HAOND et al., 2007). Os 25,89% do total dos acessos que expressaram perfis genéticos idênticos representaram

614 genótipos diferentes, fazendo com que o número de distintos perfis de SNPs seja 1.757.

Foram identificados 54, 80, 84, 396, acessos duplicados no CPATSA, CPATU, CPAC e CNPMF, respectivamente. Em termos percentuais o CPATSA apresentou o maior número de acessos duplicados (40%) com perfis geneticamente semelhantes com base no polimorfismo dos SNPs. Nas demais unidades da EMBRAPA, a percentagem de duplicatas variou de 22,47% a 25,69% dos acessos de mandioca analisados (Figura 1).

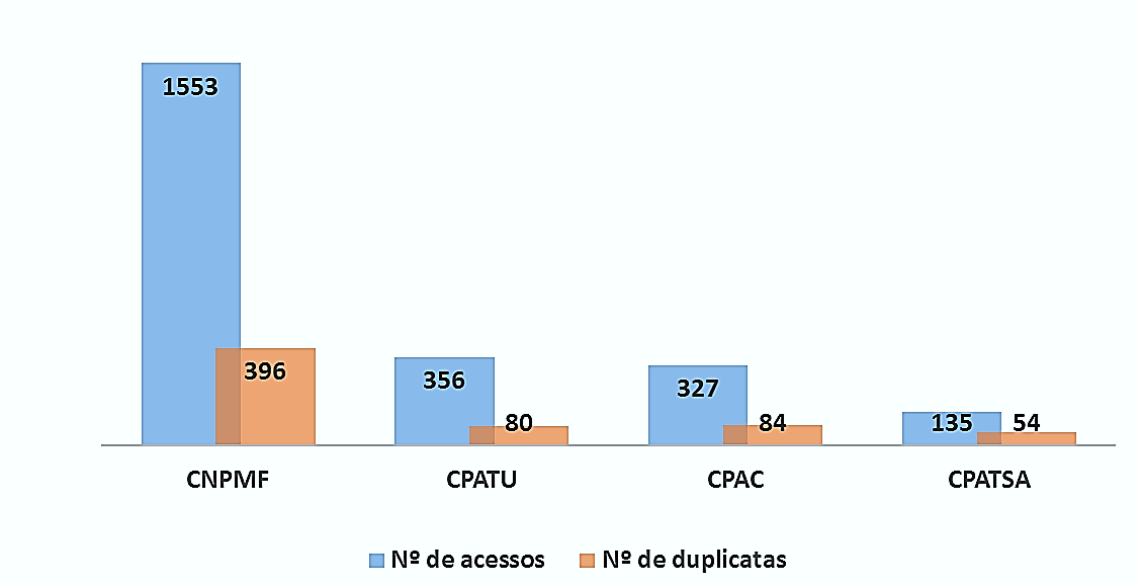


**Figura 1**. Relação entre o número de acessos conservados nos BAGs e a quantidade de genótipos que apresentam perfis genéticos semelhantes, considerados como duplicatas.

Foi possível identificar acessos com perfis genéticos semelhantes entre e dentro das diferentes coleções de mandioca da EMBRAPA. Especificamente no caso do CPATSA, todos os 135 acessos são cópias do germoplasma do CNPMF, por possuírem os mesmos MLGs (Tabela 1). Possivelmente isto ocorreu devido ao compartilhamento histórico de acessos entre essas instituições, bem como pelo fato de que muitas coletas de germoplasma na região semiárida do Brasil tem sido realizada nas regiões de abrangência de atuação do CPATSA. O CNPMF também possui em sua coleção, 87 e 82 acessos, mantidos pelo CPAC e CPATU, respectivamente, enquanto que o CPATU e CPAC apresentaram 31 acessos duplicados conservados por ambas unidades e entre CPAC e CPATSA, onde foram

encontrados somente 10 acessos duplicados. Entretanto, entre o CPATU e CPATSA não foram identificadas duplicatas mantidas por essas unidades de pesquisa (Tabela 1). Da mesma forma que no caso do CPATSA, o histórico de troca de germoplasma no Brasil é elevado, embora em muitas situações a troca de nomes, ou mesmo a falta de padronização dos dados de passaporte, podem levar a dúvidas sobre a origem dos acessos coletados. Neste caso, a análise molecular pode resolver dúvidas sobre a identidade genética dos acessos com alta confiabilidade.

**Tabela 1**. Número de acessos duplicados e multilocos (MLG) dentro e entre coleções de mandioca da EMBRAPA, identificados com base na análise partir de 20.712 marcadores SNPs.

| Unidade | CNPMF | | CPATU | | CPAC | | CPATSA | |
|---|---|---|---|---|---|---|---|---|
| | Nº duplicatas | Nº MLG | Nº duplicatas | Nº MLG | Nº duplicatas | Nº MLG | Nº duplicatas | Nº MLG |
| CNPMF | 396* | 86** | 82* | 19** | 87* | 20** | 135* | 33** |
| CPATU | - | - | 80* | 19** | 31* | 7** | 0* | 0** |
| CPAC | - | - | - | - | 84* | 17** | 10* | 2** |
| CPATSA | - | - | - | - | - | - | 54* | 2** |

* Número de duplicatas dentro de cada banco e **Número de multilocos compartilhados dentro e entre bancos. CNPMF (Embrapa Mandioca e Fruticultura); CPATU (Embrapa Amazônia Oriental); CPAC (Embrapa Cerrados) e CPATSA (Embrapa Semiárido).

Com base na análise de duplicatas, foram identificados 124 MLGs dentro dos BAGs da EMBRAPA (86, 19, 17 e 2 MLGs formados por acessos pertencentes ao CNPMF, CPATU, CPAC e CPATSA, respectivamente) (Tabela 1). Além disso, houveram acessos que compartilharam MLGs entre diferentes BAGs, sendo 33 entre CNPMF e CPATSA, 20 entre CNPMF e CPAC, 19 entre CNPMF e CPATU, sete entre CPAC e CPATSA e dois entre CPAC e CPATSA. Portanto, o maior número de MLGs entre e dentro dos BAGs foi observado no CNPMF, que tem sido um repositório de acessos para os demais bancos de germoplasma. Entretanto, observa-se nenhum compartilhamento de germoplasma entre o CPATU e CPATSA. Possivelmente isso ocorreu pela maior adversidade climática entre essas regiões, considerando que o CPATU se localiza na região Norte do Brasil (Belém-PA) cujas principais características climáticas são médias anuais de temperatura de 26,7ºC, umidade relativa 84%,

e precipitação pluviométrica de cerca de 3.000 mm (BASTOS et al., 2002), enquanto que o CPATSA se localiza na região Nordeste do Brasil (Petrolina-PE) com temperatura média anual de 26,1ºC, umidade relativa de 60%, e precipitação pluviométrica média anual de 530 mm (TEIXEIRA, 2010). Portanto, a disponibilidade de água é um dos principais fatores de diferenciação entre estas regiões do país, fazendo com que acessos provenientes da região Norte não se adaptem ao Nordeste e vice-versa.

A diversidade genética dos 2.371 acessos de mandioca, calculada como base na distância genética de Hamming, variou de 0,014 a 0,297, com média de 0,231 (Figura 2). Entretanto, a maior parte dos acessos de mandioca divergiu a uma distância maior que 0,20, indicando grande diversidade genética armazenada no germoplasma de mandioca do Brasil. Considerando o limiar máximo de 0,05 de distância genética para definição dos diferentes MLGs observou-se que os acessos BGM2004 e BGM1635 apresentaram a menor distância genética igual a 0,014, enquanto que nos demais acessos duplicados esta distância variou de 0,021 a 0,050, com média de 0,041.

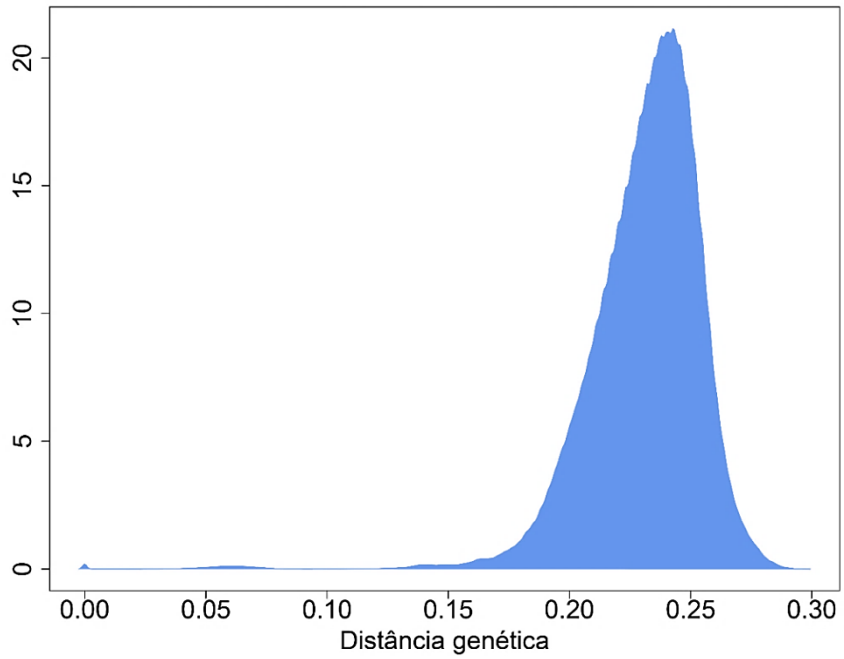


**Figura 2**. Dissimilaridade genética entre 2.371 acessos de mandioca com base em 20.712 marcadores *Single-Nucleotide Polymorphism* (SNP).

**Estrutura populacional dos acessos duplicados**

A ADCP foi utilizada para investigar o padrão de diversidade genética dos acessos de mandioca considerados duplicados. A manutenção de 250 componentes principais (CP) na etapa preliminar de transformação dos dados permitiu que a ADCP explicasse mais de 88% da variação genética total observada pelos marcadores SNPs. A definição do número de CP a serem retidos na análise é um ponto de discussão sobre o poder de redução na dimensionalidade dos dados. De modo geral, é recomendado o uso de componentes que retenham mais de 80% da variância genética (JOMBART et al., 2010). No contexto da ADCP é preciso definir um ponto de equilíbrio entre o poder de discriminação dos agrupamentos e a estabilidade de atribuições dos genótipos em cada grupo. Assim, a análise com 250 CP garantiu alto poder estatístico para se avaliar a estrutura genética dos acessos duplicados conservados nos quatro principais BAGs de mandioca do país.

A Figura 3 indica a presença de estrutura hierárquica no germoplasma duplicado de mandioca e ilustra a definição do número ideal de grupos. Com base no agrupamento *K-means*, a maior queda inicial do critério de informação bayesiano (BIC) ocorreu com $K$ = 12, sendo este o número de grupos selecionados para representar a diversidade do germoplasma duplicado de mandioca.

A probabilidade de alocação de cada acesso em determinado grupo pela ADCP foi 100% para todos os 614 acessos duplicados, não havendo, portanto, possibilidade de compartilhamento genômico entre os diferentes grupos da ADCP. Assim, o agrupamento dos acessos duplicados de mandioca, com base nas duas primeiras funções discriminantes (LD) revelou uma clara separação dos 12 grupos de diversidade (Figura 4). A variação no número de acessos por grupos foi de 12 (Grupo 6) a 129 (Grupo 11), com média de 51 acessos por grupo. Por outro lado, o número de MLGs variou de 4 (Grupo 4) a 51 (Grupo 11), com média de 16 MLGs por grupo formado pela ADCP (Tabela 2). Portanto, os acessos do Grupo 11 ainda armazenam uma ampla diversidade genética dentro do grupo.

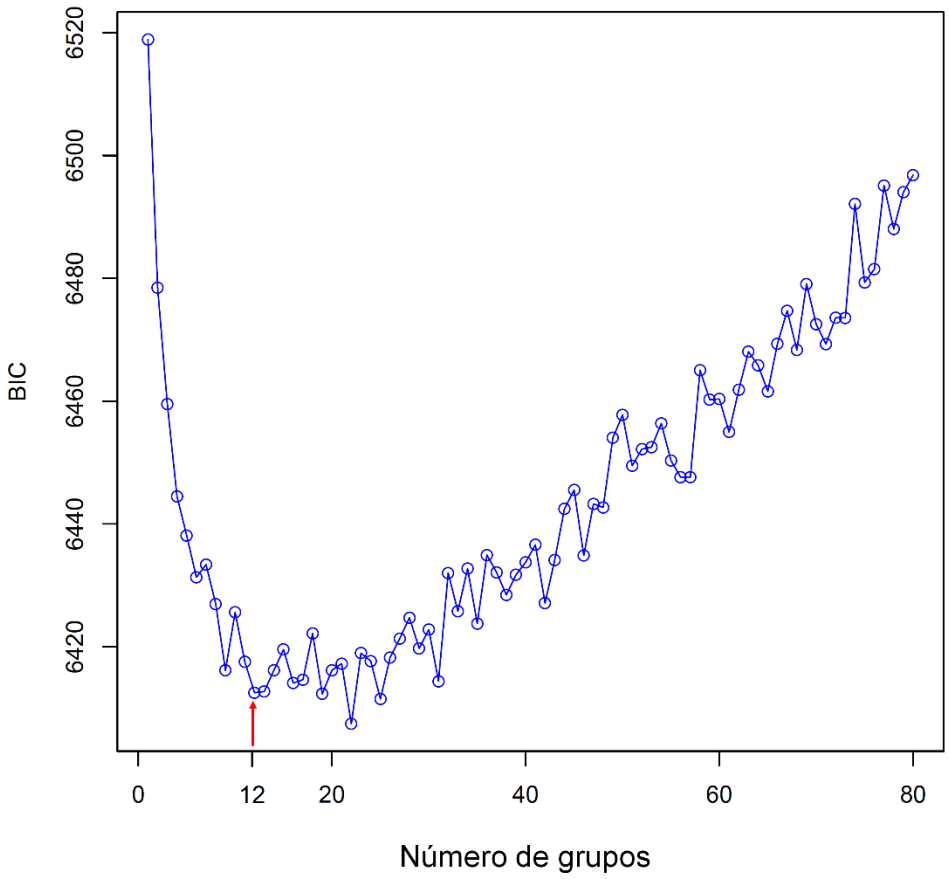


**Figura 3.** Distribuição dos valores do critério de informação bayesiano (BIC) em função do número de grupos, com base na análise de 614 acessos duplicados de mandioca com 20.712 marcadores Single-Nucleotide Polymorphism (SNP). A seta em vermelho indica o número de grupos escolhido para a análise de agrupamento.

Em função do grande intercâmbio e coleta dos mesmos acessos em uma mesma região por diferentes unidades de pesquisa da EMBRAPA, os agrupamentos formados pela ADCP não foram formados por acessos pertencentes exclusivamente a uma única unidade de pesquisa (Tabela 2). De modo geral, os grupos da ADCP foram compostos por acessos pertencentes a duas (Grupos 3 e 6), três (Grupos 2, 4, 5, 7, 8, 10 e 11) ou quatro unidades da Embrapa (Grupos 1, 9 e 12). Todos os MLGs foram alocados dentro de um mesmo grupo da ADCP, o que demonstra a concordância no agrupamento populacional dos acessos de mandioca com base em duas abordagens genéticas completamente diferentes (ou seja, análise MLG e ADCP). Esta

informação reflete a elevada acurácia na definição da similaridade das duplicadas de acessos identificadas no germoplasma de mandioca.

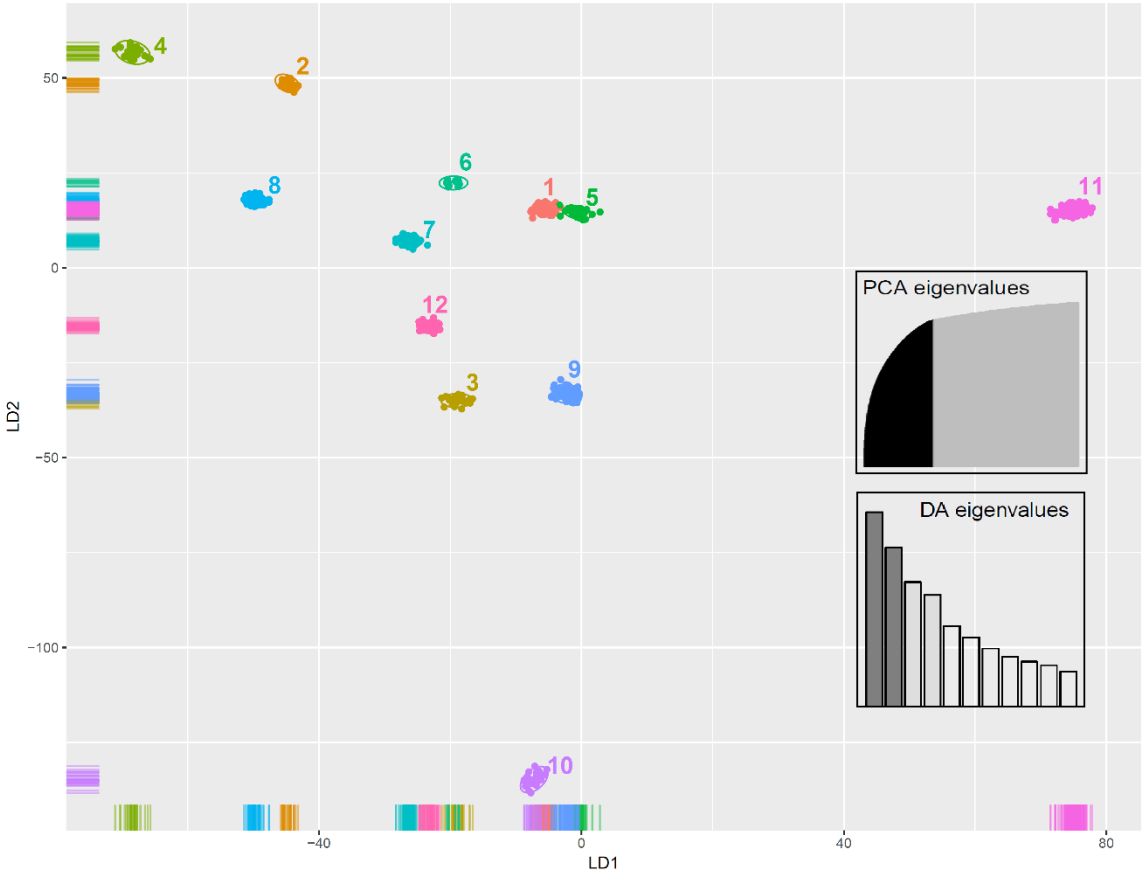


**Figura 4.** Dispersão do primeiro de segundo componente da análise discriminante de componentes principais (ADCP) com base na análise de 614 acessos de mandioca considerados duplicados, com uso de 20.712 marcadores *Single-Nucleotide Polymorphism* (SNP). Os grupos estão representados por cores de acordo com a legenda. O gráfico acima à direita representa a contribuição dos autovalores (*eigenvalues*) dos componentes principais selecionados na ADCP, enquanto o gráfico abaixo à direita indica a variância explicada pelos autovalores das duas funções discriminantes utilizadas no *scatterplot*.

A distância média dentro do agrupamento da ADCP variou de 0,037 a 0,045 nos Grupos 1 e 12, respectivamente. Por outro lado, a maior variação ocorreu nos grupos 6 (0,021 a 0,050) e 10 (0,014 a 0,050) (Tabela 2). Isto indica que a máxima variabilidade genética foi efetivamente dividida entre os diversos grupos da ADCP, considerando que o princípio da análise

discriminante (AD), apresenta como diferencial, uma melhor separação e visualização dos grupos, enquanto busca resumir toda diferenciação genética minimizando entre indivíduos e maximizando entre grupos.

**Tabela 2**. Número de acessos e genótipos multilocos de mandioca (MLG) pertencentes aos 12 grupos de diversidade genética identificados pela análise discriminante de componentes principais (ADCP) com base na análise partir de 20.712 marcadores *Single-Nucleotide Polymorphism* (SNP).

| Grupo ADCP | Nº de acessos | CNPMF | CPATSA | CPATU | CPAC | Nº MLG | Variação da distância genética | Distância genética média |
|---|---|---|---|---|---|---|---|---|
| 1 | 95 | 68 | 16 | 9 | 2 | 31 | 0,025 a 0,049 | 0,037 |
| 2 | 25 | 20 | - | 4 | 1 | 8 | 0,033 a 0,049 | 0,042 |
| 3 | 31 | 25 | 6 | - | - | 5 | 0,024 a 0,050 | 0,041 |
| 4 | 27 | 16 | 10 | - | 1 | 4 | 0,025 a 0,050 | 0,042 |
| 5 | 41 | 36 | 3 | - | 2 | 11 | 0,025 a 0,050 | 0,044 |
| 6 | 12 | 7 | 5 | - | - | 5 | 0,021 a 0,050 | 0,038 |
| 7 | 60 | 32 | - | 2 | 26 | 23 | 0,026 a 0,050 | 0,043 |
| 8 | 42 | 25 | 7 | - | 10 | 14 | 0,029 a 0,050 | 0,044 |
| 9 | 68 | 45 | 3 | 6 | 14 | 17 | 0,023 a 0,050 | 0,043 |
| 10 | 30 | 24 | 1 | 5 | - | 5 | 0,014 a 0,050 | 0,042 |
| 11 | 129 | 78 | - | 44 | 7 | 51 | 0,024 a 0,050 | 0,043 |
| 12 | 54 | 20 | 3 | 10 | 21 | 18 | 0,034 a 0,050 | 0,045 |

CNPMF (Embrapa Mandioca e Fruticultura); CPATSA (Embrapa Semiárido); CPATU (Embrapa Amazônia Oriental); CPAC (Embrapa Cerrados) e MLG (Genótipo multilocos).

## DISCUSSÃO

### Identificação de acessos duplicados com base em genótipos multilocos

Historicamente, a movimentação de germoplasma entre a Amazônia e a região central do Brasil, permitiu a domesticação da espécie e a formação de diversas variedades durante as migrações dos povos nativos, permitindo a hibridação entre as cultivares e parentes selvagens presentes, sobretudo na região do Brasil Central (NASSAR, 2002). Embora hibridações ainda ocorram na natureza e de forma controlada nos programas de melhoramento, o germoplasma de mandioca é quase exclusivamente mantido por propagação vegetativa, em condições de campo e *in vitro*. Além disso, nos dias atuais, ainda ocorre intensa troca de material propagativo entre diferentes regiões, via hastes ou manivas, que frequentemente apresenta problemas de registro e

rotulagem dos acessos que nem sempre são padronizadas, fazendo com que as entradas tenham informações limitadas sobre sua correta identidade. Além disso, os agricultores tendem a adotar diferentes nomes aos acessos intercambiados. Portanto, na formação dos bancos de germoplasma de mandioca, certamente foram coletados distintos acessos com o mesmo nome e diferentes acessos com o mesmo nome. Esta é uma realidade que acomete a maioria dos bancos de germoplasma de espécies vegetais, e que precisa ser resolvido.

Considerando o Brasil como centro de origem e diversidade da mandioca (OLSEN, 2004), a EMBRAPA tem procurado manter uma coleção diversificada de recursos genéticos de *M. esculenta* Crantz para uso imediato e futuro. A manutenção efetiva da diversidade genética do germoplasma de mandioca é particularmente importante no contexto da manutenção da agricultura moderna, que visa garantir elevados patamares de produtividade de forma competitiva, pela busca de melhores variedades e com características diferenciais àquelas usadas pelos agricultores. Portanto, a introdução de diversidade genética adicional, por meio de polinização cruzada com variedades locais e mesmo parentes selvagens, deve ser feita tendo como material base o germoplasma da espécie, que é uma fonte importante de nova variação. Contudo, o custo da conservação dos acessos em condições de campo e *in vitro*, são extremamente elevados, o que impõe limitações na manutenção e expansão do germoplasma. Com isso, é fundamental que a incorporação de novos acessos seja feita de forma estratégica para capturar a diversidade genética adicional, que efetivamente possui valor potencial para os usuários.

A presença de duplicatas de acessos certamente contribui para o aumento no custo da preservação do germoplasma de mandioca. Com isso, os curadores precisam utilizar diferentes ferramentas para identificar e reduzir a duplicação de acessos. A análise do perfil molecular dos 2.371 acessos de mandioca pertencentes às diferentes coleções da EMBRAPA revelou a presença de mais de 25% de duplicatas nos acessos destas coleções que foram coletados ao longo de mais de 40 anos no Brasil (prioritariamente). A variação na percentagem de duplicatas entre os diferentes BAGs (22,47% - CPATU a 40,00% - CPATSA), provavelmente se deve à prática de intercâmbio

entre o germoplasma nacional, bem como pela coleta de acessos em uma mesma zona agroecológica pelas diferentes coleções da EMBRAPA. Ao mesmo tempo em que estas redundâncias precisam ser reduzidas entre as diferentes coleções de mandioca, é preciso incentivar a preservação dos acessos de mandioca em zonas agroecológicas que são bastante diversas, a exemplo daquelas situadas nas regiões Norte, Nordeste e Sul do Brasil; sobretudo para garantir a sobrevivência de acessos com adaptação específica a estas regiões.

Em outros estudos com mandioca, marcadores morfológicos e isoenzimáticos, foram inicialmente utilizados para identificação de acessos com alta similaridade na coleção do CIAT, enquanto que a confirmação efetiva de duplicatas de acessos (81 em 744 acessos) foi realizada com uso de marcadores RFLP (JIMENEZ-NIETO, 1994). Outros trabalhos na coleção do CIAT indicaram valores de redundância variando de 20,0 a 25,0% da coleção (OCAMPO et al., 1993; OCAMPO et al., 1995). Chavarriaga-Aguirre et al. (1999) utilizaram marcadores isoenzimáticos, microssatélites e AFLP, juntamente com sete descritores morfológicos para identificação de duplicatas em uma coleção nuclear composta por 521 acessos de mandioca, cujos resultados demonstraram haver pelo menos 1,34% de duplicatas nesta coleção nuclear.

Embora haja diferenças entre o germoplasma armazenado no CIAT em relação ao germoplasma do Brasil, bem como ao uso de diferentes estratégias e técnicas para detecção de duplicatas, verificou-se valores semelhantes de acessos redundantes (em torno de 25%). Independente das abordagens, a redução dos acessos duplicados nos bancos de germoplasma e a comparação das novas entradas de acessos com aqueles já existentes na coleção para evitar aumento da redundância das coleções, são estratégias que devem ser implementadas na rotina dos bancos de germoplasma de mandioca.

Para Gross et al. (2012) a identificação de amostras geneticamente idênticas com elevada precisão exige que os marcadores selecionados sejam suficientemente variáveis e que levem em consideração a realidade da estrutura populacional e a relação entre indivíduos. Além disso, Arnaud-Haond et al. (2007) utilizaram o procedimento Monte Carlo para assegurar que um conjunto de marcadores microssatélites que tivessem poder discriminatório

suficiente para identificar todos os MLGs presentes na amostragem de 220 genótipos da erva marinha *Cymodocea nodosa*. Estes autores relataram que apenas sete locos microssatélites foram suficientes para determinar o número de MLGs presentes na amostra com elevada confiabilidade. Além do número, um dos problemas mais comuns que afetam as estimativas de diversidade e relacionamento entre indivíduos é o polimorfismo limitado dos marcadores utilizados. Isso impede a discriminação precisa de distintos acessos que podem ser considerados idênticos, com base no conjunto de marcadores usados, levando à superestimação da ocorrência de duplicatas. Portanto, em termos numéricos e de polimorfismo, é esperado que os 20.712 SNPs utilizados no presente trabalho sejam bastante adequados para determinar os diferentes MLGs com base no perfil genético dos 2.371 acessos de mandioca.

A análise de duplicatas no germoplasma de mandioca, considerando os diferentes MLGs, foi feita com base na função *mlg.filter* do pacote poppr. De acordo com Kamvar et al. (2014) a definição dos MLGs foi desenvolvida para populações clonais e/ou parcialmente clonais, sob um critério rigoroso que leva em consideração possíveis erros de genotipagem, principalmente para marcadores identificados pelas plataformas *Next-Generation Sequencing* (NGS). A função é aplicada a indivíduos diploides e a marcadores SNPs, não levando em consideração dados perdidos, também permitindo agrupar os acessos com base na escolha do agrupamento que melhor se adequam a natureza dos dados, considerando um limiar mínimo para distinguir os acessos e identificar as duplicatas (KAMVAR et al., 2015).

No processo de identificação de duplicatas, também é preciso considerar a superestimação do número de genótipos diferentes em função da presença de múltiplos MLGs pertencentes ao mesmo genótipo, cuja ocorrência se deve à existência de mutações somáticas ou erros de genotipagem (DOUHOVNIKOFF; DODD, 2003). Portanto, de acordo com Arnaud-Haond et al. (2007) a adoção de um limiar de distância genética pode ser definida, abaixo da qual a hipótese de que MLGs distintos pertencem ao mesmo genótipo, não deva ser rejeitada. Porém, alguns autores mencionaram que a quantidade de diversidade genética aceitável entre acessos geneticamente semelhantes ainda não está bem definida (LUND et al., 2003). De fato, Fu et al. (2006) estudaram a variação genética entre seis acessos de trigo tipo Marquis

e seis de cevada tipo Thorpe canadense com uso de marcadores AFLP para desenvolver um valor limite para declaração de duplicidade dos acessos. Esses autores observaram que as verdadeiras duplicatas das duas espécies apresentaram variação de até 5% nos fragmentos de AFLP. Portanto, a definição dos MLGs no germoplasma de mandioca com base na distância genética máxima de 0,05 entre acessos, parece ser bastante realista considerando que a automação da genotipagem dos SNPs por GBS pode reduzir ainda mais a taxa de erro na definição do perfil genético de cada acesso. De fato, de acordo com Zhou et al. (2015), a natureza bialélica dos SNPs facilita a redução na taxa de erro na genotipagem dos acessos, além de promover a compatibilidade dos resultados entre laboratórios. Além disso, a possibilidade de se identificar duplicatas com alta confiabilidade, acurácia e automação em conjunto informações com dados perdidos e erros de genotipagem, em quantidades limitadas, são vantagens importantes da análise.

**Estrutura populacional dos acessos duplicados de mandioca**

O uso de marcadores moleculares de alta cobertura genômica como os SNPs, é relativamente recente nos programas de melhoramento genético da mandioca. Além disso, SNPs para identificação de duplicatas de acessos têm sido pouco explorados, por isso este é um dos primeiros estudos sobre o entendimento da estrutura populacional em acessos de mandioca com elevada redundância genética. Com base na ADCP nos 124 MLGs foram identificados 12 grupos de diversidade. Houve uma forte correlação do agrupamento com os MLGs, pois a ADCP procedeu o agrupamento dos 124 MLGs em 12 diferentes grupos, onde todos os indivíduos alocados nos agrupamentos MLGs também permaneceram nos mesmos grupos formados pela ADCP.

A ADCP revelou uma clara separação entre os 12 diferentes grupos de acessos duplicados, considerando a sensibilidade desta técnica na detecção de subestrutura em modelos hierárquicos (JOMBART et al., 2010). A estrutura da população pode ser influenciada pela presença de alelos exclusivos de determinados grupos que possui influência desproporcional na ADCP. Com isso a maioria dos grupos são claramente separados dos demais com uso apenas das duas primeiras funções discriminantes, a exemplo do que ocorreu no agrupamento dos acessos de mandioca duplicados (Figura 4). Exceção

ocorreu apenas nos Grupos 1 e 5, que apresentaram ligeira sobreposição de alguns acessos.

Segundo Jombart et al. (2010), a ADCP pode ser útil para uma grande variedade de organismos, independentemente da sua ploidia e taxa de recombinação genética, pois a metodologia é independente de qualquer modelo de genética de populações sendo, portanto, livre de suposições sobre equilíbrio de Hardy-Weinberg ou desequilíbrio de ligação. Além disso, esta análise pode ser aplicada a grandes conjuntos de dados exigindo demanda computacional inferior à abordagem bayesiana.

A existência de elevados níveis de diversidade constitui um fator importante na interpretação da diferenciação dos agrupamentos e definição de contrastes genéticos dentro de determinados grupos de duplicatas. Portanto, estes 12 grupos formados pela ADCP podem não somente orientar as estratégias de conservação do germoplasma de mandioca, mas também contribuir para definição de grupos genéticos que possam ser cruzados para geração de populações segregantes no âmbito dos programas de melhoramento genético da espécie.

A capacidade de identificar populações de melhoramento com o elevado desempenho agronômico é efetivamente um objetivo no desenvolvimento de novas variedades de mandioca (OLIVEIRA et al., 2015). Além disso, de acordo com Semagn et al. (2012) é reconhecido que grupos de diversidade próximos tendem a aumentar a redundância em diversos programas de melhoramento e, portanto, seu uso em cruzamentos, pode resultar em desperdício de recursos, pois o cruzamento entre parentais com baixa complementariedade genética pode gerar progênies de baixo desempenho (DIAS et al., 2013; BENIN et al., 2012). Por outro lado, o cruzamento de parentais geneticamente divergentes, pode resultar em elevada variação fenotípica nas progênies. Assim, a estrutura populacional e a classificação dos acessos de mandioca em diferentes grupos com base nos marcadores SNPs podem encorajar os melhoristas a planejarem melhor seus cruzamentos para maximizar as chances de obtenção de progênies com elevado contraste fenotípicos para diversas características agronômicas.

**Perspectivas para otimização e uso do germoplasma de mandioca**

De acordo com a FAO (2010), dos 7,4 milhões de acessos armazenados em 1.700 bancos de germoplasma, estima-se que entre 70 a 75% sejam duplicatas com base em dados genotípicos e de passaporte. Além disso, a manutenção do germoplasma de mandioca no campo tem um alto custo, cerca de US$92,30 anuais por acesso de acordo com Shands (2010). Por isso, é preciso investir em métodos adequados de identificação inequívoca de duplicatas para garantir a manutenção contínua dos recursos genéticos de maior prioridade usando os limitados recursos financeiros disponibilizados, sobretudo pelo setor público.

De modo geral, o objetivo de conservação de germoplasma é manter a maior diversidade genética possível para uma dada espécie de planta em particular. Até algum tempo atrás, a inexistência de estratégias e métodos eficientes de identificação de duplicatas em bancos de germoplasma, fazia com que os curadores/melhoristas mantivessem todos os acessos na coleção, mesmo com alguma evidência da existência de duplicatas com base em descritores morfoagronômicos. Entretanto, atualmente, a possibilidade de se analisar diferenças diretamente em nível de DNA, e a disponibilidade de diversas ferramentas computacionais para associar as informações fenotípicas e genotípicas, fazem com que o descarte de acessos seja feito de forma mais segura e confiável. Contudo, a padronização na identificação das amostras duplicadas com uso de um conjunto de marcadores comuns, além de problemas na reprodutibilidade dos dados de marcadores moleculares entre diferentes laboratórios, têm dificultado o avanço da técnica e a comparação entre diferentes coleções da espécie. Por outro lado, abordagens padronizadas de NGS, a exemplo da GBS, podem identificar SNPs úteis para sobrepor esses problemas e assim representar uma plataforma de informação ideal para reduzir a redundância no germoplasma (KILIAN, GRANER, 2012). Certamente, estas abordagens mais modernas sobre a conservação dos recursos genéticos geram subsídios importantes para que os curadores/melhoristas priorizem a manutenção de genótipos únicos.

Com as duplicatas identificadas em estudos como este, os curadores podem avaliar a manutenção ou o descarte destes acessos para otimizar a conservação do germoplasma e ainda introduzir novas amostras. Entretanto,

os acessos identificados como idênticos, com base nos marcadores SNPs, devem ainda ser caracterizados no nível fenotípico para determinar se possuem características únicas ou se eles podem ser semelhantes o suficiente para serem consideradas sinonímias na coleção. Estas informações certamente darão maior confiabilidade para o descarte dos acessos redundantes, considerando a possibilidade da existência de variação somaclonal (mutações) no germoplasma de mandioca ao longo do processo de cultivo e domesticação da espécie. De fato, observações desta natureza têm sido relatadas em outras espécies de propagação clonal, a exemplo do abacaxizeiro, cuja análise de SNPs resultou na identificação de 64 acessos únicos em 170 avaliados (ZHOU et al., 2015). Além disso, os autores relataram que alguns acessos dentro do mesmo grupo de duplicatas apresentaram diferenças morfológicas aparentes, apesar de apresentarem perfis únicos de SNP, a exemplo do acesso Cayenne 7898 QC que possui cor amarela da polpa, enquanto o Cayenne 7898 4N possui cor branca.

Diante destas considerações, a caracterização dos acessos de mandioca pertencentes ao mesmo MLG com base em descritores fenotípicos ainda é essencial para complementar o perfil molecular obtido pelos marcadores SNPs para definição precisa da redundância dos acessos para fins de descarte final. Portanto, além de melhorar a eficiência das atividades rotineiras do banco de germoplasma de mandioca, as informações geradas neste trabalho podem ser úteis para a verificação e controle da origem das variedades, direcionando a introdução de novos acessos e mesmo contribuindo para minimizar problemas de registro e recomendação de variedades sinonímias.

## CONCLUSÃO

Este estudo fornece uma visão abrangente da redundância do germoplasma cultivado de *M. esculenta* no Brasil, um país considerado o centro de origem e diversidade da espécie (OLSEN, 2004). O uso de mais de 20 mil SNPs possibilitou a identificação de mais de 25% de duplicatas nos acessos deste germoplasma e uma redundância dentro de algumas coleções da EMBRAPA, como resultado do intercâmbio institucional e coleta de acessos comuns em diferentes regiões geográficas. Neste último caso, a intensa troca

de material propagativo de mandioca entre os agricultores no Brasil favorece à disseminação de genótipos de interesse em todo o território nacional. Geralmente este intercâmbio é acompanhado da troca de nomes, o que certamente faz com que os curadores tenham dúvida da origem e decidam manter os acessos na coleção até que algum tipo de caracterização possa definir o seu correto background genético.

Apesar da elevada confiabilidade dos SNPs para definição das duplicatas de acessos, a ocorrência de mutações espontâneas comumente mencionadas na literatura em espécies de propagação vegetativa, faz com que alterações em um único gene possam gerar variações importantes para a cultura que precisam ser preservadas. Portanto, a decisão final de descarte dos acessos será feita após uma completa caracterização fenotípica com base em dados produtivos, de resistência a doenças e estresses abióticos (déficit hídrico e deterioração fisiológica pós-colheita), e qualidade de raízes e amido.

O agrupamento bastante claro dos acessos de mandioca com base na ADCP foi fortemente correlacionado com os diferentes MLGs, de tal forma que todos os 124 MLGs foram mantidos nos mesmos grupos da ADCP. Isto demonstra a elevada confiabilidade na definição das duplicatas de acessos, considerando que ambos os métodos utilizam abordagens diferentes para o agrupamento dos acessos. Além disso, ampla diversidade genética foi verificada nos acessos duplicados considerando que apenas as duas primeiras funções discriminantes foram capazes de distinguir, com elevada probabilidade de alocação, os acessos de mandioca. Todas estas informações serão úteis para melhorar a eficiência no manejo do germoplasma de mandioca e no seu uso em benefício dos usuários destes recursos genéticos.

## REFERÊNCIAS

ARNAUD-HAOND, S.; DUARTE, C. M.; ALBERTO, F.; SERRÃO, E. A. Standardizing methods to address clonality in population studies. **Molecular Ecology**, v. 16, p. 5115-5139, 2007.

BANDILLO, N.; JARQUIN, D.; SONG, Q.; NELSON, R.; CREGAN, P.; SPECHT, J.; LORENZ, A. A population structure and genome-wide association

analysis on the USDA soybean germplasm collection. **The Plant Genome**, v. 8, p. 1-13, 2015.

BASTOS, T.X.; PACHECO, N.A.; NECHET, D.; ABREU SÁ, T.D. Aspectos climáticos de Belém nos últimos cem anos. Belém: Embrapa Amazônia Oriental, 2002. 31p. (Embrapa Amazônia Oriental. Documentos, 128).

BENIN, G.; MATEI, G.; COSTA DE OLIVEIRA, A.; SILVA, G. O.; HAGEMANN, T. R.; LEMES DA SILVA, C.; PAGLIOSA, E. S.; BECHE, E. Relationships between four measures of genetic distance and breeding behavior in spring wheat. **Genetic and Molecular Research**, v. 16, p. 2390-2400, 2012.

BRADBURY, P. J.; ZHANG, Z.; KROON, D. E.; CASSTEVENS, T. M.; RAMDOSS, Y.; BUCKLER, E. S. TASSEL: software for association mapping of complex traits in diverse samples. **Bioinformatics Applications Note**, v. 23, p. 2633–2635, 2007.

CHAVARRIAGA-AGUIRRE, P.; MAYA, M.M.; TOHME, J.; DUQUE, M.C.; IGLESIAS, C.; BONIERBALE, M.W.; KRESOVICH, S.; KOCHERT, G. Using microsatellites, isozymes and AFLPs 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.

CHEN, X.; XIA, Z.; FU, Y.; LU, C.; WANG, W. Constructing a genetic linkage map using an $F_1$ population of non-inbred parents in cassava (*Manihot esculenta* Crantz). **Plant Molecular Biology Reporter**, v. 28, p.676–683, 2010.

DIAS, L. A. S.; MARITA, J.; CRUZ C. D.; BARROS, E. G.; SALOMÃO, T. M. F. Genetic distance and its association with heterosis in cacao. **Brazilian Archives of Biology and Technology,** v. 46, p. 339-347, 2003.

DOUHOVNIKOFF, V.; DODD, R. S. Intra-clonal variation and a similarity threshold for identification of clones: application to *Salix exigua* using AFLP

molecular markers. **Theoretical and Applied Genetics**, v. 106, p. 1307–1315, 2003.

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, v.19, p.11-15, 1987.

EL-SHARKAWY, M. A. Cassava biology and Physiology. **Plant Molecular Biology**, v. 56, p. 481-501, 2004.

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

FAO (2010) The State of ex situ conservation. In: FAO (ed). **Second Report on the state of the world's plant genetic resources for food and agriculture**. Food and Agriculture Organisation, Rome, Italy, p. 55-90.

FREGENE M, ANGEL F, GÓMEZ R, RODRIGUEZ F, CHAVARRIAGA P, ROCA W, TOHME J, BONIERBALE M (1997) A molecular genetic map of cassava (*Manihot esculenta* Crantz) **Theoretical and Applied Genetics**, v. 95, p. 431–441, 1997.

FU, Y-B.; RICHARDS, K. W.; PETERSON, G.W. Genetic variability in multiple accessions of two Canadian heritage crop cultivars as revealed by AFLP markers. **Communications in Biometry and Crop Science**, v. 1, p.1-10, 2006.

GREENE, S. L.; PEDERSON, G. A. Eliminating duplicates in germplasm collections: a white clover example. **Crop Science**, v.36, p.1398–1400, 1996.

GROSS, B. L.; VOLK, G. M.; RICHARDS, C. M.; FORSLINE, P. L.; FAZIO, G.; CHAO, C.T. Identification of ‘‘duplicate’’ accessions within the USDA-ARS National Plant Germplasm System *Malus* Collection. **Journal of the American Society for Horticultural Science**, v. 137, p. 333–342, 2012.

HAMBLIN, M. T.; RABBI, I. Y. The effects of restriction-enzyme choice on properties of genotyping-by-sequencing libraries: A study in cassava (*Manihot esculenta*). **Crop Science**, v. 54, p. 1–6, 2014.

HORNA, D.; DEBOUCK, D.; CIPRIAN, A.; CUERVO, M.; ESCOBAR, R.; HERNANDEZ, A.; MAFLA, G.; OCAMPO, C.; SANTOS, L. G.; TORO, O. Conservation and management of genetic resources of beans, cassava and tropical forages in the CIAT genebank. In: (Section 3)

HORNA, D.; DEBOUCK, D.; DUMET, D.; HANSON, J.; PAYNE, T.; SACKVILLE-HAMILTON, R.; SANCHEZ, I.; UPADHYAYA, H. D.; VAN DEN HOUWE I. **Evaluating cost effectiveness of collection management: *ex-situ* conservation of plant genetic resources in the CG system**. Consultative Group on International Agricultural Research, 2010, p. 24-35.

JIMENEZ-NIETO, A. Identificación de duplicados del banco de germoplasma de yuca (*Manihot esculenta* Crantz) del CIAT, B. S. Thesis. Universidad Nacional de Colombia, Palmira, Colombia, 1994.

JOMBART, T.; DEVILLARD, S.; BALLOUX, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. **BioMed Central Genetic**, v. 11, p. 1-15, 2010.

KAMVAR, Z. N.; BROOKS, J. C.; GRUNWALD, N. J. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality., v. 6, p. 1-10, 2015.

KAMVAR, Z. N.; TABIMA, J. F.; GRUNWALD, N. J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. **PeerJ**, v. 2, p. 1-14, 2014.

KILIAN, B.; GRANER, A. NGS technologies for analyzing germplasm diversity in genebanks. **Briefings in Functional Genomics**, v. 11, p. 38-50, 2012.

KUNKEAW, S.; TANGPHATSORNRUANG S., SMITH D. R., TRIWITAYAKORN K. Genetic linkage map of cassava (*Manihot esculenta* Crantz) based on AFLP and SSR markers. **Plant Breeding**, v. 129, p. 112–115, 2010.

LOPEZ, C. E.; QUESADA-OCAMPO, L. M.; BOHORQUEZ, A.; DUQUE, M. C.; VARGAS, J.; TOHME, J.; VERDIER, V. (2007) Mapping EST-derived SSRs and ESTs involved in resistance to bacterial blight in *Manihot esculenta.* **Genome**, v. 50, p. 1078–1088, 2007.

LUND, B.; ORTIZ, R.; SKOVGAARD, I. M.; WAUGH, R.; ANDERSEN, S. B. Analysis of potential duplicates in barley gene bank collections using re-sampling of microsatellite data. **Theoretical and Applied Genetics**, v. 106, p.1129–1138, 2003.

MBA, R. E. C.; STEPHENSON, P.; EDWARDS, K.; MELZER, S.; NKUMBIRA, J.; GULLBERG, U.; APEL, K.; GALE, M.; TOHME, J.; FREGENE, M. simple sequence repeat (SSR) markers survey of the cassava (*Manihot esculenta* Crantz) genome: towards an SSR-based molecular genetic map of cassava. **Theoretical and Applied Genetics**, v. 102, p. 21–31, 2001.

MAMMADOV, J.; AGGARWAL, R.; BUYYARAPU, R.; KUMPATLA, S. SNP markers and their impact on plant breeding. **International Journal of Plant Genomics**, v. 2012, p. 1-12, 2012.

MELO, A. T. O.; GUTHRIE, R. S.; HALE, I. GBS-Based deconvolution of the surviving north American collection of cold-hardy kiwifruit (*Actinidia* spp.) Germplasm. **Plos One**, v. 12, p. 1-21, 2017.

MOURA, F. E.; FARIAS NETO, T. J.; SAMPAIO, E. J.; SILVA, T. D.; RAMALHO, F. G. Identification of duplicates of cassava accessions sampled on the North Region of Brazil using microsatellite markers. **Acta Amazonica**, v. 43, p. 461-468, 2013.

MOURA, E. F.; SOUSA, N. R.; MOURA, M. F.; DIAS, M. C.; SOUZA, E. D.; FARIAS NETO, J. T.; SAMPAIO, J. E. Molecular characterization of accessions of a rare genetic resource: sugary cassava (*Manihot esculenta* Crantz) from Brazilian Amazon. **Genetic Resources and Crop Evolution**, v. 63, p. 583-593, 2016.

MTUNGUJA, M. K.; RANJAN, A.; LASWAI, H. S.; MUZANILA, Y.; NDUNGURU, J.; SINHA, N. R. Genetic diversity of farmer-preferred cassava landraces in Tanzania based on morphological descriptors and single nucleotide polymorphisms. **Plant Genetic Resources: Characterization and Utilization**, v. 15, p. 138-146, 2017.

NASSAR, N. M. Cassava, *Manihot esculenta* Crantz, genetic resources: origin of the crop, its evolution and relationships with wild relatives. **Genetic Molecular Research**, v. 1, p. 298–305, 2002.

OCAMPO, C.; ANGEL, F.; JIMÉNEZ, A.; JARAMILLO, G.; HERSHEY, G.; GRANADOS, E.; IGLESIAS, C. DNA fingerprinting to confirm possible genetic duplicates in cassava germplasm. In: **The Cassava Biotechnology Network: Proceedings second international scientific meeting**. Bogor, Indonesia, 22–26 August 1994, Centro Internacional de Agricultura Tropical, Cali, Colombia, 1995, p. 145–151.

OCAMPO, C.; HERSHEY, C.; IGLESIAS, C.; IWANAGA, M. Esterase isozyme fingerprinting of the cassava germplasm collection held at CIAT. In: ROCA, W. M.; THRO, A. M (Eds.) **Proceedings first international scientific meeting of the Cassava Biotechnology Network**. Cartagena, Colombia, 25–28 August 1992, Centro Internacional de Agricultura Tropical, Cali, Colombia, 1993.

OLIVEIRA, E. J.; SANTANA, F. A.; OLIVEIRA, L. A.; SANTOS, V. S. Genotypic variation of traits related to quality of cassava roots using affinity propagation algorithm. **Scientia Agricola**, v. 72, p. 53-61, 2015.

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

ORTIZ, A. H. T.; ROCHA, V. P. C.; MOIANA, L. D.; GONÇALVES-VIDIGAL, M. C.; GALVÁN, M. Z.; VIDIGAL FILHO, P. S. Population structure and genetic diversity in sweet cassava cultivars from Paraná, Brazil. **Plant Molecular Biology Reporter**, v. 34, p. 1153-1166, 2016.

R Development Core Team (2017). R: A language and environment for statistical computing, reference index version 3.3.4. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0 http://www.R-project.org.

RABBI, I. Y.; KULAKOW, P. A.; MANU-ADUENING, J. A.; DANKYI, A. A.; ASIBUO, J. Y.; PARKES, E. Y.; ABDOULAYE, T.; GIRMA, G.; GEDIL, M. A.; RAMU, P.; REYES, B.; MAREDIA, M. K. Tracking crop varieties using genotyping-by-sequencing markers: a case study using cassava (*Manihot esculenta* Crantz). **BioMedCentral Genetics**, v. 16, p. 1-11, 2015.

SEMAGN, K.; MAGOROKOSHO, C.; VIVEK, B. S.; MAKUMBI, D.; BEYENE, Y.; MUGO, S.; PRASANNA, B. M; WARBURTON, M. L. Molecular characterization of diverse CIMMYT maize inbred lines from eastern and southern Africa using single-nucleotide polymorphic markers. **BioMedCentral Genomics**, v. 13, p. 1-13, 2012.

SHANDS, H.; HAWTIN, G.; MACNEIL, G. The cost to the CGIAR centres of maintaining and distributing germplasm. **CGIAR Consortium**, p. 81, 2010.

SONG, Q.; HYTEN, D. L.; JIA, G.; QUIGLEY, C. V.; FICKUS, E. W.; NELSON, R. L.; CREGAN, P. B. Fingerprinting soybean germplasm and its utility in genomic research. **G3: Genes, Genomes, Genetics**, v. 5, p. 1999-2006, 2015.

TEIXEIRA, A. H. C. Informações agrometeorológicas do polo Petrolina, PE/Juazeiro - 1963 a 2009/Antônio Heriberto de Castro Teixeira. Petrolina: Embrapa Semiárido, 2010, 21 p. (Embrapa Semiárido. Documentos, 233).

TREUREN, R.; HINTUM, T. J. L. Marker-assisted reduction of redundancy in germplasm collections: genetic and economic aspects. **Acta Horticulturae**, v. 623, p.139–149, 2003.

VASCONCELOS, L. M.; BRITO, A. C.; CARMO, C. D.; OLIVEIRA, E. J. Polymorphism of starch pathway genes in cassava. **Genetics and Molecular Research**, v. 15, p. 2-15, 2017.

VAN HINTUM T. J. L.; BOUKEMA, I. W.; VISSER, D. L. Reduction of duplication in a *Brassica oleracea* germplasm collection. **Genetic Resources and Crop Evolution**, v. 43, p.343–349, 1996.

VAN HINTUM, T. J. L.; KNÜPFFER, H. Duplication within and between germplasm collections. I. Identifying duplication and the basis of passport data. **Genetic Resources and Crop Evolution**, v. 42, p. 127–133, 1995.

VILAS-BOAS, S. A.; HOHENFELD, C. S.; OLIVEIRA, S. A. S.; OLIVEIRA, E. J. Sources of resistance to cassava root rot caused by *Fusarium* spp.: a genotypic approach. **Euphytica**, v. 209, p. 237-251, 2016.

ZHOU, L.; MATSUMOTO, T.; TAN, H-W.; MEINHARDT, L.W.; MISCHKE, S.; WANG, B.; ZHANG, D. Developing single nucleotide polymorphism markers for the identification of pineapple (*Ananas comosus*) germplasm. **Horticulture Research**, v. 2, p. 1-12, 2015.

WANG, B.; GUO, X.; ZHAO, P.; RUAN, M.; YU, X.; ZOU, L.; YANG, Y.; LI, X.; DENG, D.; XIAO, J.; XIAO, Y.; HU, C.; WANG, X.; WANG, W.; PENG, M. Molecular diversity analysis, drought related marker-traits association mapping and discovery of excellent alleles for 100-day old plants by EST-SSRs in cassava germplasms (*Manihot esculenta* Crantz). **Plos One**, v. 12, p. 1-23, 2017.

## CONSIDERAÇÕES FINAIS

A aplicação de novas técnicas no melhoramento genético da mandioca tem crescido nos últimos anos, sobretudo com a geração de um grande número de informações moleculares de baixo custo a exemplo da identificação de SNPs por GBS. Especificamente no presente estudo, o uso dos marcadores SNPs proporcionou um melhor entendimento sobre a diversidade genética e estrutura populacional conservada no BAG-Mandioca.

Com o entendimento da diversidade genética do germoplasma de mandioca, será possível identificar parentais contrastantes para cruzamentos e desenvolvimento de populações segregantes com máxima variabilidade genética, a fim de ampliar o uso do germoplasma conservado. A identificação da endogamia presente no BAG-Mandioca constitui-se uma excelente oportunidade para explorar os benefícios da maior homozigosidade dos genótipos, uma vez que os acessos mais homozigóticos podem ser autofecundados para rápida obtenção e seleção de linhas puras, tendo como foco a exploração do vigor hibrido e de novos fenótipos com expressão em homozigose recessiva, a exemplo do amido ceroso.

A formação de 22 grupos de diversidade molecular pelo método ADCP reflete a grande variabilidade genética existente no germoplasma de mandioca do Brasil, fruto da intensa domesticação e seleção histórica neste país. Entretanto, a falta de associação entre os agrupamentos formados com base nas informações moleculares e fenotípicas demonstrou a necessidade de esforços adicionais na aquisição de outras informações fenotípicas, visando uma melhor classificação do germoplasma.

Em relação ao estudo de duplicatas, sugere-se que os acessos considerados como duplicatas sejam caracterizados fenotipicamente e somente após a confirmação da redundância genética com base em dados moleculares e fenotípicos seja realizada o seu descarte final, para manter conservada uma maior representatividade de todo recurso genético disponível da espécie.