
**UNIVERSIDADE ESTADUAL DE CAMPINAS**

**INSTITUTO DE BIOLOGIA**

**PAULA ARIELLE MENDES RIBEIRO VALDISSER**

# GENÔMICA POPULACIONAL E ANÁLISE DE ASSOCIAÇÃO GENÔMICA AMPLA (GWAS) PARA TOLERÂNCIA À SECA E PRODUTIVIDADE EM FEIJOEIRO COMUM (*Phaseolus vulgaris* L.)

# POPULATION GENOMICS AND GENOME-WIDE ASSOCIATION STUDIES (GWAS) FOR DROUGHT TOLERANCE AND YIELD IN COMMON BEAN (*Phaseolus vulgaris* L.)

**CAMPINAS**

**2017**

**PAULA ARIELLE MENDES RIBEIRO VALDISSER**

# GENÔMICA POPULACIONAL E ANÁLISE DE ASSOCIAÇÃO GENÔMICA AMPLA (GWAS) PARA TOLERÂNCIA À SECA E PRODUTIVIDADE EM FEIJOEIRO COMUM (*Phaseolus vulgaris* L.)

# POPULATION GENOMICS AND GENOME-WIDE ASSOCIATION STUDIES (GWAS) FOR DROUGHT TOLERANCE AND YIELD IN COMMON BEAN (*Phaseolus vulgaris* L.)

Dissertação apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do Título de Mestra em Genética e Biologia Molecular na área de Bioinformática.

Dissertation presented to the Institute of Biology of the University of Campinas in partial fulfillment of the requirements for the Master's degree in Genetics and Molecular Biology in the field of Bioinformatics.

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELA ALUNA PAULA ARIELLE MENDES RIBEIRO VALDISSER E ORIENTADA PELA DRA. MARIA IMACULADA ZUCCHI.

Orientadora: Maria Imaculada Zucchi

Coorientadora: Rosana Pereira Vianello

CAMPINAS

2017

**Agência(s) de fomento e nº(s) de processo(s):** Não se aplica.

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

V233g

Valdisser, Paula Arielle Mendes Ribeiro, 1983-
Genômica populacional e análise de associação genômica ampla (GWAS) para tolerância à seca e produtividade em feijoeiro comum (*Phaseolus vulgaris* L.) / Paula Arielle Mendes Ribeiro Valdisser. – Campinas, SP : [s.n.], 2017.

Orientador: Maria Imaculada Zucchi.
Coorientador: Rosana Pereira Vianello.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Feijão comum. 2. Polimorfismo de nucleotídeo único. 3. Plantas - Tolerância à seca. I. Zucchi, Maria Imaculada. II. Vianello, Rosana Pereira. III. Universidade Estadual de Campinas. Instituto de Biologia. IV. Título.

Informações para Biblioteca Digital

**Título em outro idioma:** Population genomics and genome-wide association studies (GWAS) for drought tolerance and yield in common bean (*Phaseolus vulgaris* L.)
**Palavras-chave em inglês:**
Common bean
Single nucleotide polymorphism
Plants - Drought tolerance
**Área de concentração:** Bioinformática
**Titulação:** Mestra em Genética e Biologia Molecular
**Banca examinadora:**
Maria Imaculada Zucchi [Orientador]
Anete Pereira de Souza
Gabriel Rodrigues Alves Margarido
**Data de defesa:** 22-02-2017
**Programa de Pós-Graduação:** Genética e Biologia Molecular

**Campinas, 22/02/2017.**

**COMISSÃO EXAMINADORA**

Prof.(a) Dr.(a). Maria Imaculada Zucchi

Prof.(a). Dr.(a) Anete Pereira de Souza

Prof.(a) Dr(a). Gabriel Rodrigues Alves Margarido

*Os membros da Comissão Examinadora acima assinaram a Ata de Defesa, que se encontra no processo de vida acadêmica do aluno.*

## DEDICATÓRIA

À Deus, por ter abençoado e me dado forças durante toda minha trajetória.

Aos meus pais, meus maiores incentivadores nos estudos.

À minha irmã e meu irmão, companheiros e amigos pra toda vida.

A meu esposo pelo amor, companheirismo e apoio.

À minha filha Luiza por todo amor, carinho e fofura.

## AGRADECIMENTOS

Primeiramente, agradeço à Deus pela bênçãos, saúde e proteção durante todo o percurso do meu mestrado.

Agradeço à minha família. Aos meus pais (Delcides e Maria Helena) por toda educação e valores morais que me deram e por serem meus exemplos de determinação, coragem e dedicação. Ao meu esposo, Thiago, por todo amor, apoio, compreensão e incentivo. À minha filha, Luiza, por toda compreensão, motivação e momentos de alegria e descontração. Aos meus irmãos (Daniel e Karla), cunhados (Renata, Wender, Fábio, Juliana e Diselma), sobrinhos (Matheus, Isabela, Kelly e Lara) e tia Magda, por todo carinho, apoio e ajuda nos momentos necessários.

Às minhas orientadoras Maria Imaculada Zucchi e Rosana Pereira Vianello por toda confiança depositada em mim, por todos os ensinamentos, sugestões, contribuições, críticas e palavras de apoio e incentivo. Vocês são exemplos e inspiração para mim de profissionais comprometidas, dedicadas e sempre dispostas a ajudar. Foi uma honra ser orientada por vocês. Serei eternamente grata.

À Embrapa por ter me concedido a licença, a bolsa de estudo, a infraestrutura e recursos financeiros necessários para realização deste trabalho.

À UNICAMP, em especial ao programa de pós-graduação do Instituto de Biologia, representados por todos os professores e funcionários, pela oportunidade de realização do curso de Mestrado.

A todos os co-autores e colaboradores deste trabalho pela ajuda nas análises, contribuições na leitura e escrita e por todos os ensinamentos compartilhados. Muito obrigada pelo comprometimento de todos vocês. Obrigada Wendell, Jâneo, Bárbara, Gesimária, Ivandilson, Claudio, João Antônio, Tereza, Odilon, Cléber, João Paulo, Jaison, Anna Cristina, Alessandra, Leandro, Alexandre.

Ao Laboratório de Biotecnologia, pela estrutura física e recursos fornecidos para extração e quantificação de DNA, e às alunas Lorraynne e Ariadna que me auxiliaram nesta etapa.

Ao João Antônio e a todos os estudantes que me auxiliaram na coleta de dados fenotípicos: Douglas, Rodrigo, Ariadna, Milena, Luann.

Ao BAG (Dr. Jaison, Divino e Neide) pelo treinamento na caracterização morfológica dos grãos de feijão e no fornecimento de informações sobre os materiais utilizados.

Ao Melhoramento de Feijão por todo o apoio na montagem dos experimentos de campo, apoio nos trabalhos de multiplicação de sementes em telados e pela infraestrutura disponibilizada para coleta dos dados fenotípicos.

Ao Dr. Cléber e toda equipe de Porangatu/GO que conduziram os experimentos de campo.

A todos os colegas da Embrapa Arroz e Feijão, que direta ou indiretamente contribuíram para o desenvolvimento deste trabalho. Por todas as palavras de apoio e incentivo e por todo conhecimento compartilhado. Em especial, ao Dr. Claudio Brondani, Dra. Tereza Cristina, João Antônio, José Simião, Dra. Gesimária Ribeiro, Dra. Luana Alves, Dra. Anna Cristina.

A todos os professores e alunos do Departamento de Genética e Biologia Molecular da UNICAMP e da Escola de Agronomia de Universidade Federal de Goiás por toda transmissão e compartilhamento de conhecimentos.

A todos os professores e pesquisadores que fizeram parte do exame de qualificação e pré-banca por todas as excelentes contribuições neste trabalho. Obrigada Dra. Anete Pereira de Souza, Dr. Alisson Fernando Chiorato, Dra. Glyn Mara Figueira e Dra. Regina Helena Geribello Priolli.

Enfim, agradeço a todos aqui citados pelas contribuições na minha formação. Toda essa trajetória só foi possível graças ao apoio de todos vocês. Serei eternamente grata.

## RESUMO

O feijoeiro comum é uma importante fonte de proteínas e fibras, sendo considerado o mais importante legume de grãos para o consumo humano direto no mundo. Fatores como o aumento da população, mudanças climáticas globais, diminuição da disponibilidade de água para a agricultura e redução das áreas cultiváveis têm causado grande preocupação quanto à estabilidade e manutenção da produção agrícola em níveis que possam atender à demanda de consumo do grão. Diante disto, se torna fundamental a busca por alternativas que possam incrementar a produção, tais como o desenvolvimento de novas cultivares com características vantajosas e com maior potencial produtivo. O uso dos recursos genéticos presentes nos bancos de germoplasma é uma importante estratégia para busca de novos alelos e incorporação de variabilidade genética nos programas de melhoramento, permitindo o desenvolvimento de cultivares com combinações alélicas novas e favoráveis. Entretanto, para a conservação e uso eficiente destes recursos genéticos é necessária uma adequada caracterização molecular e fenotípica. Atualmente, a caracterização molecular de alta densidade pode ser realizada a custos bastante reduzidos, graças à utilização da metodologia de genotipagem por sequenciamento (GbS), a partir da qual milhares de polimorfismos de base única (SNPs) são identificados e utilizados na genotipagem de acessos de bancos de germoplasma. Associado a esse nível de resolução de genoma, estão disponíveis métodos de mensuração, em larga escala, de fenótipos relacionados a caracteres de importância econômica, como os componentes de produtividade em condições de seca. Através da correlação entre os fenótipos e genótipos é possível explorar a associação não aleatória entre alelos de diferentes locos, a determinado caráter de interesse, através do estudo de associação genômica ampla (GWAS). Neste contexto, o objetivo deste trabalho foi caracterizar, molecularmente, variedades de feijão pertencentes à Coleção Nuclear de Feijão da Embrapa (CONFE) e identificar, através da análise de GWAS, regiões do genoma potencialmente associadas ao controle da tolerância a seca em feijão. No primeiro capítulo, um total de 6286 SNPs foram obtidos através da metodologia DArTseq, sendo utilizados para estimar a diversidade ($H_E$ = 0,442) e divergência genética ($F_{ST}$ = 0,747) nos acessos da CONFE, o decaimento do desequilíbrio de ligação (88 kb, variando de ~ 395 kb no Andino a ~ 130 kb no Mesoamericano), os locos que, possivelmente, estão sob seleção devido aos processos de domesticação e seleção artificial (59 SNPs entre os pools gênicos), além da seleção de um conjunto de 560 SNPs para compor um painel altamente informativo, para uso na caracterização do germoplasma de origens diversas (Andina e Mesoamericana). No segundo capítulo, utilizando um conjunto de 8789 SNPs, derivados das metodologias DArTseq e Capture-Seq, foram identificados, através de GWAS, 212 SNPs significativamente associados aos componentes de produtividade coletados em ambientes com e sem deficiência hídrica. Estes SNPs estão dentro ou, em provável, desequilíbrio de ligação com genes que codificam proteínas putativas que participam dos processos moleculares relacionados a tolerância à seca. Adicionalmente, foram identificados genótipos com desempenho superior em ambientes com e sem déficit hídrico. Ao final destes estudos, foi possível obter: 1) um painel de 560 SNPs para caracterizar amplamente o germoplasma de feijoeiro; 2) um painel de 343 indivíduos genotipados que podem ser utilizados em estudos de GWAS para caracteres de interesse dos programas de melhoramento; 3) um conjunto de SNPs com potencial de uso na seleção assistida para tolerância à seca.

## ABSTRACT

The common bean is an important source of protein and fiber and it is considered the most important grain legume for direct human consumption in the world. Factors such as population growth, global climate change, scarcity water for agriculture and reduction of arable land have caused great concern about the stability and maintenance of agricultural production at levels that can meet the consumption demand of the grain. Because of this, the search for alternatives that can increase production is fundamental, such as the development of new cultivars with advantageous traits and with greater productive potential. The use of the genetic resources present in germplasm banks is an important strategy to search for new alleles and to incorporate genetic variability in breeding programs, allowing the development of cultivars with new and favorable allelic combinations. However, for the conservation and efficient use of these genetic resources an adequate molecular and phenotypic characterization is necessary. Nowadays, the high-density molecular characterization can be realized at low cost because of the use of genotyping by sequencing (GbS) methodology, from which thousands of single nucleotide polymorphisms (SNPs) are identified and genotyped. Associated with this level of genome resolution, methods for large-scale measurement of phenotypes related to economically important traits are available, such as productivity components under drought conditions. Through the correlation between phenotypes and genotypes, it is possible to explore the non-random association between alleles of different loci, to the trait of interest, through the genome-wide association studies (GWAS). In this context, the aim of this work is to characterize molecularly bean varieties belonging to the Core Collection of Beans (CONFE) of the Embrapa and to identify genome regions potentially associated with drought tolerance in common bean through GWAS. In the first chapter, a total of 6,286 SNPs were obtained from DArTseq methodology, being used to estimate the diversity ($H_E$ = 0.442) and genetic divergence ($F_{ST}$ = 0.747) in the CONFE accessions, to determine the decay of the linkage disequilibrium (88 kb, ranging from ~ 395 kb in the Andean to ~ 130 kb in the Mesoamerican) and to detect the loci that are possibly under selection due to the processes of domestication and artificial selection (59 SNPs between the gene pools). Besides, it was selected a set of 560 SNPs to compose a highly informative panel for use in the characterization of germplasm from diverse origins (Andean and Mesoamerican). In the second chapter, using a set of 8,789 SNPs derived from the DArTseq and Capture-Seq methodologies, 212 SNPs significantly associated with the productivity components collected in environments with and without water deficiency were identified through GWAS. These SNPs are within or likely in linkage disequilibrium with genes encoding putative proteins that participate in molecular processes related to drought tolerance. In addition, genotypes with superior performance were identified in environments with and without water deficit. At the end of these studies it was possible to obtain: 1) a panel of 560 SNPs to broadly characterize common bean germplasm; 2) a panel of 343 genotyped individuals that can be used in GWAS for traits of interest to breeding programs; 3) a set of SNPs with potential use in marker-assisted selection for drought tolerance.

**SUMÁRIO**

## INTRODUÇÃO GERAL

O feijoeiro comum pertence à família das leguminosas (Fabacea) e gênero Phaseolus (Doyle e Luckow 2003). As leguminosas são plantas vitais na agricultura, pois formam associações com bactérias fixadoras de nitrogênio, aumentando a fertilidade dos solos. Esta é a principal razão pela qual as leguminosas são mais ricas em proteínas do que todas as outras plantas (Broughton et al. 2003). O gênero Phaseolus contém aproximadamente 70 espécies, das quais cinco são amplamente cultivadas: *P. vulgaris* L., *P. lunatus* L., *P. coccineus* L., *P. acutifolius* A. Gray var. *latifolius* Freeman e *P. polyanthus* Greenman (Freytag e Debouck 2002). Dentre essas espécies, o feijoeiro comum (*Phaseolus vulgaris* L.) destaca-se por ser a mais consumida, equivalendo à metade das leguminosas de grãos consumidas no mundo inteiro (Broughton et al. 2003). Esta espécie fornece até 15% do total de calorias diárias e 36% de proteína total diária para populações de regiões da África e das Américas (Schmutz et al 2014).

De acordo com as estatísticas da FAO (Food and Agriculture Organization of the United Nations), a produção mundial de feijão em 2014 foi de aproximadamente 26,5 milhões de toneladas, cultivada em 30,6 milhões de hectares, sendo o Brasil considerado o terceiro maior produtor mundial, com produção de aproximadamente 3,3 milhões de toneladas (FAO 2014). No Brasil, o feijoeiro comum é cultivado em, praticamente, todos os estados brasileiros, nas mais variadas condições edafoclimáticas e em diferentes épocas e sistemas de cultivo (Carneiro 2002).

### Origem, domesticação e melhoramento

O feijoeiro comum é originário das Américas e é organizado em dois *pools* gênicos (Mesoamericano e Andino) que divergiram de um ancestral silvestre em comum há mais de 100000 anos atrás (Mamidi et al. 2013). Eventos de domesticação independentes ocorreram dentro de cada *pool* gênico, há aproximadamente 8000 anos, provocando uma alta diferenciação genética entre estes e grandes alterações morfológicas, incluindo o tamanho e cores das sementes, tamanho das folhas e variações no hábito de crescimento (Singh et al. 1991, McClean et al. 2002). Outra teoria sobre a história evolutiva do feijoeiro comum sugere que o *pool* gênico Andino foi derivado de uma população Mesoamericana silvestre, a partir de uma população de fundação de apenas alguns milhares de indivíduos, sofrendo, portanto, um

gargalo genético que ocasionou a redução da diversidade genética neste *pool* gênico em comparação ao Mesoamericano (Bitocchi et al. 2012).

A domesticação em feijoeiro comum promoveu várias alterações fisiológicas e morfológicas, como o hábito de crescimento (determinado x indeterminado), dormência da semente (presente x ausente), sensibilidade ao fotoperíodo, forma, cor e tamanho das sementes. Todas essas modificações diferenciaram as plantas cultivadas de seus ancestrais silvestres, tornando-as mais adaptadas às diferentes condições edafoclimáticas (Gepts e Papa 2002). Os processos de domesticação em feijoeiro comum têm sido amplamente estudados e os principais genes associados têm sido mapeados e caracterizados (Koinange et al. 1996, Repinski et al. 2012, Schmutz et al. 2014).

O Brasil é considerado um centro secundário de diversidade do feijoeiro comum, devido à grande variabilidade genética existente dentro da espécie no país (Burle et al. 2010). Freitas et al. (2006), com base em dados de faseolina obtidos da análise de uma amostra arqueológica de feijão encontrada em uma caverna no Norte do estado de Minas Gerais (datada de 1660 a 1738), provavelmente cultivada pelos indígenas, gerou evidência de maior relação com as variedades de feijão encontradas no Norte da América do Sul e México (*pool* gênico Mesoamericano). Burle et al. (2010), em um estudo de diversidade utilizando marcadores microssatélites (Single Sequence Repeats, SSRs), concluiu que, apesar dos dois *pools* gênicos estarem presentes no Brasil, existe um forte predomínio de variedades tradicionais do germoplasma Mesoamericano.

No Brasil, os programas de melhoramento genético do feijoeiro comum são focados, principalmente, no aumento da produtividade da cultura, resistência às principais pragas, doenças e fatores abióticos, além dos esforços para agregar valor de mercado às cultivares, por meio do aumento de teores nutricionais de ferro e zinco no grão (Perseguini 2013).

**Germoplasma**

Apesar da extensa variabilidade genética do conjunto de acessos de feijoeiro comum armazenados em bancos de germoplasma no Brasil, as cultivares atuais apresentam baixa variabilidade genética entre si, principalmente devido à pressão do mercado por tipos específicos de grãos, o que as tornam mais vulneráveis às pragas, doenças e à ocorrência de estresses abióticos, como a seca, que no conjunto resultam em menor produtividade de grão (Abdurakhmonov e Abdukarimov 2008). Isso demonstra a importância de estudos e do uso

dos recursos genéticos disponíveis nos bancos de germoplasma para ampliação da variabilidade genética disponível aos programas de melhoramento, permitindo o surgimento de combinações alélicas novas, e com grande potencial de obtenção de linhagens e cultivares com maior capacidade adaptativa e maior estabilidade de produção.

Os bancos de germoplasmas são reservatórios das variações genéticas de uma espécie e englobam materiais silvestres, variedades crioulas, linhagens, híbridos e cultivares comerciais. Para o enriquecimento dos bancos, são implementadas estratégias de intercâmbio de germoplasma, que consiste na transferência de patrimônio genético vegetal de um país para outro, consolidado no princípio de reciprocidade e obedecendo a regras específicas, que visam, além de preservar o patrimônio genético de cada país, atender as normas de segurança biológica e leis fitossanitárias, para evitar a introdução de pragas no país. Para isso, os materiais introduzidos são submetidos à quarentena de pós-entrada, na qual eles são analisados para verificar se existe alguma praga que ainda não ocorra no país (EMBRAPA 2016).

Para acessar a variabilidade genética existente nos bancos de germoplasma, é necessária sua caracterização agromorfológica e molecular. Dependendo do tamanho do banco de germoplasma, a caracterização de todos os acessos se torna inviável e, por isso, são estabelecidas as coleções, que são conjuntos de genótipos representativos da variabilidade genética da espécie e podem ser criadas para diferentes fins como, por exemplo, as coleções nucleares (Bespalhok et al. 2007). Estas coleções foram definidas por Frankel (1984), como uma amostra com número limitado de acessos que representam a máxima variabilidade genética da espécie conservada, com um mínimo de genótipos geneticamente semelhantes. As coleções nucleares não são estabelecidas com o objetivo de substituir a coleção original de germoplasma, mas sim para possibilitar que estudos de caracterização de uma série de variáveis sejam permitidos pelo menor número de acessos.

A Embrapa possui um banco de germoplasma de feijão com, aproximadamente, 17345 acessos, dos quais 600 acessos, representativos da variabilidade genética da coleção principal, foram selecionados para compor uma coleção nuclear, denominada Coleção Nuclear de Feijão da Embrapa (CONFE), que é formada por acessos pertencentes aos *pools* gênicos Andino e Mesoamericano e, estão distribuídos em três estratos: a) variedades tradicionais do Brasil; b) linhagens/cultivares melhoradas do Brasil; e c) linhagens/cultivares introduzidas (originárias de outros países) (Rangel et al. 2013).

Estudos de diversidade genética em coleções de germoplasma se fazem necessários para viabilizar a manutenção da variabilidade disponível por meio do

conhecimento da extensão dessa variabilidade, e possibilitando maior utilização dos acessos nos programas de melhoramento, principalmente nos casos em que se deseja combinar alelos favoráveis em um único indivíduo (Nass et al. 2001). Os estudos de diversidade genética têm sido realizados em germoplasmas de feijoeiro comum tanto com marcadores morfológicos (Sharma et al. 2013, Kumar et al. 2014), quanto com marcadores moleculares (Khaidizar et al. 2012, Cardoso et al. 2013). Estes últimos têm sido amplamente empregados em análises genéticas devido ao maior poder informativo e estabilidade, já que não sofrem influência do meio ambiente, além da facilidade de manuseio, rapidez e reprodutibilidade (Kumar et al. 2009). Além da caracterização dos acessos do germoplasma, o uso de marcadores moleculares na busca por alelos de interesse que contribuem para a expressão de caracteres de importância agronômica também é relevante para a efetiva utilização destes recursos genéticos nos programas de melhoramento.

**Genômica e marcadores moleculares**

O feijoeiro comum é uma planta autógama, com taxa de fecundação cruzada estimada entre 3 e 5% (Burle et al. 2010), possui genoma diplóide com 11 cromossomos (2n = 22) e tamanho estimado de 588 Mbp (Bennett e Leitch 1995, 2012).

O desenvolvimento de técnicas moleculares para análises genéticas tem proporcionado um grande avanço no conhecimento da genética e estrutura do genoma do feijoeiro comum ao longo dos últimos 20 anos (Ferreira et al. 2010, Hanai et al. 2010, Perseguini et al. 2011, Cardoso et al. 2013). Dentre estas técnicas, destacam-se os marcadores moleculares, os quais têm sido amplamente utilizados para monitorar as variações genéticas entre os indivíduos desta espécie. Estabelecendo uma linha do tempo, já foram utilizados na genotipagem do feijoeiro comum os marcadores RFLP (Restriction Fragment Length Polymorphism) e RAPD (Random Amplified Polymorphic DNA) (Freyre et al. 1998), SSRs (Grisi et al. 2007, Müller et al. 2014) e, mais recentemente, marcadores SNPs (Single Nucleotide Polymorphisms) (Galeano et al. 2011, 2012; Müller et al. 2015; Valdisser et al. 2016). Cada uma dessas tecnologias permitiu, a seu tempo, importantes avanços na compreensão da variabilidade genética e caracterização de germoplasmas, identificação e *fingerprint* de genótipos, estimativas da distância genética entre os acessos, construção de mapas de ligação e detecção de locos de caracteres quantitativos (Quantitative trait loci, QTLs), seleção assistida por marcadores, identificação de sequências de genes candidatos importantes (Freyre et al. 1998, Mukeshimana et al. 2014, Zhang et al. 2015, Müller et al.

2015, Song et al. 2015, Valdisser et al. 2016). Atualmente, os SNPs são amplamente empregados em análise genética, devido a maior abundância no genoma (Brookes 1999), além de serem adequados para o desenvolvimento de métodos de genotipagem de *high throughput* com facilidade para automação (Ding e Jin 2009).

Atualmente, a caracterização molecular de alta densidade tem permitido a identificação de milhares de marcadores moleculares a custos reduzidos como, por exemplo, com a metodologia de genotipagem por sequenciamento (GbS), baseada no sequenciamento de nova geração (Next Generation Sequencing, NGS). Como o custo do sequenciamento reduziu consideravelmente, o NGS viabilizou o sequenciamento de grandes conjuntos de genótipos, abrindo novas oportunidades para explorar a diversidade genética existente nos bancos de germoplasma e aumentando a chance de descoberta de genes pelo aumento da densidade de marcadores e, consequentemente, a maior possibilidade de detecção de QTLs, fornecendo a base para o estudo das relações complexas entre genótipos e fenótipos em todo o genoma. Essas tecnologias permitem o sequenciamento de milhares de bases nucleotídicas utilizando plataformas automatizadas capazes de produzir centenas de gigabases de dados em uma única corrida, em curto período de tempo e com baixo custo por base sequenciada (Varshney et al. 2009). A análise dos dados gerados pelo NGS, através de métodos de bioinformática, permite a descoberta de novos genes, suas posições e sequências reguladoras, além da disponibilização de uma vasta quantidade de marcadores moleculares (Pérez-de-Castro et al. 2012), que podem ser reunidos em painéis de genotipagem de marcadores SNPs (chips), constituídos por SNPs gênicos e não gênicos bem distribuídos ao longo do genoma (Song et al. 2015). Recentemente, diferentes abordagens de genotipagem por sequenciamento em plantas estão sendo disponibilizadas, tais como RADseq (Willing et al. 2011) e DArTseq (Cruz et al. 2013).

A tecnologia DArTseq (Figura 1) consiste no princípio de redução da complexidade genômica das amostras de DNA total utilizando combinações específicas de enzimas de restrição, otimizadas para cada espécie. Após a digestão com enzima de restrição, cada amostra recebe uma sequência adaptadora indexadora única (barcode) que permite, posteriormente, rastrear as sequências geradas para cada amostra. Os fragmentos selecionados resultantes são submetidos ao sequenciamento NGS para a geração de milhões de sequências curtas (em geral 150 bases). As sequências provenientes das diferentes amostras são alinhadas, permitindo a identificação dos SNPs entre as amostras para os mesmos fragmentos de DNA (Sansaloni 2012). O método DArT foi desenvolvido inicialmente em arroz (Jaccoud et al. 2001) e, posteriormente, utilizado em dezenas de espécies de plantas, incluindo os

legumes ervilha (Yang et al. 2011), grão de bico (Thudi et al. 2011) e o feijoeiro comum (Brinêz et al. 2012). A integração da tecnologia de DArT e GbS foi recentemente desenvolvida na Austrália na empresa DArTPty, a qual combina o método de redução de complexidade do genoma com o sequenciamento NGS. O sucesso do emprego dessa estratégia foi demonstrado em eucalipto, no qual foram identificados milhares de marcadores SNPs possibilitando uma ampla amostragem do genoma dessa espécie (Sansaloni 2012).

A metodologia Capture-Seq (Figura 2) consiste na identificação de polimorfismos presentes em regiões genômicas específicas, geralmente contendo genes alvos, através da hibridização dessas regiões com sondas de captura, seguida pelo NGS dos fragmentos capturados (Neves et al. 2013). Resumidamente, a metodologia consiste na etapa inicial de seleção e desenho de sondas de captura, que sejam capazes de hibridizar em regiões únicas e específicas de interesse, com base em um genoma de referência. Em seguida, as amostras de DNA são fragmentadas mecanicamente e, os fragmentos gerados são ligados aos adaptadores específicos, enriquecidos por PCR e hibridizados com as sondas, para posterior sequenciamento NGS.

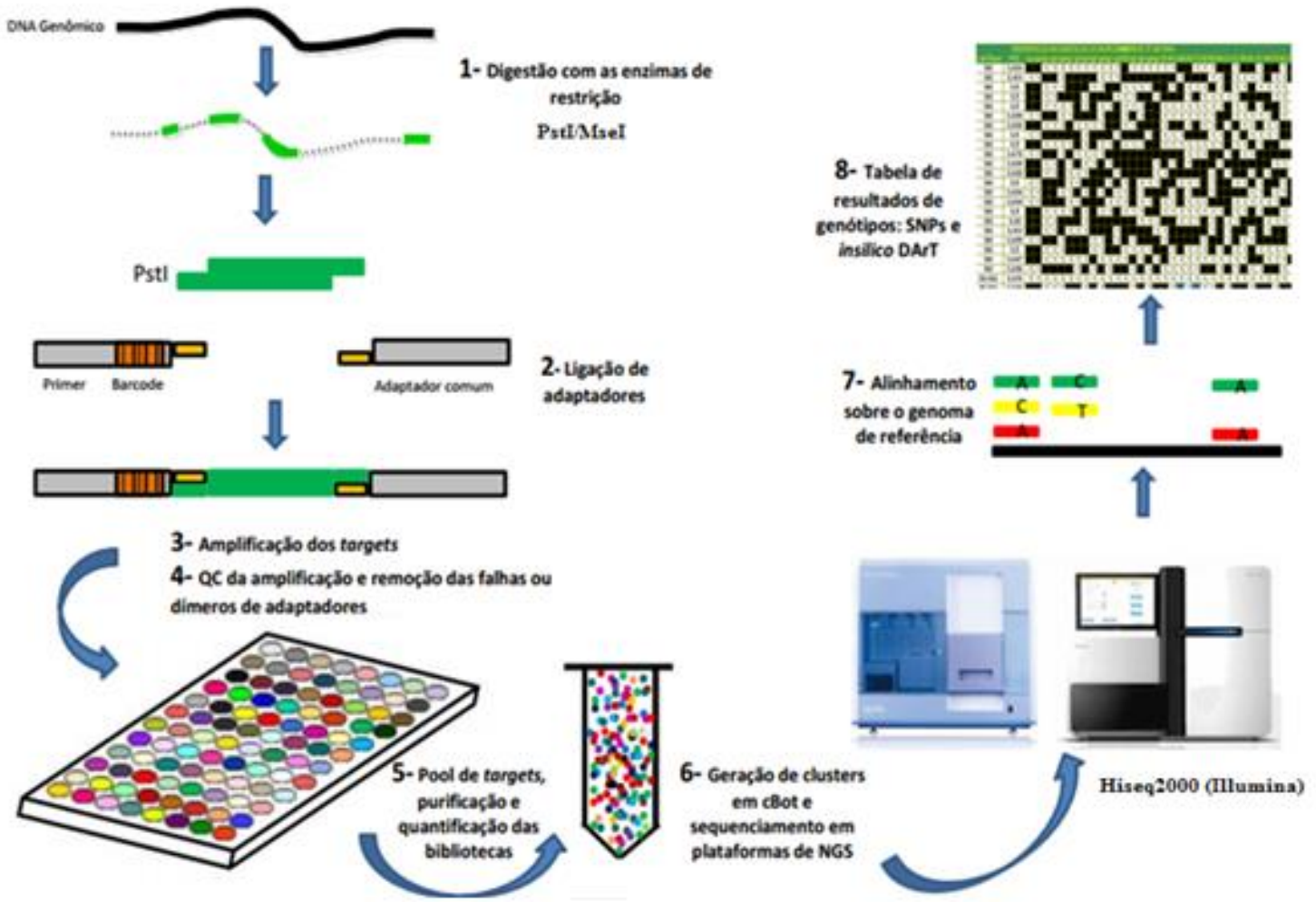



Figura 1. Fluxograma da metodologia DArTseq.
Fonte: Sansaloni 2012.

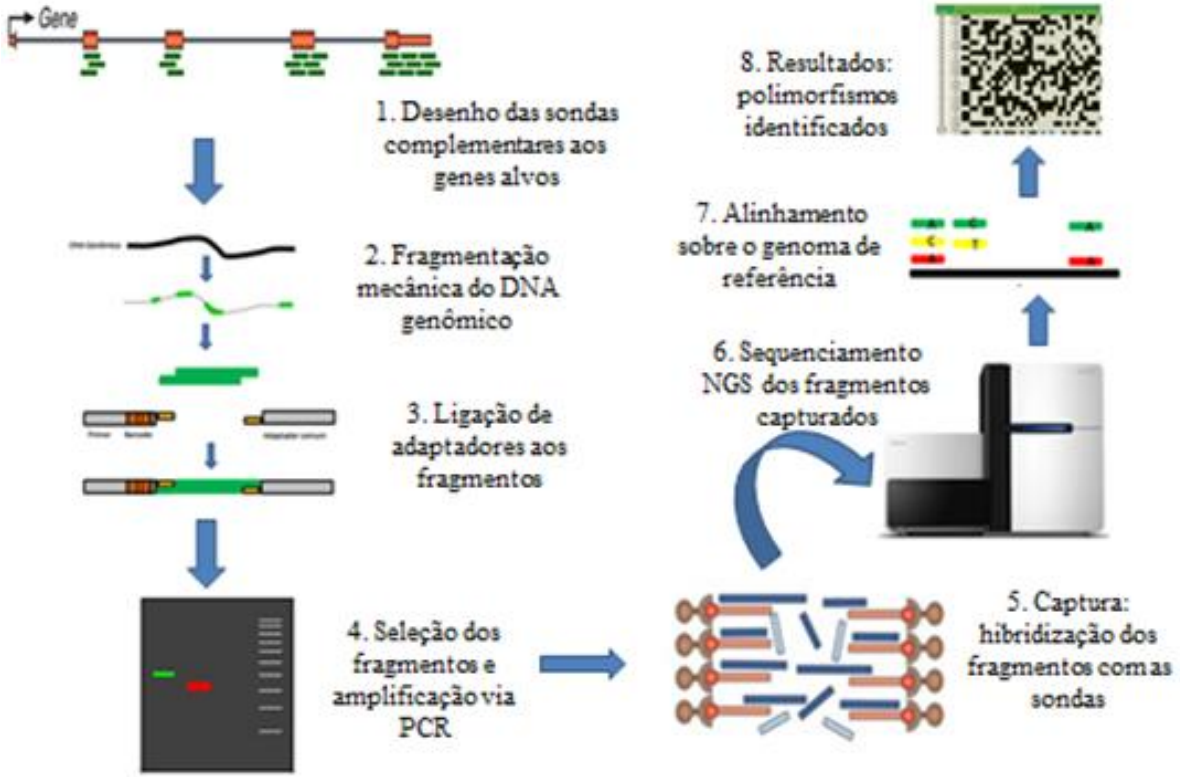



Figura 2. Fluxograma da metodologia Capture-Seq.

A disponibilidade de uma densidade maior de marcadores SNPs distribuídos ao longo de todo o genoma, obtida pelas tecnologias derivadas de NGS, também torna possível a aplicação dos Estudos de Associação Genômica Ampla (Genome-Wide Association Studies, GWAS) com maior eficiência, facilitando o estudo do genótipo e sua relação com o fenótipo. Estes estudos utilizam grande quantidade de indivíduos fenotipados para caracteres de interesse e genotipados para identificação de polimorfismos de DNA para posterior estabelecimento de associações estatísticas, com o objetivo de correlacionar as variações do DNA com as diferenças fenotípicas, visando identificar regiões genômicas onde os genes responsáveis por caracteres de interesse estão localizados (Pérez-de-Castro et al 2012).

**Deficiência hídrica**

As previsões ambientais apontam para o aumento do aquecimento global nas próximas décadas e, consequentemente, este fenômeno será acompanhado por períodos de seca mais intensos e abundantes (Lesk et al. 2016). Tal cenário tende ao agravamento devido ao crescimento populacional e consequente demanda por mais alimentos, principalmente nos países em desenvolvimento (FAO 2016). Neste contexto, o desenvolvimento de cultivares mais tolerantes aos períodos de deficiência hídrica, bem como o investimento em novas tecnologias que auxiliem as plantas a superar períodos prolongados de seca, será fundamental na manutenção da produção agrícola brasileira e mundial em níveis que possam atender uma população que está em constante crescimento (Nepomuceno et al. 2001).

O feijoeiro comum é considerado uma planta sensível à seca, principalmente devido ao sistema radicular pouco desenvolvido e à baixa capacidade de recuperação após períodos de deficiência hídrica (Guimarães 1992). O déficit hídrico é o segundo maior causador da queda de produtividade em feijoeiro comum, após doenças (Singh 1995), afetando cerca de 60% das áreas de produção e é resultante tanto de períodos de estiagem, durante o ciclo de cultivo da espécie, quanto da irregularidade nas precipitações (Graham e Ranalli 1997, McClean et al. 2011). Quando submetido ao déficit hídrico, o feijoeiro comum apresenta redução da área foliar e aumento da resistência estomática (Didonet e Silva 2004), sendo a fase reprodutiva a mais sensível, com maior vulnerabilidade do início da floração até o início da formação de vagens (Fageria et al. 1991).

A resistência à seca é fisiologicamente e geneticamente uma característica complexa, de natureza quantitativa e herança poligênica, resultado da ação conjunta de muitos genes cuja expressão está sob forte influência ambiental (Miklas et al. 2006). Sob condições

de déficit hídrico, as plantas sofrem grandes desordens fisiológicas nas suas células, tecidos e órgãos e, como consequência, ocorre decréscimo das taxas de crescimento, produtividade e qualidade de grãos (Rizhsky et al. 2004, Mittler 2006, Barnabás et al. 2008, Ribeiro et al. 2008, Prasad et al. 2008). Os danos causados pela seca podem ser reversíveis e/ou irreversíveis, dependendo da eficiência dos mecanismos fisiológicos e/ou morfológicos da planta (Ribeiro et al. 2004). Assim, diversas respostas fisiológicas como, distribuição dos estômatos nas folhas, regulação de trocas gasosas, redução de área foliar, arquitetura de raiz, ajuste fenológico, *status* hídrico e ajuste osmótico representam mecanismos de adaptação aos estresses abióticos em plantas de feijoeiro comum (Rosales et al. 2012, Omae et al. 2012, Lanna et al. 2016, Polania et al. 2016).

A maior limitação para o melhoramento genético voltado para tolerância à seca é o conhecimento insuficiente sobre as bases fisiológicas, moleculares e genéticas das respostas das plantas submetidas ao déficit hídrico. Isso reforça a necessidade de um enfoque contínuo nos estudos de tolerância ao déficit hídrico, com maior ênfase nas diferenças genéticas entre os genótipos (Nepomuceno et al. 2001).

Devido à complexidade da resposta das plantas frente às condições de déficit hídrico, o uso de tecnologias de genômica avançada (NGS, GWAS) tem o potencial de identificar, com maior eficiência e precisão, as regiões genômicas e/ou genes responsáveis pelos mecanismos de tolerância à seca, além de fornecer ferramentas, como marcadores moleculares, capazes de explorar a diversidade de genótipos de feijoeiro comum, selecionando aqueles de desempenho superior, e viabilizando a incorporação dos alelos vantajosos por meio da seleção assistida por marcadores nos programas de melhoramento genético.

**Análise de Associação Genômica Ampla**

Uma das estratégias utilizadas para identificar genes associados a fenótipos específicos é o mapeamento de ligação para detecção de QTLs, o qual permite identificar as regiões genômicas que contêm um ou mais genes responsáveis por parte da variação de determinado caráter quantitativo, através da associação entre o fenótipo e o genótipo de indivíduos de uma população segregante. Para o feijão, têm sido conduzidos estudos de detecção de QTLs relacionados a caracteres diversos, tais como características fenológicas, tamanho e qualidade de sementes (Pérez-Vega et al. 2010), tempo de cozimento dos grãos (Garcia et al. 2012) e tolerância à seca (Mukeshimana et al. 2014, Trapp et al. 2015, Villordo-

Pineda et al. 2015). Porém, geralmente, a cobertura genômica utilizada é reduzida, resultando em pouca precisão quanto ao posicionamento dos QTLs. Além disso, o intervalo de um QTL pode ser grande e conter muitos genes, dificultando a identificação dos genes diretamente associados ao caráter avaliado (Alzate-Marin et al. 2005).

A análise de GWAS consegue superar algumas das limitações do mapeamento de ligação, pois, ao utilizar populações naturais (coleções de germoplasma), os eventos de recombinação acumulados ao longo de inúmeras gerações, reduzem o desequilíbrio de ligação, permitindo obter estimativas mais precisas da localização dos genes de interesse (Zhu et al. 2008). Diante disso, o mapeamento por associação tem, como vantagens, a possibilidade de detecção de associações genéticas válidas para toda a população e não somente para um cruzamento, ampliando a variação alélica detectada e conferindo maior precisão e resolução na localização dos QTLs (Flint-Garcia et al. 2003).

Os estudos de GWAS buscam estimar quantos e quais os QTLs responsáveis pela variação nas características fenotípicas, localizar sua posição no genoma e estimar seus efeitos. A identificação destas regiões genômicas associadas a determinado fenótipo possibilita o uso de estratégias para reduzir o tempo e recursos financeiros gastos no desenvolvimento de novas cultivares, como a seleção assistida por marcadores (SAM) e a clonagem gênica.

Recentemente, estudos de GWAS têm sido aplicados em feijoeiro comum, utilizando milhares de SNPs para identificação de regiões genômicas relacionadas a diferentes caracteres agronômicos que afetam a produção e a qualidade nutricional, tais como dias para florescimento, hábito de crescimento, altura da copa, acamamento e peso de sementes (Schmutz et al. 2014, Moghaddam et al. 2016), tempo de cocção (Cichy et al. 2015), fixação biológica de nitrogênio (Kamfwa et al. 2015), resistência à antracnose e mancha angular da folha (Zuiderveen et al. 2016, Perseguini et al. 2016) e tolerância à seca (Hoyos-Villegas et al. 2016).

## OBJETIVO GERAL

Caracterizar molecularmente os acessos da CONFE através da metodologia de genotipagem por sequenciamento e identificar marcadores SNPs associados a genes relacionados à tolerância à seca, por meio da análise de mapeamento associativo (GWAS).

**Objetivos específicos:**

1) Caracterizar um conjunto de 188 acessos (Andinos e Mesoamericanos) da CONFE, por meio de marcadores SNPs gerados pela metodologia DArTseq.

2) Avaliar um conjunto de 343 acessos Mesoamericanos, quanto aos componentes de produtividade (massa de 100 grãos e produtividade) em ambientes com e sem deficiência hídrica.

3) Identificar marcadores SNPs, associados aos genes relacionados à tolerância à seca, por meio da análise de mapeamento associativo, usando um conjunto de 8789 marcadores SNPs gerados pelas tecnologias DArTseq e Capture-Seq.

## REFERÊNCIAS BIBLIOGRÁFICAS

Abdurakhmonov IY, Abdukarimov A (2008) Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int J Plant Genomics. doi: 10.1155/2008/574927

Alzate-Marin AL, Cervigni GDL, Moreira AM, Barros EG (2005) Seleção assistida por marcadores moleculares visando ao desenvolvimento de plantas resistentes a doenças, com ênfase em feijoeiro e soja. Fitopatol Bras 30:333–342. doi: 10.1590/S0100-41582005000400001

Barnabás B, Jäger K, Fehér A (2008) The effect of drought and heat stress on reproductive processes in cereals. Plant, Cell Environ 31:11–38. doi: 10.1111/j.1365-3040.2007.01727.x

Bennett MD, Leitch IJ (1995) Nuclear DNA amounts in angiosperms. Ann Bot 76:113-116.

Bennett MD, Leitch IJ (2012) Angiosperm DNA C-Values database. Release 8.0. Disponível em: www.kew.org/cvalues/.

Bespalhok JCF, Guerra EP, Oliveira R (2007) Uso e conservação do germoplasma. In: Bespalhok JCF, Guerra EP, Oliveira R. Melhoramento de Plantas. Disponível em: http:// www.bespa.agrarias.ufpr.br/conteudo.

Bitocchi E, Nanni L, Bellucci E, et al (2012) Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. Pnas 109:788–796.

Briñez B, Blair MW, Kilian A, et al (2012) A whole genome DArT assay to assess germplasm collection diversity in common beans. Mol Breed 30:181–193. doi: 10.1007/s11032-011-9609-3

Brookes AJ (1999) The essence of SNPs. Gene 234:177–186. doi: 10.1016/S0378-1119(99)00219-X

Broughton WJ, Hernández G, Blair M, et al (2003) Beans (*Phaseolus* spp.) - Model food legumes. Plant Soil 252:55–128. doi: 10.1023/A:1024146710611

Burle ML, Fonseca JR, Kami JA, Gepts P (2010) Microsatellite diversity and genetic structure among common bean (*Phaseolus vulgaris* L.) landraces in Brazil, a secondary center of diversity. Theor Appl Genet 121:801–813. doi: 10.1007/s00122-010-1350-5

Cardoso PCB, Veiga MM, de Menezes IPP, et al (2013) Molecular characterization of high performance inbred lines of Brazilian common beans. Genet Mol Res 12:5467–84. doi: 10.4238/2013.February.6.4

Carneiro JES (2002). Alternativas para obtenção e escolha de populações segregantes no feijoeiro. Tese (Doutorado) - Universidade Federal de Lavras, Lavras. 134p.

Cichy KA, Wiesinger JA, Mendoza FA (2015) Genetic diversity and genome-wide association analysis of cooking time in dry bean (*Phaseolus vulgaris* L.). Theor Appl Genet 128:1555–1567. doi: 10.1007/s00122-015-2531-z

Cruz VM V, Kilian A, Dierig DA (2013) Development of DArT Marker Platforms and Genetic Diversity Assessment of the U.S. Collection of the New Oilseed Crop Lesquerella and Related Species. PLoS One 8:1–13. doi: 10.1371/journal.pone.0064062

Ding C, Jin S (2009) High-throughput methods for SNP genotyping. Methods Mol Biol 578:245-254. doi: 10.1007/978-1-60327-411-1_16.

Doyle JJ, Luckow MA (2003) The Rest of the Iceberg . Legume Diversity and Evolution in a Phylogenetic Context 1. Plant Physiol 131:900–910. doi: 10.1104/pp.102.018150.groups

EMBRAPA. Intercâmbio de germoplasma. Disponível em: www.embrapa.br/recursos-geneticos-e-biotecnologia/pesquisa-e-desenvolvimento/intercambio-de-germoplasma. Acesso: 15 de dezembro de 2016.

Fageria NK, Baligar VC, Jones CA (1991) Common bean and cowpea. In: Fageria NK, Baligar VC, Jones CA (Ed.). Growth and mineral nutrition of field crops. New York : M. Dekker, p.280-318.

FAO (2014). Faostat. Disponível em:www.fao.org/faostat/en/#data/QC. Acesso: 10 de dezembro de 2016.

FAO (2016). Water. Disponível em: www.fao.org/water/en/. Acesso: 12 de dezembro de 2016.

Ferreira LG, Buso GSC, Brondani RPV, et al (2010) Genetic map of the common bean using a breeding population derived from the Mesoamerican gene pool. Crop Breeding and

Applied Biotechnology 10:1-8.

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357-374.

Frankel OH (1984) Genetic perspectives of germplasm conservation. In W. K. Arber et al. (ed.) Genetic Manipulation: Impact on Man and Society. Cambridge Univ. Press, Cambridge, p. 161-170.

Freitas FDO (2006) Evidências genético-arqueológicas sobre a origem do feijão comum no Brasil. Pesqui Agropecu Bras 41:1199–1203. doi: 10.1590/S0100-204X2006000700018

Freyre R, Skroch PW, Geffroy V, et al (1998) Towards an integrated linkage map of common bean. 4. Development of a core linkage map and alignment of RFLP maps. Theor Appl Genet 97:847-856.

Freytag GF, Debouck DG (2002) Taxonomy, distribution, and ecology of the genus phaseolus (Leguminosae-papilionoideae) in North America, Mexico and Central America. Botanical Research Institute of Texas (BRIT), Forth Worth, TX, USA. 298 p.

Galeano CH, Fernandez AC, Franco-Herrera N, Cichy KA, McClean PE, Vanderleyden J, et al. (2011) Saturation of an Intra-Gene Pool Linkage Map: Towards a Unified Consensus Linkage Map for Fine Mapping and Synteny Analysis in Common Bean. PLoS One 6(12): e28135. doi:10.1371/journal.pone.0028135

Galeano CH, Cortés AJ, Fernández AC, et al (2012) Gene-Based Single Nucleotide Polymorphism Markers for Genetic and Association Mapping in Common Bean. BMC Genet 13:48. doi: 10.1186/1471-2156-13-48

Garcia RAV, Rangel PN, Bassinello PZ, et al (2012) QTL mapping for the cooking time of common beans. Euphytica 186:779–792. doi: 10.1007/s10681-011-0587-7

Gepts P, Papa R (2002) Evolution during Domestication. Encycl Life Sci 1–7. doi: 10.1038/npg.els.0003071

Graham PH, Ranalli P. 1997. Common bean (*Phaseolus vulgaris* L.). Field Crops Research 53: 131-146.

Grisi MCM, Blair MW, Gepts P, Brondani C, Pereira PAA, Brondani RPV (2007) Genetic mapping of a new set of microsatellite markers in a reference common bean (*Phaseolus vulgaris*) population BAT93 x Jalo EEP558. Geneti Mol Research 3:691-706.

Guimarães CM (1992) Características morfo-fisiológicas do feijoeiro (*Phaseolus vulgaris* L.) relacionadas com a resistência à seca. Tese (Doutorado) – Universidade Estadual de Campinas - UNICAMP, Campinas, 131 p.

Hoyos-Villegas V, Song Q, Kelly JD (2016) Genome-wide Association Analysis for Drought

Tolerance and Associated Traits in Common Bean. The Plant Genome 10:1-17.

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:E25. doi: 10.1093/nar/29.4.e25

Kamfwa K, Cichy KA, Kelly JD (2015) Genome-wide association analysis of symbiotic nitrogen fixation in common bean. Theor Appl Genet 128:1999–2017. doi: 10.1007/s00122-015-2562-5

Khaidizar MI, Haliloglu K, Elkoca E, Aydin M,Kantar F (2012) Genetic diversity of common bean (*Phaseolus vulgaris* L.) landraces grown in northeast Anatolia of Turkey assessedwith simple sequence repeat markers. J Field Crops 17:145–150.

Koinange EMK, Koinange EMK, Singh SP, et al (1996) Genetic control of the domestication syndrome in common-bean. Crop Sci 36:1037–1045. doi: 10.2135/cropsci1996.0011183X003600040037x

Kumar P, Gupta VK, Misra AK, et al (2009) Potential of Molecular Markers in Plant Biotechnology. Plant Omics Journal 2:141-162.

Kumar S, Verma AK, Sharma A, et al (2014) Phaseolin: A 47.5 kDa protein of red kidney bean (*Phaseolus vulgaris* L.) plays a pivotal role in hypersensitivity induction. Int Immunopharmacol 19:178–190. doi: 10.1016/j.intimp.2014.01.014

Lanna AC, Mitsuzono ST, Gledson T, et al (2016) Physiological characterization of common bean (*Phaseolus vulgaris* L.) genotypes, water- stress induced with contrasting response towards drought. AJCS 10:1–6.

Lesk C, Rowhani P, Ramankutty N (2016) Influence of extreme weather disasters on global crop production. Nature 529:84–87. doi: 10.1038/nature16467

Müller BS de F, Sakamoto T, Menezes IPP de, et al (2014) Analysis of BAC-end sequences in common bean (*Phaseolus vulgaris* L.) towards the development and characterization of long motifs SSRs. Plant Mol Biol 86:455–470. doi: 10.1007/s11103-014-0240-7

Müller BS de F, Pappas GJ, Valdisser PAMR, et al (2015) An Operational SNP Panel Integrated to SSR Marker for the Assessment of Genetic Diversity and Population Structure of the Common Bean. Plant Mol Biol Report 33:1697–1711. doi: 10.1007/s11105-015-0866-x

Mamidi S, Rossi M, Moghaddam SM, et al (2013) Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L. Heredity (Edinb) 110:267–76. doi: 10.1038/hdy.2012.82

Mcclean PE, Lee RK, Otto C, et al (2002) Molecular and Phenotypic Mapping of Genes

Controlling Seed Coat Pattern and Color in Common Bean (*Phaseolus vulgaris* L.). The Journal of Heredity 93:148–152.

McClean PE, Burridge J, Beebe S, et al (2011) Crop improvement in the era of climate change: An integrated, multi-disciplinary approach for common bean *(Phaseolus vulgaris*). Funct Plant Biol 38:927–933. doi: 10.1071/FP11102

Miklas PN, Kelly JD, Beebe SE, Blair MW (2006) Common bean breeding for resistance against biotic and abiotic stresses: From classical to MAS breeding. Euphytica 147:105–131. doi: 10.1007/s10681-006-4600-5

Mittler R (2006) Abiotic stress, the field environment and stress combination. Trends Plant Sci 11:15–19. doi: 10.1016/j.tplants.2005.11.002

Moghaddam SM, Mamidi S, Osorno JM, et al (2016) Genome-Wide Association Study Identifies Candidate Loci Underlying Agronomic Traits in a Middle American Diversity Panel of Common Bean. Plant Genome 9:1-21. doi: 10.3835/plantgenome2016.02.0012

Mukeshimana G, Butare L, Cregan PB, et al (2014) Quantitative trait loci associated with drought tolerance in common bean. Crop Sci 54:923–938. doi: 10.2135/cropsci2013.06.0427

Nass LL, Valois ACC, Melo IS, Valadares-Inglis MC (2001) Recursos genéticos e melhoramento de plantas, Rondonópolis: Fundação MT, 1183p.

Nepomuceno AL, Neumaier N, Farias JRB, Oya T (2001) Tolerância à seca em plantas. Biotecnologia Ciência e Desenvolvimento 23:12-18.

Neves LG, Davis JM, Barbazuk WB, Kirst M (2013) Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J 75:146–156. doi: 10.1111/tpj.12193

Omae H, Kumar A, Shono M (2012) Adaptation to high temperature and water deficit in the common bean (*Phaseolus vulgaris* L.) during the reproductive period. J Bot 2012:Article-803413. doi: 10.1155/2012/803413

Pérez-de-Castro AM, Vilanova S, Cañizares J, et al (2012) Application of genomic tools in plant breeding. Curr Genomics 13:179–95. doi: 10.2174/138920212800543084

Pérez-Vega E, Pañeda A, Rodríguez-Suárez C, et al (2010) Mapping of QTLs for morpho-agronomic and seed quality traits in a RIL population of common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 120:1367–1380. doi: 10.1007/s00122-010-1261-5

Perseguini JMKC, Chioratto AF, Zucchi MI, et al. (2011) Genetic diversity in cultivated carioca common beans based on molecular marker analysis. Genet Mol Biol 34:88-102.

Perseguini JMKC (2013) Estudo do desequilíbrio de ligação em *Phaseolus vulgaris* L. visando o mapeamento associativo de fatores bióticos e abióticos. Tese (Doutorado) –

Universidade Estadual de Campinas, Campinas. 134 p.

Perseguini JMKC, Oblessuc PR, Rosa JRBF, et al (2016) Genome-Wide Association Studies of Anthracnose and Angular Leaf Spot Resistance in Common Bean (*Phaseolus vulgaris* L.). PLoS One 11:1–19. doi: 10.1371/journal.pone.0150506

Polania J, Rao IM, Cajiao C, et al (2016) Physiological traits associated with drought resistance in Andean and Mesoamerican genotypes of common bean (*Phaseolus vulgaris* L.). Euphytica 210:17–29. doi: 10.1007/s10681-016-1691-5

Prasad PVV, Staggenborg S, Ristic Z (2008) Impacts of drought and/or heat stress on physiological, developmental, growth, and yield processes of crop plants, in Response of Crops to Limited Water: Understanding and Modeling Water Stress Effects on Plant Growth Processes: Advances in Agricultural Systems Modeling Series 1, eds L. R. Ahuja, V. R. Reddy, S. A. Saseendran, and Q. Yu (Madison, WI: ASA-CSSA-SSSA), 301–356.

Rangel PHN, Oliveira JP de, Costa JGC da, et al (2013) Banco ativo de germoplasma de arroz e feijão: passado, presente e futuro. Santo Antônio de Goiás: Embrapa Arroz e Feijão. Documentos 288, 59 p.

Repinski SL, Kwak M, Gepts P (2012) The common bean growth habit gene PvTFL1y is a functional homolog of Arabidopsis TFL1. Theor Appl Genet 124:1539–1547. doi: 10.1007/s00122-012-1808-8

Ribeiro RV, Santos MG, Souza GM, et al (2004). Environmental effects on photosynthetic capacityof bean genotypes. Piracicaba Pesq. agropec. bras., Brasília 39:615-623.

Ribeiro RV, Santos MG, Machado EC, et al (2008) Photochemical heat-shock response in common bean leaves as affected by previous water deficit. Russian Journal of Plant Physiology: a Comprehensive Russian Journal on Modern Phytophysiology 55: 350–358.

Rizhsky L, Liang H, Shuman J, et al (2004) When Defense Pathways Collide. The Response of Arabidopsis to a Combination of Drought and Heat Stress. Plant Physiol 134:1683–1696. doi: 10.1104/pp.103.033431.1

Rosales MA, Ocampo E, Rodríguez-Valentín R, et al (2012) Physiological analysis of common bean (Phaseolus vulgaris L.) cultivars uncovers characteristics related to terminal drought resistance. Plant Physiol Biochem 56:24–34. doi: 10.1016/j.plaphy.2012.04.007

Sansaloni CP (2012) Desenvolvimento e aplicações de DArT (Diversity ArraysTechnology) e genotipagem por sequenciamento (Genotyping-by-Sequencing) para análise genética em

Eucalyptus. Tese (Doutorado). Universidade de Brasília, Brasília. 145 p.

Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–13. doi: 10.1038/ng.3008

Sharma PN, Díaz LM, Blair MW (2013) Genetic diversity of two Indian common bean germplasm collections based on morphological and microsatellite markers. Plant Genetic Resources 11:121–130. doi: 10.1017/S1479262112000469.

Singh SP, Nodari R, Gepts P (1991) Genetic Diversity in Cultivated Common Bean: I. Allozymes. Crop Sci 31:19–23. doi: 10.2135/cropsci1991.0011183X003100010004x

Singh SP (1995) Selection for water - stress tolerance in interracial populations of commom bean. Crop Science 35:118-124.

Song Q, Jia G, Hyten DL, et al (2015) SNP Assay Development for Linkage Map Construction, Anchoring Whole Genome Sequence and Other Genetic and Genomic Applications in Common Bean. G3 Genes|Genomes|Genetics 5:g3.115.020594. doi: 10.1534/g3.115.020594

Thudi M, Bohra A, Nayak SN, et al (2011) Novel SSR Markers from BAC-End Sequences, DArT Arrays and a Comprehensive Genetic Map with 1,291 Marker Loci for Chickpea (*Cicer arietinum* L.). PLoS One. doi: 10.1371/journal.pone.0027275

Trapp JJ, Urrea C a., Cregan PB, Miklas PN (2015) Quantitative Trait Loci for Yield under Multiple Stress and Drought Conditions in a Dry Bean Population. Crop Sci 55:1596. doi: 10.2135/cropsci2014.11.0792

Valdisser PAMR, Pappas GJ, de Menezes IPP, et al (2016) SNP discovery in common bean by restriction-associated DNA (RAD) sequencing for genetic diversity and population structure analysis. Mol Genet Genomics 291:1277–1291. doi: 10.1007/s00438-016-1182-3

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522–530. doi: 10.1016/j.tibtech.2009.05.006

Villordo-Pineda E, González-Chavira MM, Giraldo-Carbajo P, et al (2015) Identification of novel drought-tolerant-associated SNPs in common bean (*Phaseolus vulgaris*). Front Plant Sci 6:1–9. doi: 10.3389/fpls.2015.00546

Willing EM, Hoffmann M, Klein JD, et al (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics 27:2187–2193. doi: 10.1093/bioinformatics/btr346

Yang SY, Saxena RK, Kulwal PL, et al (2011) The first genetic map of pigeon pea based on diversity arrays technology (DArT) markers. J Genet 90:103–109. doi: 10.1007/s12041-011-0050-5

Zhang L, Cai R, Yuan M, et al (2015) Genetic diversity and DNA fingerprinting in jute (Corchorus spp.) based on SSR markers. Crop J 3:416–422. doi: 10.1016/j.cj.2015.05.005

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and Prospects of Association Mapping in Plants. Plant Genome 1:5. doi: 10.3835/plantgenome2008.02.0089

Zuiderveen GH, Padder BA, Kamfwa K, et al (2016) Genome-Wide association study of anthracnose resistance in andean beans (*Phaseolus vulgaris*). PLoS One 11:1–17. doi: 10.1371/journal.pone.0156391

# CAPÍTULO I

Valdisser PAMR, et al. In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping.

Artigo submetido à revista BMC Genomics

## RESUMO

A diversidade genética, estrutura populacional e as análises de desequilíbrio de ligação (DL) são essenciais para a caracterização eficaz de uma coleção nuclear de feijoeiro comum. Um total de 6286 SNPs, derivados da metodologia DArTseq, forneceu uma plataforma robusta com uma alta densidade de SNPs para genotipagem de feijoeiro comum (1 SNP / 86,5 kb). Destes, 72,25% e 27,75% foram identificados em regiões gênicas e intergênicas, respectivamente. Quanto aos efeitos putativos dos SNPs, foi previsto que 7% tem impacto moderado e 0,05% impacto alto, estando relacionados com mecanismos de crescimento celular e estresses diversos. A estruturação por *pool* gênico (K = 2) foi predominante, seguida por outras subdivisões (K = 3 e K = 5), principalmente relacionada ao tipo de grão comercial. Um total de 83% e 91% de todos os SNPs foram polimórficos dentro dos *pools* gênicos Andino e Mesoamericano, respectivamente, e 26% diferenciaram os dois pools gênicos. A análise de diversidade genética revelou uma média de $H_E$ de 0,442, com 0,102 para Andinos e 0,168 para Mesoamericanos ($F_{ST}$ = 0,747 entre os *pools* gênicos); 0,440 para cultivar/linhagem e 0,448 para variedades tradicionais ($F_{ST}$ = 0,002). SNPs sob seleção foram identificados dentro dos *pools* gênicos entre os germoplasmas variedades tradicionais e cultivar/linhagem (Andino: 18; Mesoamericano: 69) e entre os *pools* gênicos (59 SNPs), predominantemente nos cromossomos 1 e 9. O DL corrigido para estrutura populacional e parentesco ($r^2_{SV}$) foi de aproximadamente 88 kb, variando de ~ 395 kb no Andino a ~ 130 kb no Mesoamericano. Além disso, 82% dos SNPs foram capturados em blocos de haplótipos cobrindo ~71% do genoma. Foi desenvolvido um painel de 560 SNPs em equilíbrio de ligação, capazes de capturar uma grande parte da diversidade dentro e entre os *pools* gênicos ($H_E$ = 0,401; $F_{ST}$ = 0,411). As propriedades genéticas dos SNPs identificados neste estudo são de grande relevância para propósitos de melhoramento genético do feijoeiro.

# In-depth genome characterization of a Brazilian common bean core collection using DArTseq high-density SNP genotyping

Paula Arielle M. R. Valdisser[1,5], Wendell J. Pereira[2], Jâneo E. de Almeida Filho[3], Bárbara S. F. Müller[2], Gesimária R. C. Coelho[1], Ivandilson P. P. de Menezes[4], João Paulo G. Viana[5], Maria I. Zucchi[5], Anna Cristina Lanna[1], Alexandre S. G. Coelho[6], Jaison P. de Oliveira[1], Alessandra C. Moraes[1], Claudio Brondani[1], Rosana P. Vianello[1].

**1** Embrapa Arroz e Feijão (CNPAF), Santo Antônio de Goiás-GO, Brasil;

**2** Programa de Pós-Graduação em Biologia Molecular, Universidade de Brasília (UnB), Brasília-DF, Brasil;

**3** Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF), Campos dos Goytacazes-RJ, Brasil;

**4** Laboratório de Genética e Biologia Molecular, Departamento de Biologia, Instituto Federal Goiano (IF Goiano), Urutaí-GO, Brasil;

**5** Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Estadual de Campinas (UNICAMP), Campinas-SP, Brasil;

**6** Escola de Agronomia, Universidade Federal de Goiás (UFG), Goiânia, GO, Brazil.

** Corresponding author:*

Rosana Pereira Vianello

Embrapa Arroz e Feijão

Phone: +55 62 3533 2154

Fax: +55 62 3533 2100

E-mail address: rosana.vianello@embrapa.br

*Key words:* common bean, DArTseq, diversity analysis, linkage disequilibrium, loci under selection

## ABSTRACT

Genetic diversity, population structure and linkage disequilibrium (LD) analyses are essential for the effective characterization of core collections. DArTseq technology provides a robust and high-density SNP genotyping platform for common bean genotyping (1 SNP/86.5 Kbp). Using this technology, 6,286 SNPs were genotyped. Of these, 72.25% were identified in genic and 27.75% in intergenic regions. Regarding their putative effects, 7% were predicted to have moderate and 0.05% high impact on processes, related to cell growth mechanisms and stress responses. The genetic subdivision associated to the common bean gene pool structure (K = 2) predominated, followed by other subdivisions that were mainly related to grain types (K = 3 and K= 5). A total of 83% and 91% of all SNPs were polymorphic within the Andean and Mesoamerican gene pools, respectively, and 26% differentiated the gene pools. Genetic diversity analysis revealed an average $H_E$ of 0.442 for the whole collection, of 0.102 for Andean and 0.168 for Mesoamerican gene pools ($F_{ST}$ = 0.747 between gene pools), of 0.440 for the group of cultivars and lines, and 0.448 for the group of landrace accessions ($F_{ST}$ = 0.002 between groups). SNPs outliers were identified within gene pools comparing landrace and cultivar/line germplasm groups (Andean: 18; Mesoamerican: 69) and between the gene pools (59 SNPs), predominantly on chromosomes 1 and 9. The LD extension estimate corrected for population structure and relatedness ($r^2_{SV}$) was ~ 88 kbp, ranging from ~ 395 kbp in the Andean to ~ 130 kbp in Mesoamerican gene pools. A fraction of 82% of SNPs was captured in haplotype blocks covering ~ 71% of the genome. A panel of 560 SNPs in linkage equilibrium, capturing a large portion of the intra and inter gene pools diversity ($H_E$ = 0.401, $F_{ST}$ = 0.411), was developed. The SNPs identified in this study and their genetic properties are of great relevance for breeding purposes.

## INTRODUCTION

It is estimated that approximately 150 plant species are grown directly for human consumption or animal feed worldwide, and 30 of them contribute to 95% of the calories and protein in the human diet (Füleky 2009). Legumes, along with grasses, are the main source of human food. Among legumes with edible dry seeds (pulses), over 80 species are widely cultivated, including the common bean (*Phaseolus vulgaris* L) (Tiwari et al. 2011). Common bean is a very important crop for food security and sustainable agriculture. The species is considered the most important grain legume available for human consumption (Broughton et al. 2003), being cultivated in 126 countries with an annual planted area estimated to be 30.6 million hectares (FAO 2014) and representing 37% of all legumes consumed in the world. To ensure the preservation of the extensive genetic diversity of common bean, national and international gene banks were created. The International Common Bean Gene Bank at CIAT (International Center for Tropical Agriculture, Colombia) has more than 37,000 accessions, of which approximately 90% are cultivated *Phaseolus vulgaris* varieties (CIAT 2014). In Brazil, the gene bank at Embrapa Rice and Beans has ~17,345 accessions, of which approximately 3.5% (~600 accessions) were selected to compose the core collection (acronym CONFE), which is made up of three strata: a) landraces from Brazil; b) cultivars/lines improved in Brazil; and c) introduced cultivars/lines, all of Andean and Mesoamerican origin. The seed samples are publicly available for research institutions in Brazil and abroad and are stored at Global Seeds Banking of Svalbard, located in Longyearbyen, Norway.

There is consensus regarding the predominant genetic structure of common bean in the Andean and Mesoamerican gene pools (Rossi et al. 2009; Mamidi et al. 2011), due to a divergence estimated to have occurred since 165,000 years ago (Schmutz et al. 2014). Genes related to agronomic traits of great interest to current breeding programs, such as flowering, plant height, and nitrogen metabolism, were identified as being under selection during the domestication process (Schmutz et al. 2014). The common bean landraces from Brazil, a secondary center of domestication, are adapted to diverse soil and climate conditions and present broad genetic diversity (Burle et al. 2010). It is expected that several adaptive mechanisms selected over generations of domestication remain unknown (McCouch et al. 2013) and can be used as an important source of useful genes for breeding programs (Dwivedi et al. 2016). A large proportion of plant genetic resources remains unexplored. This situation is changing due to efforts in breeding programs to increase the available genetic diversity among the set of genitors used in crosses (Blair and Lorigados 2016; Rodriguez et al. 2016).

Through pre-breeding programs, work to identify favorable alleles of genes related to important agronomic traits in wild germplasm and landraces, with subsequent incorporation into improved crops, has been reviewed (Porch et al. 2013). The availability of common bean reference genomes (Schmutz et al. 2014; Vlasova et al. 2016), in addition to predicted functions for thousands of genes, extends the possibilities for marker-assisted selection, and to increase the efficiency of genetic breeding programs (Meziadi et al. 2016).

Molecular markers have been very helpful in efforts to detect gaps and redundancies in germplasm collections (Cruz et al. 2013), to elucidate the genetic diversity in both wild germplasm (Blair et al. 2012) and in landraces and cultivars/lines (Burle et al. 2010; Cardoso et al. 2014), to explore the effects of selection in the domestication process and to evaluate the dynamics of gene flow and genetic structure due to geographic distribution (Papa and Gepts 2003; Papa et al. 2007). Many of these studies were conducted using SSR markers (Blair et al. 2009; Gill-Langarica et al. 2011; Perseguini et al. 2011; Müller et al. 2014, 2015). In recent years, SNP markers have been increasingly developed and applied in common bean genetic analysis (Blair et al. 2013; Müller et al. 2015; Valdisser et al. 2016). Based on 131 SNPs, Rodriguez et al. (2016) analyzed a set of 577 wild and domesticated common bean accessions and drew conclusions about the genetic structure along the domestication sites and identified geographic regions that were hotspots of genetic diversity. More recently, a 6,000 SNP chip was developed (BARCBean6K_3) and successfully used in linkage and genome-wide association mapping studies (Song et al. 2015; Cichy et al. 2015).

High-density genotyping, combining genome complexity reduction with next-generation sequencing (NGS), allows the identification of an almost unlimited number of SNPs for any species at low cost. The strategies of restriction site-associated DNA sequencing (RADseq, Willing et al. 2011) and genotyping by DNA sequencing (GBS, Elshire et al. 2011) allow researchers to identify and genotype thousands of SNPs in several plant species, including common bean (Valdisser et al. 2016; Schröder et al. 2016). The Diversity Arrays Technology (DArT), based on genome complexity reduction and SNP detection through hybridization of PCR fragments (Jaccoud et al. 2001), has been used in the construction of dense linkage maps, mapping quantitative trait loci (QTL), genome-wide association studies (GWAS), and studies of genetic diversity and population structure (Raman et al. 2014; Hahn and Wurschum 2014; Ren et al. 2015). In legumes, DArT markers were used to detect QTLs associated with the genetic resistance to angular leaf spot (Briñez et al. 2012; Oblessuc et al. 2013). In the last years, DArT technology was modified to incorporate the advantages of the

genotyping by sequencing approach (DArTseq™; Cruz et al. 2013; Zou et al. 2014; Sánchez-Sevilla et al. 2015).

In this study, DArTseq derived SNPs were used for the genetic analysis of a common bean germplasm collection of Andean and Mesoamerican origins, being each origin further stratified into cultivars/lines and landraces groups. This study also made advances in the detection and characterization of genomic regions with signals of selection imposed by the domestication and breeding of common bean in Brazil. In addition, a set of SNPs with high discriminatory value between gene pools, as well as between groups (landraces and cultivars/lines) within gene pools, was proposed for routine use for the characterization of gene bank accessions and in breeding programs.

## MATERIAL AND METHODS

### Plant material

A total of 188 common bean accessions, including 91 landraces and 97 Brazilian and international cultivars/lines belonging to the Andean and Mesoamerican gene pools, were used (Supplementary Material 1). The accessions were planted in a greenhouse and multiplied via selfing in order to ensure homogeneity for genetic analysis. DNA from individual plants was extracted using the Invisorb Spin Plant Mini Kit (Stratec Molecular, Berlin, Germany), followed by shipment to a DArTseq analysis facility (DArT Pty Ltd., Bruce, Australia).

### Genotyping using DArTseq

DArTseq™ represents a combination of DArT complexity reduction methods based on methyl filtration and next-generation sequencing platforms (Kilian et al. 2012). The technology was optimized for common bean considering both the size of the representation and the fraction of the genome selected for analysis. The complexity reduction method was based on PstI-MseI. DNA samples were processed before and after sequencing as described by Sánchez-Sevilla et al. (2015). The amplification products were sequenced on the Illumina Hiseq2000 platform. Approximately 2,000,000 sequences per barcode/sample were identified and used in marker calling. Identical sequences were collapsed into fastqcall files. These files were used in the secondary pipeline for DArT PL's proprietary SNP-calling algorithms (DArTsoft-seq).

**Structural and functional characterization of SNPs**

Genomic regions flanking SNPs were aligned against the reference genome of *P. vulgaris* v 1.0 (Schmutz et al. 2014) using BLASTN with an *E-value* ≤ 1.0E-25 (Altschul et al. 1997). Annotation and prediction of effect were performed using the SnpEff v 4.2 (Cingolani et al. 2012) based on the Phytozome database (Goodstein et al. 2012). SNPs with putative effects predicted to be moderate or high were functionally annotated using the Blast2GO tool v 3.2 (Conesa et al. 2005). The transcripts were characterized using Gene Ontology terms (Ashburner et al. 2000; Consortium 2015) and the graphs of the terms were filtered according to the node score (the node score is calculated for each GO term in the graphs and takes into account the topology of the ontology and the number of sequences belonging to a given node). KEGG (available in Blast2GO v 3.2) provided the Enzyme Code (EC) for metabolic pathways.

**Diversity analysis and genetic structure**

The genetic structure, based on the Bayesian clustering approach, was implemented by Structure v 2.3.4 (Pritchard et al. 2000). This analysis was conducted using SNPs in linkage equilibrium (LE; $r^2 < 0.5$) identified using SVS software (Golden Helix) through the LD Prunning command. A population number (K) ranging from 1 to 20, with 20 interactions each, was assumed. The admixture model was applied using a 500,000 burn-in periods followed by 1,000,000 MCMC replications. The most likely K was determined, as proposed by Evanno et al. (2005) using Structure Harvester v 0.6.93 (Earl and Vonholdt 2012), followed by analysis with CLUMPP v 1.1.2 (Jakobsson and Rosenberg 2007). The organization chart was generated in R v 3.1.3 (R Development Core Team 2015). Discriminant Analysis of Principal Components (DAPC, Jombart et al. 2010) was performed using the Adegenet package (Jombart and Ahmed 2011) to provide further support for the identified population groups. The dendrogram was constructed using the neighbor joining (NJ) method implemented by Mega v 5 (Tamura et al. 2011), based on a matrix calculated by Simple Matching Dissimilarity with 1,000 bootstrap interactions (Darwin 6.0.10; Perrier and Jacquemoud-Collet 2006). Analysis of genetic diversity was performed in GenAlex 6.501 (Peakall and Smouse 2012) using SNPs with a call rate ≥ 75% (5,531 SNPs).

**Genetic diversity and differentiation along the genome**

The $F_{ST}$ for each window of the genome (Weir and Cockerham 1984), Tajima's D (Tajima 1989), diversity from Nei ($\pi$, average pairwise differences among individuals chosen randomly from the sample population, Nei and Li 1979; Nei 1987), nucleotide diversity within the population (Hudson et al. 1992; Wakeley 1996) and Watterson's $\theta$ ($\theta_W$, estimation of population mutation rate calculated on the basis of the number of segregating sites, Watterson 1975) were estimated using non-overlapping 100 Kb sliding windows in PopGenome package (Pfeifer et al. 2014).

In total, 5,273 SNPs with call rate ≥ 75% and distributed in the 11 chromosomes were used in 2,494 windows of 100 Kb. Finally, the $F_{ST}$, Tajima's D and $\theta_W$ statistics were calculated using loess smoothing of non-overlapping 100 Kb sliding windows, to visualize patterns of variation across the two different gene pools in common bean (Andean Vs. Mesoamerican) and also between Cultivars/Lines and Landraces in each gene pool.

**SNPs outliers**

The SNPs outliers were detected using two methods: 1) Method proposed by Foll and Gaggiotti (2008) implemented in the BayeScan 2.0, which estimates the probability of each locus to be under selection using MCMC. The analysis was performed using 20 pilot runs with 5,000 interactions, burn-in of 100,000 followed by 100,000 interactions ("thinning interval" equal to 20 and sample size of 5,000), with a probability > 1. The analysis was performed three times to ensure robustness and only the outliers loci identified across all the runs were considered. 2) Hierarchical method of Excoffier et al. (2009) implemented in Arlequin v 3.5.2.2 (Excoffier and Lischer 2010), which identified *outliers* by comparing the levels of genetic diversity and differentiation among populations. The hierarchical island model was simulated with two groups (Andean and Mesoamerican), two demes per group with 20,000 simulations to generate an $F_{ST}$ joint distribution versus heterozygosity. Those loci that fall outside the 95% confidence interval were considered *outliers*.

**Linkage disequilibrium (LD) and haplotype blocks**

LD was estimated using SNPs with MAF ≥ 0.05 and the pairwise LD measures were calculated by the usual method ($r^2$) and corrected for bias due to population structure (K

= 2) and relatedness ($r^2_{SV}$) using the LDcorSV R-package (Desrousseaux et al. 2013). The Genetic Relationship Matrix (GRM) was estimated using the algorithm proposed by Yang et al. (2010) using GCTA software (Yang et al. 2011). LD decay (half of the maximum value) was explained by the nonlinear model proposed by Hill and Weir (1988) and adjusted to the nls function in R (R Development Core Team 2015). Haplotypic blocks were identified using Haploview 4.2 (Barrett et al. 2005) based on the confidence interval method described by Gabriel et al. (2002): MAF ≥ 0.05 and call rate ≥ 75%. Heterozygous loci were considered missing data.

**Genetic diversity distribution based on temperature and rainfall maps**

**Genetic diversity of landraces heatmap:** Spatial analysis of genetic diversity ($H_E$) was performed applying an individual-centered approach as described by Manel et al. (2007) and adapted from the Wombling method (Womble 1951). $H_E$ estimates were obtained using a hierarchical procedure, with a 150 km neighborhood grid used to avoid spatial autocorrelation between groups. In cases in which only one accession was represented in a given region, $H_E$ represents diversity only for this accession. This analysis was performed using the "sHe" function of the R package "biotools" (Silva 2016).

**Georeferencing landraces in thematic maps of climate in Brazil:** The Brazilian maps were derived from the Brazilian Institute of Geography and Statistics (IBGE, Department of Cartography, 2016). Data from rainfall and climate/temperature were obtained from the Institute of Forest Research and Studies (IPEF). The software ArcGIS, based on Geographic Information System (SIG), was used to define areas on the maps. Landraces were geographically placed on the maps using the associated coordinate information.

**RESULTS**

**Genotyping using DArTseq**

The 188 beans analyzed by DArTseq comprised a mini core group derived from the Brazilian common bean core collection (600 accessions) and are representative of the most genetically diverse accessions identified by microsatellite markers analysis (data not shown). For the SNP markers generated in DArTseq, robust parameters were implemented

(Supplementary Material 2): (1) *Call rate* ranging from 0.50 to 1.00, with an average of 92%, in other words, only ~8% missing data for each marker; and (2) high reproducibility, ranging from 96.85 to 100%. The averages of homozygotes and heterozygotes were 0.88 and 0.04, respectively. Polymorphism content (PIC) ranged from 0.23 to 0.5, with an average of 0.44, and the minor-allele frequency (MAF) ranged from 0.13 to 0.5, with an average of 0.35. A total of 6,286 SNPs were obtained from 181 accessions, of which only seven genotypes (3.72%) failed to generate sequence information.

**Structural and functional characterization of SNPs**

From the 6,286 SNP-flanking regions, 5,961 (94.82%) showed alignment in the genome, of which 5,311 (89.09%) aligned to a single region and 650 (10.90%) presented multiple alignments (ranging from two to 88). The sequences aligned to the 11 chromosomes and 12 scaffolds. The average number of SNPs per chromosome was 541, ranging from 389 on chromosome 4 to 792 on chromosome 2 (Table 1, Figure 1). Based on *Phytozome* database, 15 SNPs aligned with 12 scaffolds and 325 SNPs did not align with the genome. An average of one SNP every 86,503 base pairs was estimated. Regarding the polymorphism types, transition (Ts) was the most abundant (3,299 events, 55.30%), being most frequently cytosine to thymine (923), followed by transversions (Tv) with 2,655 events (44.70%). The ratio of Ts/Tv was 1.24. Most SNPs were in genes (72.25%), including a window of 5 kb upstream and downstream of the gene.

**Table 1.** SNPs-DArTseq distribution by common bean chromosomes.

| Chromosome | Number of SNPs | Chromosome size (kbp)* | Mean of SNP per Mbp | Number of genes** |
|---|---|---|---|---|
| 1 | 533 | 52183.50 | 10.21 | 2694 |
| 2 | 792 | 49033.70 | 16.15 | 3338 |
| 3 | 623 | 52218.60 | 11.93 | 2973 |
| 4 | 389 | 45793.20 | 8.49 | 1789 |
| 5 | 431 | 40237.50 | 10.71 | 1863 |
| 6 | 532 | 31973.20 | 16.64 | 2221 |
| 7 | 537 | 51698.40 | 10.39 | 2812 |
| 8 | 656 | 59634.60 | 11.00 | 2932 |
| 9 | 523 | 37399.60 | 13.98 | 2633 |
| 10 | 401 | 43213.20 | 9.28 | 1659 |
| 11 | 529 | 50203.60 | 10.54 | 2168 |
| Total | 5946 | 513589.10 | 11.58 | 27352 |

* Schumtz et al. 2014

** *P. vulgaris* v 1.0 (*Phytozome* database)

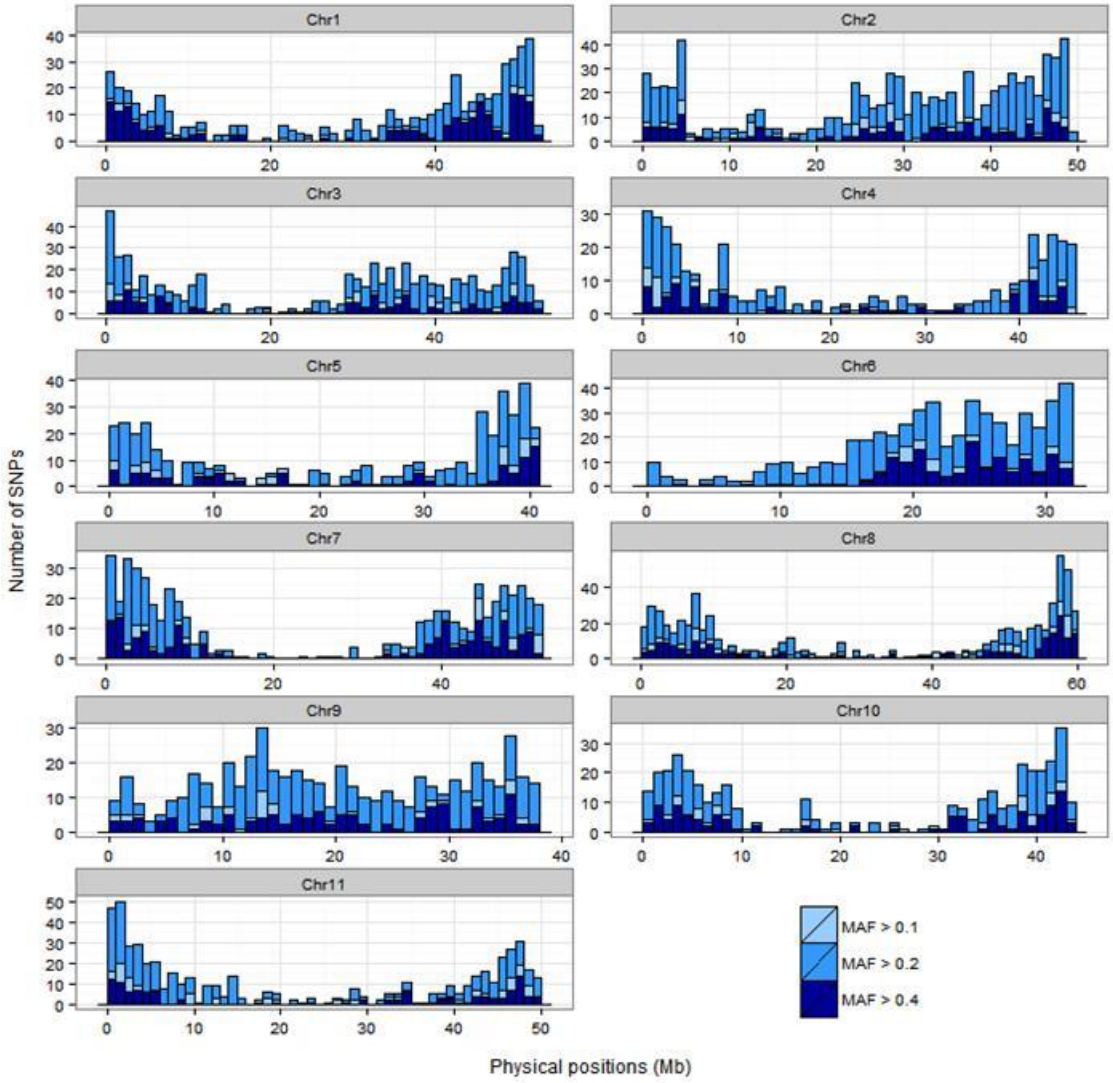



**Figure 1.** Distribution and positioning of the SNPs-DArTseq in *P. vulgaris* chromosomes.

A total of 12,217 effects were predicted for 5,954 SNPs, providing information on the location of all isoforms, genic, and intergenic regions. The predicted effects were of modifier type (77.8%), low impact (14.22%), moderate impact (7.92%), and high impact (0.05%). Most SNPs with predicted effects were observed in genic regions (6,950), of which 20.82% and 17.30% were observed within exons and introns, respectively, with the remaining in non-translated regions. In genic flanking sequences (5 kb window) 5,267 effects were identified, of which 58.21% and 41.79% occurred in downstream and upstream regions, respectively. SNP effects categorized as moderate and high, which are more likely to be under strong selective pressure, were identified in 901 transcripts, of which 810 were mapped and 777 were fully noted (Supplementary Material 3). These genes were related to a variety of mechanisms, such as plant development and multiple stress response pathways (Figure 2). Among the 777 annotated sequences, 359 were identified as enzymes, mainly transferases (129) and hydrolases (125; Supplementary Material 4). Genes involved in metabolic pathways are described in Supplementary Material 5.

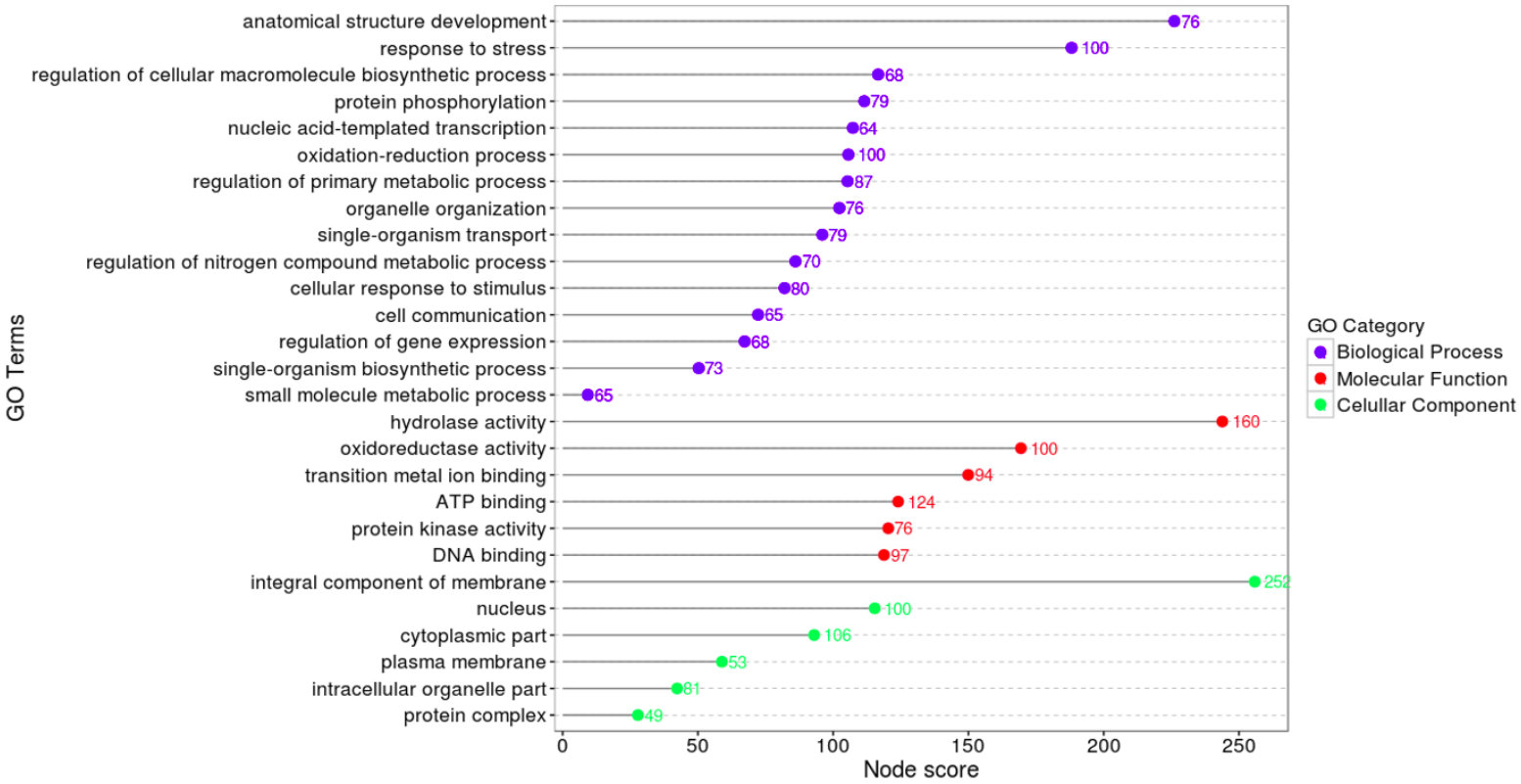
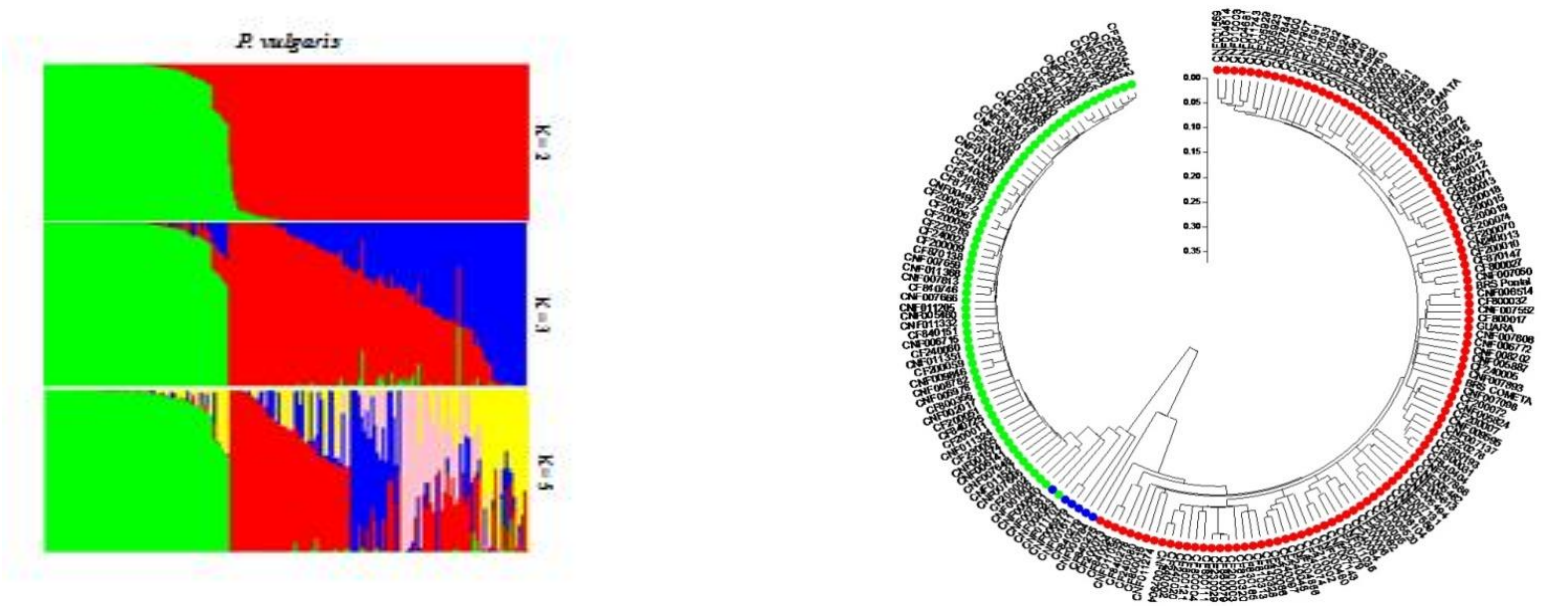



**Figure 2.** Functional annotation of SNPs with high and moderate impact predicted by SnpEff. The terms were filtered according to the node score. The numbers represent the amount of transcripts related to each term.

**Germplasm genetic structure**

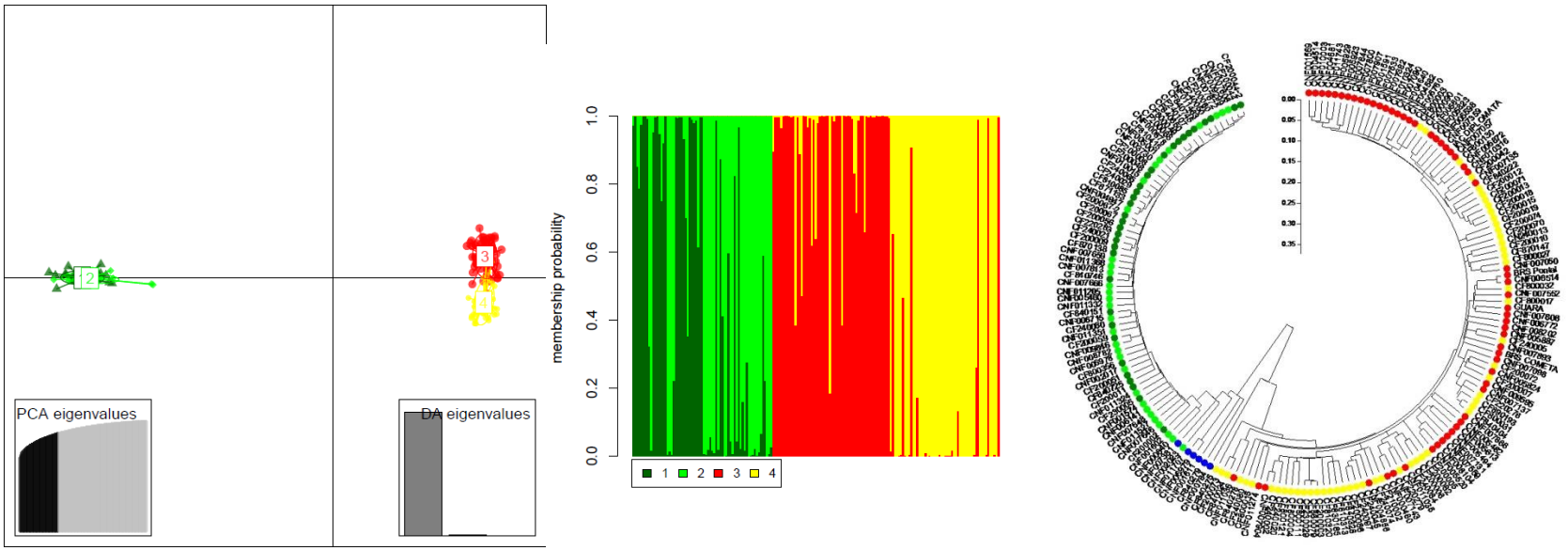
Population structure analysis performed using 580 SNPs in LE revealed K = 2 as the most likely, with the subdivision in Andean (64) and Mesoamerican (111) gene pools and six genotypes (3.87%) as admixture (Figure 3A). Five of the genotypes with admixture (ranging from 62 to 69%) were mainly from Andean origin: four cultivars/lines developed by international institutions and one Brazilian landrace from Rio Grande do Sul state (white or brindle grains). The genotype with a predominance of Mesoamerican germplasm (~ 65%) is a cultivar/line with brindle grain type from Russia (CNF000784). For K = 3, the Mesoamerican groups were fragmented in two (M1 and M2) in addition to 45 genotypes with admixture. The M1 group was composed by 46 accessions ($q \geq 0.7$) of which 74% (34) were black grain types from Brazilian and international cultivars/lines. M2 contained 20 Brazilian genotypes ($q \geq 0.7$), 17 landraces and three cultivars/lines, without grain type prevalence. For K = 5, an additional fragmentation within the Mesoamerican gene pool was observed (M1, M2, M3, and M4). M1 was formed by 28 genotypes, 20 cultivars/lines and eight landraces, with a predominance (82.14%) of the black grain type. M2 contained seven accessions from Brazil (six landraces and one cultivar/line), of which 43% were of the yellow grain type. The M3 group was represented by six Brazilian genotypes (four landraces and two cultivars/lines) with a *carioca* commercial grain type. Finally, M4, with eight genotypes (six landraces and three cultivars/lines), had different types of grain (62.5% of brown and red type). The dendrogram shows the same division found in Structure (K = 2; Figure 3B).

The tools implemented in DAPC revealed a more complex population structure in the Mesoamerican by landraces and lines/cultivars (Figure 4).




**A** **B**

**Figure 3.** Population structure. A) Population structure inferred by the Bayesian approach based on SNPs-DArTseq for K = 2, 3 and 5. Each individual is represented by a vertical line that is divided into colored segments based on the proportion of the division identified for 2, 3 and 5 subpopulations. The groups include: A1 (green): Andean; M1 (red): Mesoamerican with a prevalence of 82.14 % of the accessions with black commercial grain; M2 (blue): seven accessions from Brazil (six landraces and one cultivar/line), of which 43% was of yellow grain type; M3 (pink): six Brazilian genotypes (four landraces and two cultivars/lines) with carioca commercial grain type; M4 (yellow): nine genotypes (six landraces and three cultivars/lines) with different type of grain (62.5% of brown and red type). B) Dendrogram showing the division between the two gene pools: Andean (green), Mesoamerican (red), and Admixture (blue).




**A** **B** **C**

**Figure 4.** Population structure. A) DAPC using the set of 5531 SNPs, 1: Andean cultivars/lines (dark green); 2: Andean landraces (light green); 3: Mesoamerican cultivars/lines (red); 4: Mesoamerican landraces (yellow). B) DAPC showing the separation between Mesoamerican cultivars/lines (red) and Mesoamerican landraces (yellow). C) Dendrogram showing the division between the two gene pools: Andean cultivars/lines (dark green), Andean landraces (light green), admixture (blue), Mesoamerican cultivars/lines (red), and Mesoamerican landraces (yellow).

**Diversity analysis**

The analysis of genetic diversity (5,531 SNPs) revealed a total of 4,590 (82.99%) and 5,020 (90.76%) polymorphic SNPs in the Andean and Mesoamerican groups, respectively. The average estimates of $H_E$ for the Andean and Mesoamerican were 0.102 and 0.168, respectively, and 0.442 for all samples (Table 2). $H_O$ values were estimated to be 0.04 for the Andean, 0.035 for the Mesoamerican, and 0.037 for the entire set of samples. The inbreeding coefficient (F) over the total accessions was 0.908 (Andean: 0.561; Mesoamerican: 0.652). A set of 1,452 SNPs (~26%) distinguished the Andean (511 unique alleles) to the Mesoamerican (941 unique alleles) gene pools. The estimated power of exclusion was high, approaching 100% (a set of only 28 SNPs differentiate all genotypes), and the combined PI was estimated to be 0.00 (Table 2), with individual values ranging from 0.375 to 0.669. The estimated overall $F_{ST}$ between Andean and Mesoamerican was 0.747 ($p > 0.001$).

High numbers of polymorphic SNPs were identified for Mesoamerican (87.51%) and Andean (88.39%) cultivars/lines compared to the landraces (Mesoamerican = 90.78%. Andean = 73.49%). The $H_E$ values for the Mesoamerican group were 0.177 and 0.185 for cultivars/lines and landraces, respectively, while for the Andean, the corresponding values were 0.145 and 0.099. In both gene pools, the estimates of $F_{ST}$ between cultivars/lines and landraces were 0.031 ($p > 0.001$) and 0.012 ($p > 0.002$) for Mesoamerican and Andean, respectively. Within the Andean gene pool, cultivars/lines presented 1,217 private alleles, while in landraces it was 533 (Table 3).

**Table 2.** Genetic diversity and divergence within Andean and Mesoamerican gene pools.

| Group | S | P | NAP | $H_O$ (SE) | $H_E$ (SE) | F (SE) | $F_{ST}$ (SE) | $F_{IS}$ (SE) | $F_{IT}$ (SE) | PI | PE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Andean | 64 | 82.99% | 511 | 0.040 ±0.002 | 0.102 ±0.002 | 0.561 ±0.006 | 0.747 ±0.001 | 0.822 ±0.001 | 0.955 ±0.0031 | 1.05E-249 | 1 |
| Mesoamerican | 111 | 90.76% | 941 | 0.035 ±0.001 | 0.168 ±0.002 | 0.652 ±0.006 | | | | 0 | 1 |
| Total | 175 | 100 | - | 0.0373 ±0.001 | 0.4425 ±0.001 | 0.9082 ±0.003 | | | | 0 | 1 |

The sample size (*S*), percentage of polymorphic loci (*P*), number of private alleles (NAP), observed heterozygosity ($H_O$),gene diversity ($H_E$), inbreeding coefficient (*F*), genetic differentiation ($F_{ST}$), fixation index ($F_{IS}$), total inbreeding ($F_{IT}$), probability of identity (PI), probability of exclusion (PE), and standard deviations (SE) are presented.

**Table 3.** Genetic diversity and divergence among cultivars/lines and landraces of the Andean and Mesoamerican gene pools.

| Group | | S | P | NAP | $H_O$ (SE) | $H_E$ (SE) | F (SE) | $F_{ST}$ (SE) | $F_{IS}$ (SE) | $F_{IT}$ (SE) |
|---|---|---|---|---|---|---|---|---|---|---|
| Mesoamerican | Cult/Lines[a] | 57 | 87.51% | 463 | 0.038 ±0.001 | 0.177 ±0.003 | 0.652 ±0.006 | 0.031 ±0.001 | 0.836 ±0.001 | 0.841 ±0.001 |
| | Landraces | 54 | 90.78% | 627 | 0.040 ±0.001 | 0.185 ±0.002 | 0.646 ±0.006 | | | |
| | Total | 111 | 100.00% | - | 0.039 ±0.001 | 0.185 ±0.002 | 0.652 ±0.006 | | | |
| Andean | Cult/Lines[a] | 31 | 88.39% | 1217 | 0.046 ±0.002 | 0.145 ±0.002 | 0.647 ±0.007 | 0.012 ±0.002 | 0.738 ±0.001 | 0.741 ±0.001 |
| | Landraces | 33 | 73.49% | 533 | 0.050 ±0.002 | 0.099 ±0.002 | 0.377 ±0.007 | | | |
| | Total | 64 | 100.00% | - | 0.048 ±0.002 | 0.123 ±0.002 | 0.561 ±0.007 | | | |

The sample size (*S*), percentage of polymorphic loci (*P*), number of private alleles (NAP), observed heterozygosity ($H_O$),gene diversity ($H_E$), inbreeding coefficient (*F*), genetic differentiation ($F_{ST}$), fixationindex ($F_{IS}$), total inbreeding ($F_{IT}$), and standard deviations (SE) are presented.
[a]Cult/Lines: cultivars/lines

**Diversity analysis and distribution of genetic differentiation in the genome**

$F_{ST}$ was high between the gene pools for the majority of the chromosomes, with the highest level at chromosomes 1 and 9 (Figure 5A). The overall differentiation among cultivars/lines and landraces was lower for the Andean germplasm ($F_{ST}$ = 0.0082) compared to the Mesoamerican ($F_{ST}$ = 0.0218; Figure 6A). The average value of $\pi$ over the whole population, based on 5,241 SNPs, was 0.0171 (±0.001) and was greatly reduced for the Andean ($\pi$ = 0.0017 ± 0.0002, 3,889 SNPs) and Mesoamerican ($\pi$ = 0.0045 ± 0.0006; 3,957 SNPs) groups (Table 4). These values were consistent with an MAF > 0.3 in the whole population of 4,210 (80%) SNPs and an MAF < 0.1 for about half of SNPs into the Andean and Mesoamerican groups. Considering the germplasm stratum, the $\pi$ value was 0.0044 (± 0.0005) for the cultivars/lines and 0.0043 for the landraces of Mesoamerican origin, with a similar distribution of SNPs into MAF classes (Figure 7B). Reduced values were observed for the cultivars/lines (0.0022) and landraces (0.0013) of Andean origin, probably due to the additional set of SNPs with MAF > 0.1 and ≤ 0.2 (Figure 7B).

The Watterson's Mean θ ($\theta_W$) for all individuals was 0.00071 (± 0.00001), with a lower value estimated for the 33 Andean landraces (0.00058). The $\theta_W$, Tajima's D, and $F_{ST}$ estimates were highly variable across the *P. vulgaris* genome (Figures 5 and 6) and regions that displayed high values of $F_{ST}$ also presented elevated LD (data not shown) and reduced $\theta_W$ and Tajima's D (Figures 5A and 5B), mostly in centromeric regions. For the Andean accessions, negative Tajima's D values were observed for all chromosomes except for chromosome 4 (Supplementary Material 6), which could indicate that positive selection in the Andean group is driving divergence between the gene pools, as evidenced by the correlation of centromeres and regions of elevated $F_{ST}$ (Figure 5C). For the Mesoamerican group, Tajima's D values were variable across the genome and some regions, such as in the chromosome 5 approximately 10-20 Mbp, presented high Tajima's D values and low $F_{ST}$, indicating balancing selection (Figures 5A and 5C). Conversely, in the same chromosome 5, a region near 30 Mbp had a low Tajima's D value and high $F_{ST}$, indicating possible positive selection (Figures 5A and 5C).

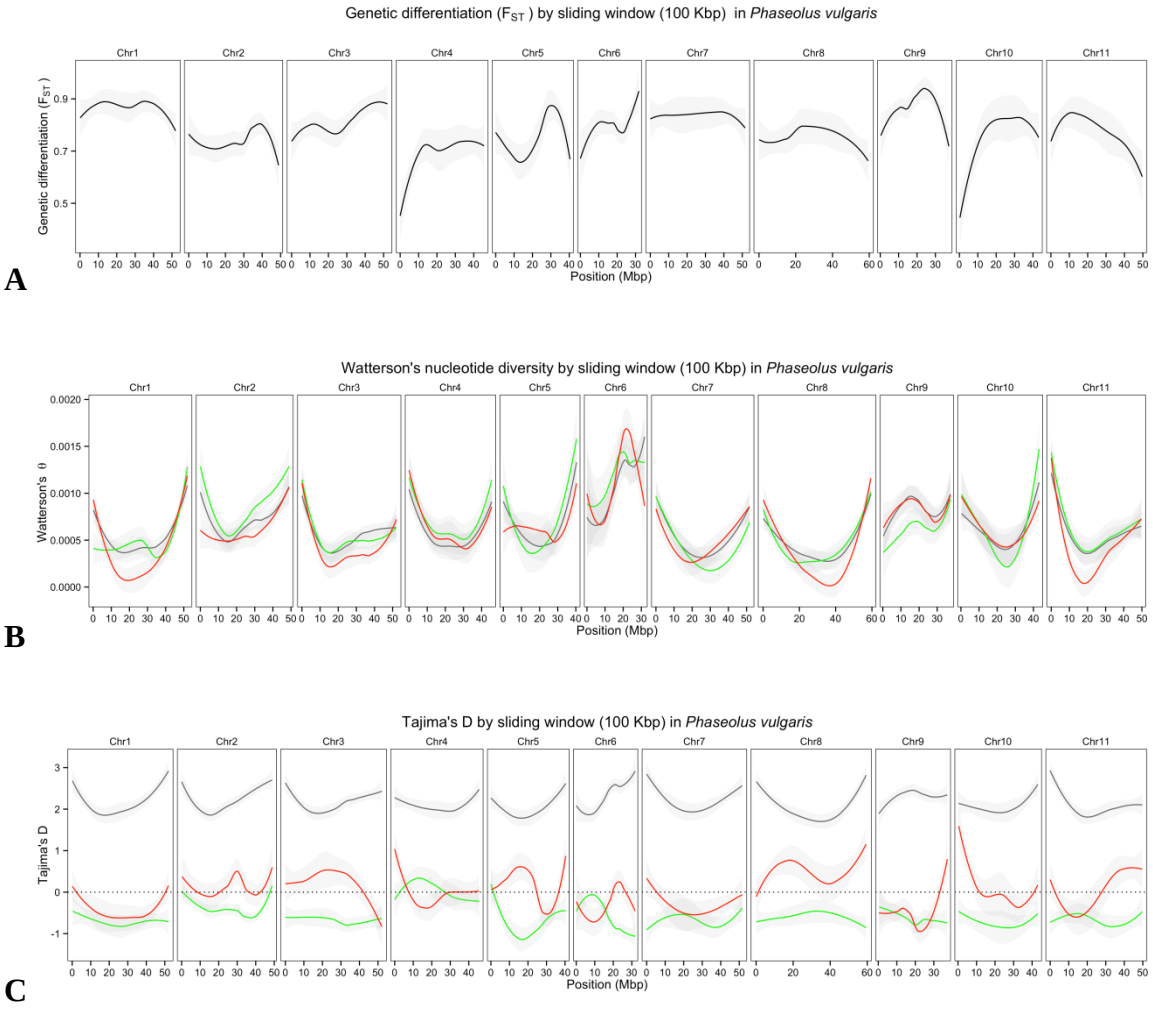
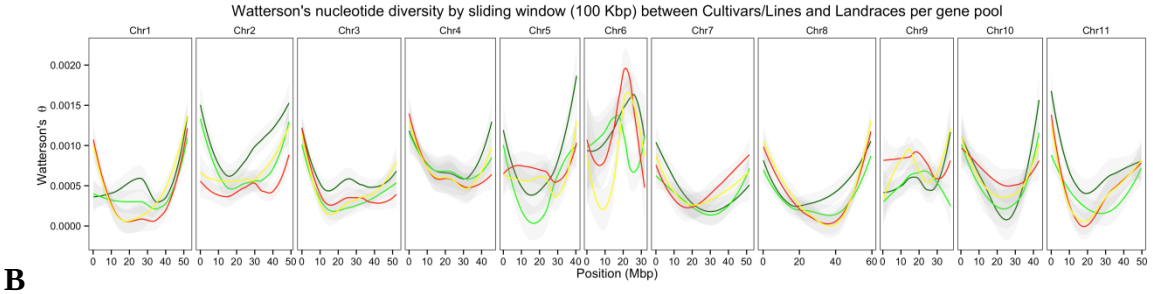
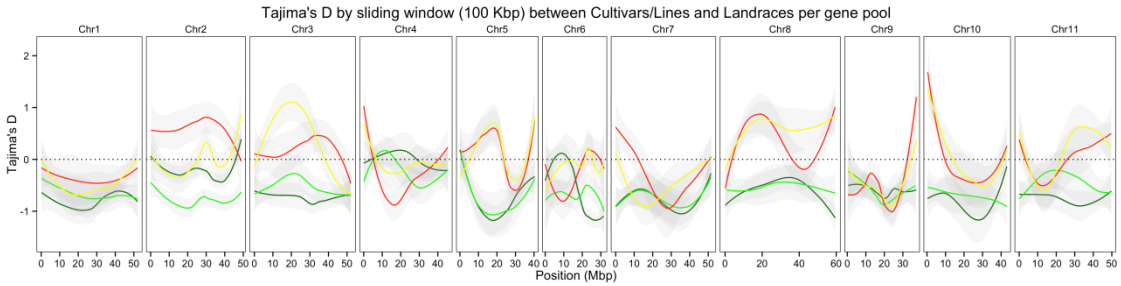



**Figure 5.** Genome-wide loess curves for genetic differentiation ($F_{ST}$) (A), Watterson's θ ($\theta_W$) (B), and Tajima's D (C) for all 11 chromosomes in the *P. vulgaris* genome for each gene pool. $F_{ST}$ is given as an average across all pairwise comparison between Andean and Mesoamerican. Results of Tajima's D and $\theta_W$ are given for each gene pool separately, Mesoamerican (red) and Andean (green), and also estimate for the whole population (grey). $F_{ST}$, Tajima's D and $\theta_W$ related summary statistics were calculated for each 100 kb non-overlapping sliding window.


**A**




**B**




**C**

**Figure 6.** Genome-wide loess curves for genetic differentiation ($F_{ST}$) (A), Watterson's θ ($\theta_W$) (B), and Tajima's D (C) for all 11 chromosomes in the *P. vulgaris* genome for each group. $F_{ST}$ is given as an average across all pairwise comparisons between Andean cultivars/lines and landraces (green), and between Mesoamerican cultivars/lines and landraces (red). Results of Tajima's D and $\theta_W$ are given for each group separately, Andean cultivars/lines (dark green) and landraces (green), and Mesoamerican cultivars/lines (red) and landraces (yellow). $F_{ST}$, Tajima's D and $\theta_W$ related summary statistics were calculated for each 100 kb non-overlapping sliding window.

**Table 4.** Summary of the *P. vulgaris* genome-wide diversity based on SNPs-DArTseq.

| Gene Pool | Group | N | S | Θw | SE – θw | NDw | SE - NDw | π | SE - π | P | M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Andean** | Cult/Lines[a] | 31 | 3506 | 0.000777 | 0.000015 | 0.000554 | 0.000014 | 0.002171 | 0.000281 | 3858 | 1415 |
| | Landraces | 33 | 2647 | 0.000580 | 0.000013 | 0.000386 | 0.000011 | 0.001357 | 0.000228 | 3190 | 2083 |
| | Total | 64 | 3889 | 0.000728 | 0.000013 | 0.000471 | 0.000012 | 0.001781 | 0.000266 | 4364 | 909 |
| **Mesoamerican** | Cult/Lines[a] | 57 | 3283 | 0.000641 | 0.000014 | 0.000665 | 0.000020 | 0.004386 | 0.000488 | 4177 | 1096 |
| | Landraces | 54 | 3460 | 0.000667 | 0.000014 | 0.000677 | 0.000019 | 0.004330 | 0.000566 | 4340 | 933 |
| | Total | 111 | 3957 | 0.000667 | 0.000012 | 0.000685 | 0.000019 | 0.004541 | 0.000572 | 4778 | 495 |
| **Whole** | All | 181 | 5241 | 0.000713 | 0.000010 | 0.002038 | 0.000029 | 0.017125 | 0.001540 | 5273 | 0 |

The number of samples (N), number of segregating sites (S), Watterson's nucleotide diversity (θw ), nucleotide diversity within (NDw), diversity from Nei (π), number of polymorphic SNPs (P), number of monomorphic SNPs (M), and standard deviations (SE) are presented.
[(a)] Cult/Lines: cultivars/lines

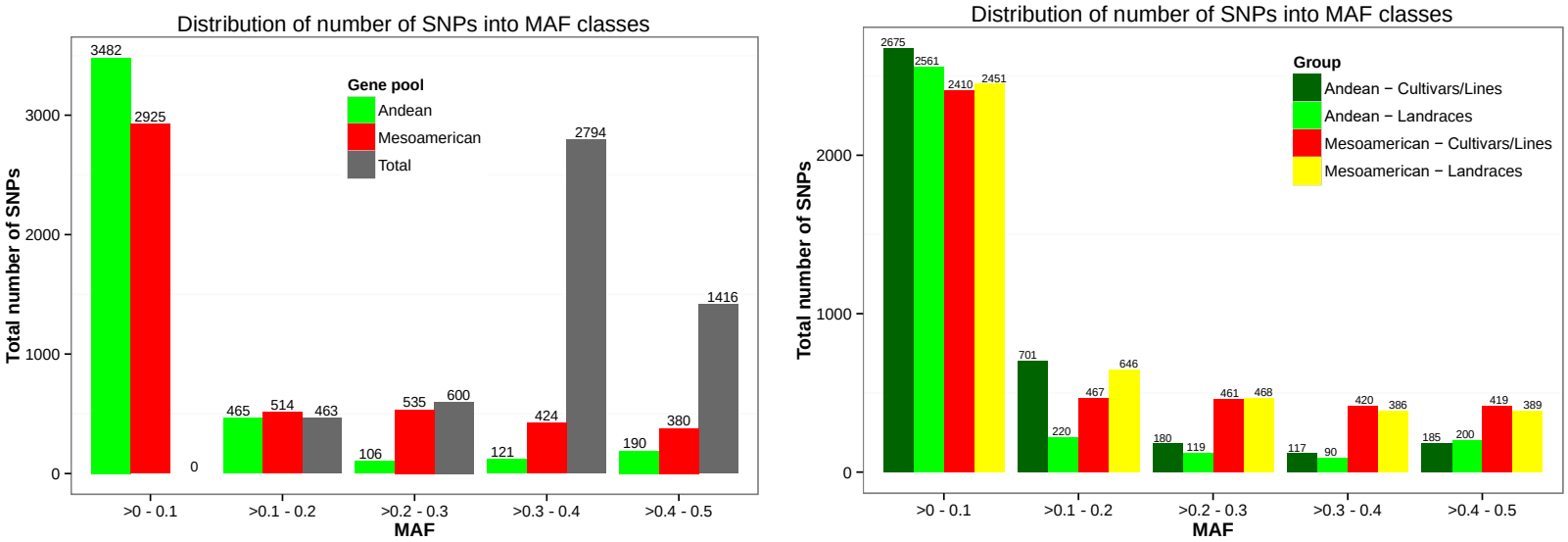



**Figure 7.** Distribution of SNPs into minor-allele frequency (MAF) classes. (A) Distribution of the number of SNPs into MAF classes for the whole population (grey), Andean (green), and Mesoamerican (red) genotypes. (B) Distribution of the number of SNPs into MAF classes for each group separately: Andean cultivars/lines (dark green) and landraces (light green) and Mesoamerican cultivars/lines (red) and landraces (yellow).

**SNPs outliers**

A total of 16 and 59 outlier SNP loci were identified based on BayeScan ($q < 0.05$) and Arlequin ($p < 0.05$), respectively, of which 16 loci were common to both analyses. From the 59 SNPs, 54 aligned over the 11 chromosomes (with the highest abundance on chromosomes 1 and 9), with an average of one SNP every 8.6 million bases. Across the genome, ~83% of SNPs were identified in genic flanking regions (in a 5 kb window) or within the genes (26.36% in the downstream region, 15.46% in the upstream region, 17.27% in introns, 20.91% in exons, and 2.73% in the 5' UTR). The Ts/Tv rate of the outliers was of 2.18. A total of 110 putative effects were predicted for 54 outliers, of which 11% were low-impact, 10% moderate and 79.09% modifier type. We identified 91 transcripts affected by 54 SNPs outliers, of which 82 presented homology to the nr (non-redundant) database and 71 were annotated (Supplementary Material 7). Based on GO, within the categories of "cellular component," "biological process," and "molecular function," most genes were assigned to "integral component of membrane, plasma membrane, and cytoplasmic part," "DNA binding, ATP binding, and ligase activity," and "DNA metabolic process, transmembrane transport, and signal transduction," respectively (Figure 8). In addition, 45 outliers SNP were identified in metabolic pathways (Supplementary Material 8).

Within the Mesoamerican gene pools (comparing between landraces and cultivars/lines), 15 outlier SNPs common to both analyses were identified distributed around chromosomes 2, 7, 8, and 9. 131 transcripts were affected by these SNPs, and 116 of these have been annotated (Supplementary Material 9). For the Andean group, a set of 18 outliers SNP, mainly in chromosome 10, were associated with 42 transcripts, of which 35 were annotated (Supplementary Material 10). Only one outlier loci (3381974_16_T_C) was common to both gene pools. The most abundant functional terms within one of the three GO categories is described in figure 9.

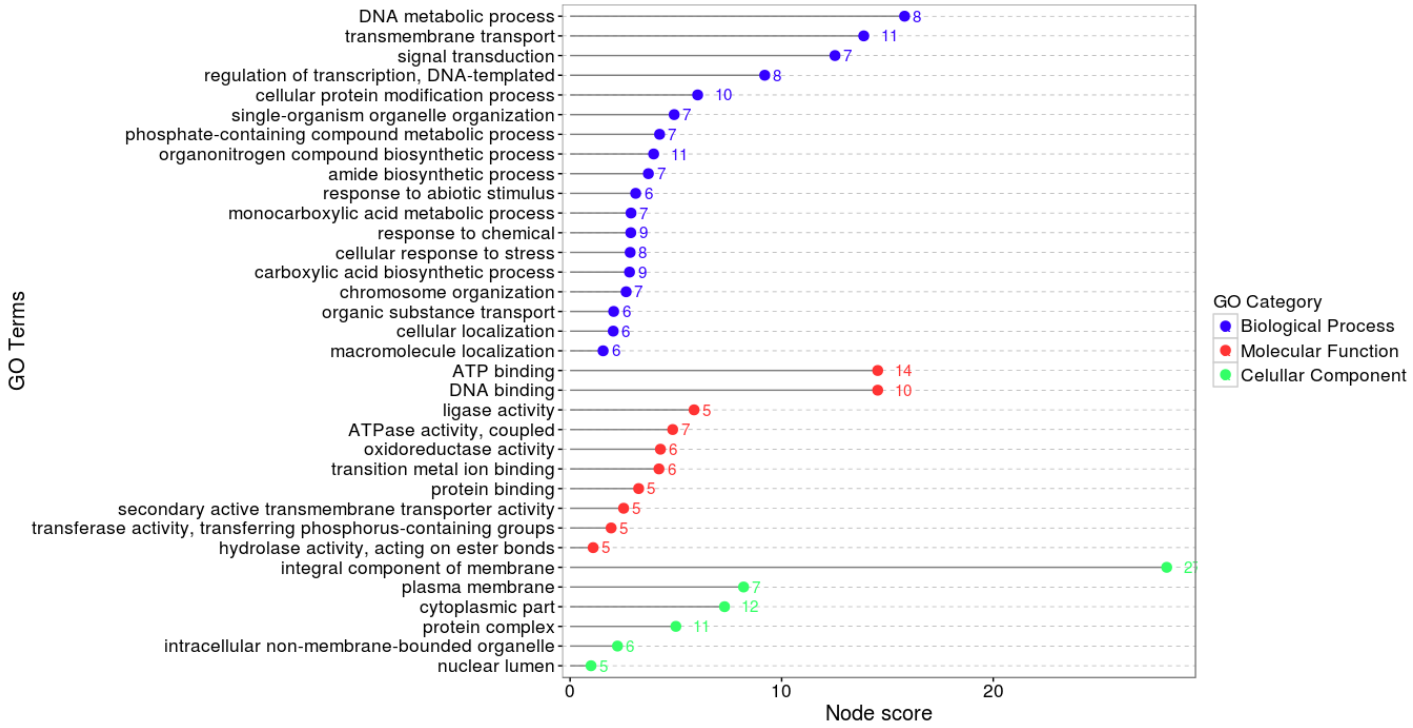
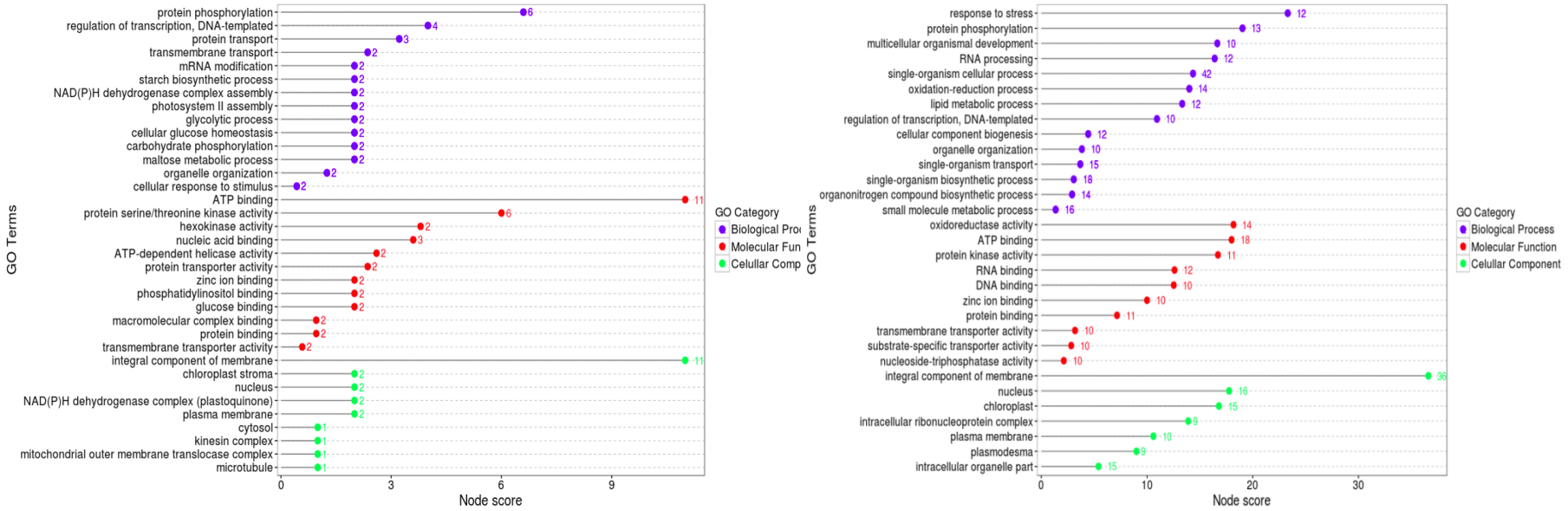



**Figure 8.** Functional annotation showing the most relevant GO terms for the outliers SNPs. The terms were filtered according to the node score. The numbers represent the amount of transcripts related to each term.




**A** **B**

**Figure 9.** Functional annotation showing the most relevant GO terms for the outliers SNPs. The terms were filtered according to the node score. The numbers represent the amount of transcripts related to each term. (A) Andean; (B) Mesoamerican.

**LD decay**

The LD decay in the Andean gene pool (Figure 10A and 10D) was slower than in the Mesoamerican gene pool (Figure 10B and 10E). For both groups, LD with correction for relatedness and structure showed a faster rate of decay and dropped to half ($r^2$ ~ 0.23) at a distance of ~2055 kb and ~395 kb for $r^2$ and $r^2_{SV}$ for Andean, respectively, while for the Mesoamerican group, half decay was observed at distances of ~312 kb and 130 kb. For accessions overall without correction ($r^2$), no decay was observed within the 3000 kb window (Figure 10C), while with correction ($r^2_{SV}$) LD decreased to half at 88 kb (Figure 10F). Within the landraces, the $r^2$ was estimated to be 1722 kb and 389 kb, for the Andean and Mesoamerican groups, respectively, and for the stratum cultivars/lines, it was 4040 kb and 428 kb.

Through haplotype analysis, a set of 437 blocks representative of the 11 chromosomes, ranging from 31 (chromosome 1) to 62 (chromosome 8) were identified. A total of 4354 SNPs (82.57%) were distributed in these blocks, with an average of ~10 SNPs per block. Chromosomes 9 (90.12%) and 4 (71.77%) had the highest and lowest percentage of SNPs within blocks, respectively. The average block size was 842.2 kb, and the largest block was in chromosome 3, with 8634 kb and 21 SNPs. The maximum and minimum frequency of haplotypes was 0.850 and 0.010, respectively, with the most common haplotype located on chromosome 7. On average, 71.66% of the genome was covered by the blocks (Table 5). A larger number of blocks were identified in the Mesoamerican group (248 blocks), compared to the Andean group (98 blocks), comprising 798 (3.18 SNPs/block) and 592 (4.6 SNPs/block) SNPs, respectively. In both gene pools, chromosome 2 presented the highest number of blocks (25 for Andean and 41 for Mesoamerican groups, Table 6).

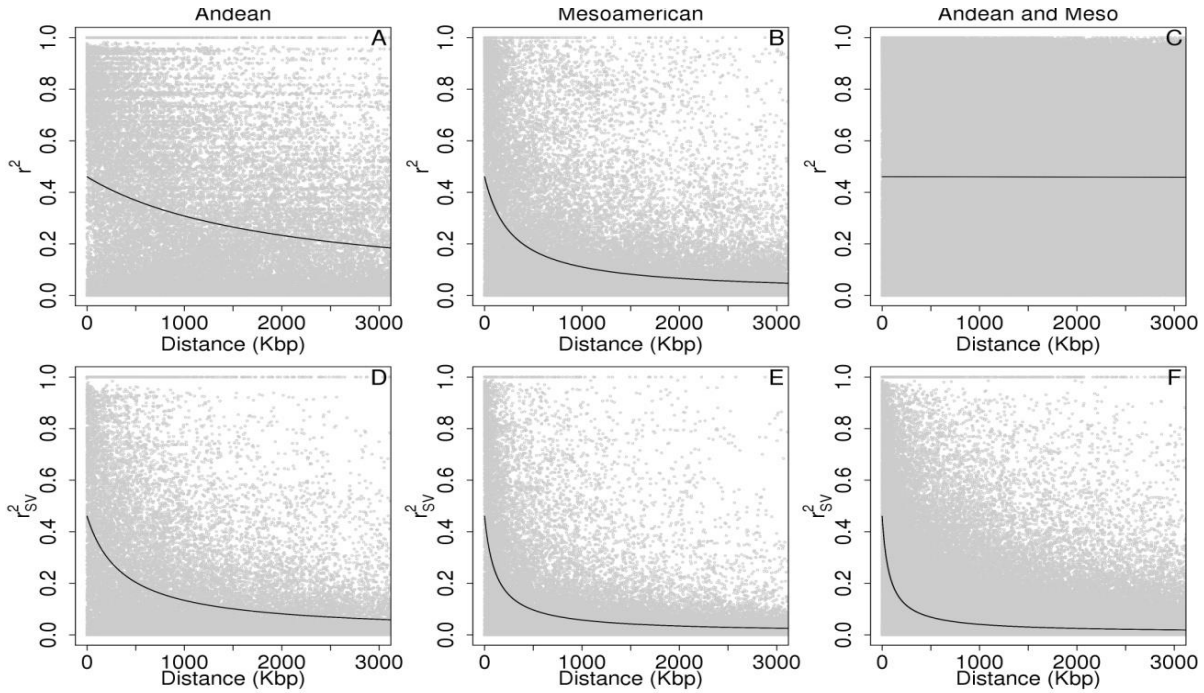



**Figure 10.** Linkage Disequilibrium decay without correction ($r^2$ , panels A, B and C) and corrected for relatedness and population structure ($r^2_{sv,}$ panels D, E and F) for the Mesoamerican (A and D), Andean (B and E) and the grouped accessions (C and F).

**Table 5.** Haplotype blocks identification based on genotyping of common bean germplasm using the SNPs-DArTseq.

| Chromosomes | Total of blocks | Total SNPs/haplotype blocks | Average SNP/block | SNPs/haplotype blocks (%) | Blocks size (kb) | Physical length/chromossomes (kb)[a] | Block genome coverage (%) |
|---|---|---|---|---|---|---|---|
| 1 | 31 | 432 | 13.94 | 86.75 | 42497.69 | 52183.50 | 81.44 |
| 2 | 56 | 600 | 10.71 | 83.92 | 38734.17 | 49033.70 | 78.99 |
| 3 | 37 | 479 | 12.95 | 84.63 | 42067.03 | 52218.60 | 80.56 |
| 4 | 35 | 239 | 6.83 | 71.77 | 30826.24 | 45793.20 | 67.32 |
| 5 | 35 | 311 | 8.89 | 79.74 | 28497.35 | 40237.50 | 70.82 |
| 6 | 40 | 423 | 10.58 | 88.68 | 24709.16 | 31973.20 | 77.28 |
| 7 | 37 | 420 | 11.35 | 87.87 | 35632.69 | 51698.40 | 68.92 |
| 8 | 62 | 428 | 6.90 | 76.98 | 35786.86 | 59634.60 | 60.01 |
| 9 | 33 | 438 | 13.27 | 90.12 | 30338.75 | 37399.60 | 81.12 |
| 10 | 36 | 248 | 6.89 | 74.92 | 26847.85 | 43213.20 | 62.13 |
| 11 | 35 | 336 | 9.60 | 75.85 | 32104.09 | 50203.60 | 63.95 |
| Total | 437 | 4354 | 9.96 | 82.57 | 368041.88 | 513589.10 | 71.66[b] |

a Schmutz et al. (2014)
b Average genome block coverage

**Table 6.** Haplotype blocks identification within Andean (AND) and Mesoamerican (MESO) gene pools.

| Chromosome | Total of blocks | | Total SNPs/blocks | | Average SNP/block | | SNPs/haplotype blocks (%) | | Blocks size (kb) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AND | MESO | AND | MESO | AND | MESO | AND | MESO | AND | MESO |
| 1 | 2 | 22 | 8 | 80 | 4.00 | 3.64 | 8.60 | 36.70 | 67.10 | 1831.76 |
| 2 | 25 | 41 | 204 | 44 | 8.16 | 1.07 | 43.59 | 14.10 | 22616.37 | 10256.22 |
| 3 | 5 | 20 | 11 | 84 | 2.20 | 4.20 | 10.09 | 38.71 | 28.27 | 6336.05 |
| 4 | 19 | 18 | 116 | 94 | 6.11 | 5.22 | 51.56 | 42.34 | 18181.43 | 20469.35 |
| 5 | 9 | 20 | 66 | 76 | 7.33 | 3.80 | 45.83 | 41.08 | 3365.23 | 10704.77 |
| 6 | 10 | 23 | 74 | 69 | 7.40 | 3.00 | 48.68 | 32.70 | 132885.33 | 1542.29 |
| 7 | 5 | 18 | 10 | 50 | 2.00 | 2.78 | 12.99 | 26.18 | 31.90 | 506.80 |
| 8 | 5 | 36 | 19 | 126 | 3.80 | 3.50 | 13.29 | 36.10 | 354.41 | 2954.98 |
| 9 | 4 | 16 | 18 | 58 | 4.50 | 3.63 | 22.22 | 35.15 | 1362.90 | 1922.26 |
| 10 | 6 | 17 | 33 | 52 | 5.50 | 3.06 | 28.45 | 30.06 | 1712.86 | 3235.07 |
| 11 | 8 | 17 | 33 | 56 | 4.13 | 3.29 | 21.85 | 27.18 | 659.16 | 2795.36 |
| TOTAL | 98 | 248 | 592 | 789 | 6.04 | 3.18 | 33.66 | 32.22 | 181264.96 | 62554.91 |

**Genetic analysis based on a low-density SNP panel**

A total of 560 SNPs (Supplementary Material 12) were selected for the assessment of genetic diversity in common bean. These SNPs were polymorphic in both gene pools, with MAF > 0.05, $r^2 < 0.5$, an average $H_E = 0.401$ and were distributed over the 11 chromosomes. The *F*-values between the Andean and Mesoamerican groups were $F_{ST} = 0.411$ (± 0.001), $F_{IS} = 0.826$ (± 0.001), and $F_{IT} = 0.897$ (± 0.001). DAPC revealed a structure similar to those obtained for the whole set of SNPs (5,531) (Figure 11). Within the Andean gene pool, 88.57% and 72.50% of SNPs were polymorphic for the landraces and cultivars/lines, respectively, with $F_{ST}$ estimated at 0.010 ± 0.001. For the Mesoamerican accessions, ~97% of SNPs were polymorphic in both strata, with an estimated $F_{ST}$ of 0.034 ± 0.001.

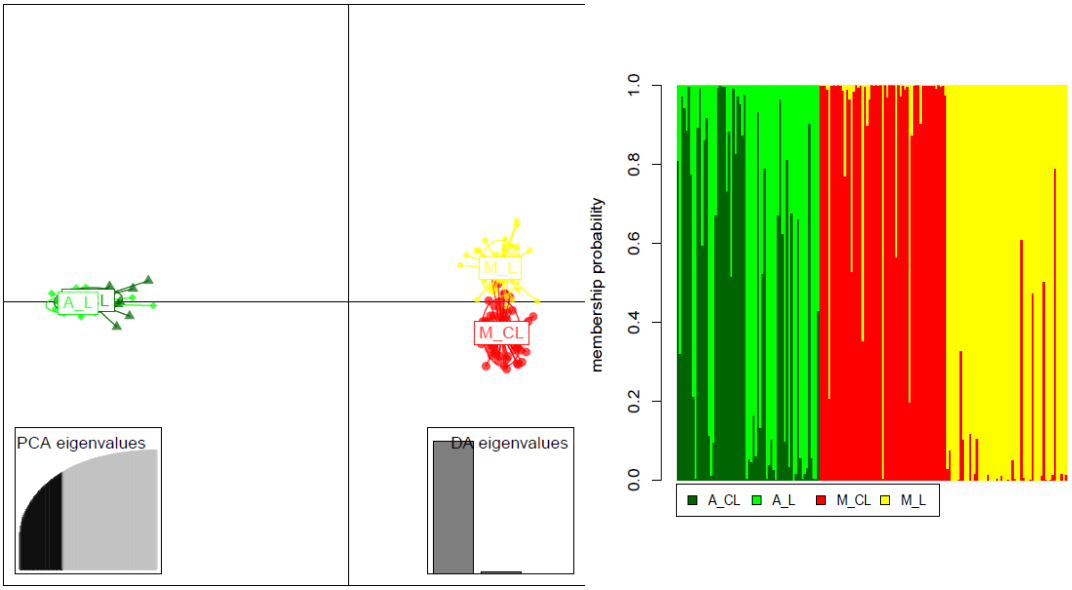
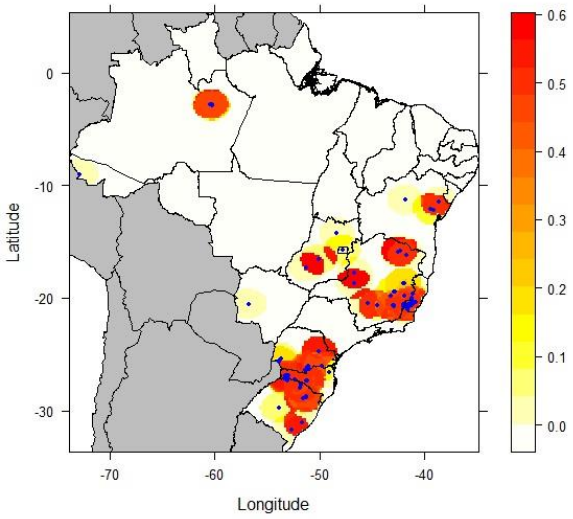


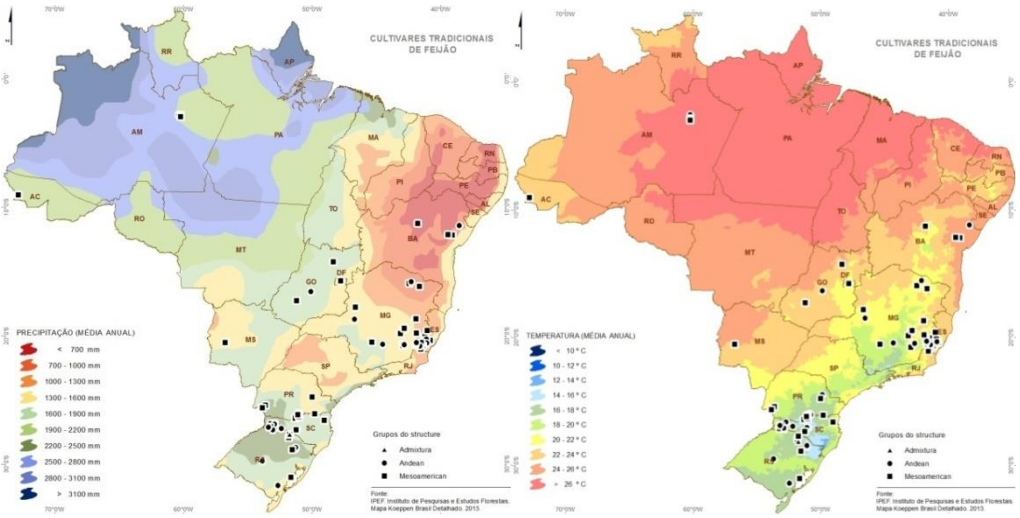

**Figure 11.** DAPC based on the 560 SNP panel showing the division between Mesoamerican cultivars/lines (red) and landraces (yellow). M_CL: Mesoamerican cultivars/lines; M_L: Mesoamerican landraces; A_CL (dark green): Andean cultivars/lines; A_L (light green): Andean landraces.

**Genetic diversity distribution based on temperature and rainfall maps**

The highest estimates of $H_E$ were observed in areas containing germplasm of Andean and Mesoamerican origin, as well as accessions characterized as admixtures by structure analysis (Figure 12). The Brazilian regions with the highest genetic diversity ($H_E$ ≥0.5) were in the states of Goiás, Bahia, Minas Gerais, Paraná, Santa Catarina and Rio Grande do Sul. The association of edaphoclimatic and genetic diversity maps provide a baseline to establish areas for in situ conservation and targets of diversity for collecting germplasm. In Bahia, which experiences annual average precipitation ranging from ≤ 700 mm to a maximum of 1,000 mm, seven accessions (five Mesoamerican and two Andean) were collected ($H_E \geq 0.1$), and showed an average genetic distance of ~0.4483. Within gene pools, the average genetic distance and $H_E$ were estimated at 0.1414 and 0.185 for the Mesoamerican and 0.054 and 0.099 for the Andean gene pools. Based on the Brazilian mean annual temperature map, three accessions were from regions with ≥ 26°C (two Mesoamerican - CF200005, CF200003 - and one Andean - CF200002) and one was from regions ranging from 14 - 16°C (CF200070). The $H_E$ among these four accessions was estimated at 0.411.




**A**




**B** **C**

**Figure 12.** Spatial distribution of the genetic diversity (A) on the maps of total annual rainfall (B) and average annual temperature (C). The circle represents accessions from Mesoamerican origin and square from Andean origin.

## DISCUSSION

### Genotyping with DArTseq

The SNP markers derived using DArTseq demonstrated that this technology is an efficient method of genotyping with broad genome coverage and can be useful for analyses of genetic diversity in a common bean germplasm pool composed of landraces and cultivars. The sequencing of two important varieties of common bean representative of the Andean and Mesoamerican groups (Schmutz et al. 2014, Vlasova et al. 2016) has allowed the identification and determination of genomic positions of SNPs with several scientific implications. Among the 6,286 SNPs identified, 94.82% were placed on the *P. vulgaris* genome, supporting the analysis of population structure, LD, and identification of genomic regions under possible selection that have an impact on crop improvement research. The average call rate was 92%, close to the value of 91.3% previously reported for watermelon (Ren et al. 2015), and the scoring reproducibility of 99.44% was consistent with the value described for wheat (98.5%; Li et al. 2015). A higher density of SNPs was obtained (SNP/86 kbp) compared to our previous report (SNP/500 kbp), which was based on RADseq technology (Valdisser et al. 2016), increasing the genome resolution for subsequent analyses. The combination of restriction enzymes used in DArTseq (PstI-MseI) resulted in the more frequent appearance of SNPs, as reported by Schröder et al. (2016). Within the Andean (83%) and Mesoamerican (91%) gene pools, a larger proportion of polymorphisms were identified, which was higher than previously reported using SNPs-RAD (Andean: 72.7% and Mesoamerican 83.3%; Valdisser et al. 2016). The considerable level of SNP polymorphism within gene pools, in addition to their wide genomic representativeness over the genome (99.78%), is favorable to reduce the ascertainment bias given a more uniform and realistic distribution of allelic frequency over the whole population.

### Genetic Diversity

DArTseq also allowed the detection of SNPs with high diversity (n=175, $H_E$ = 0.442), compared to the SNPs identified by RADseq (n = 95, $H_E$ = 0.384) and SNPs derived from the polymorphism between BAT93 and Jalo Epp558 (n = 88, $H_E$ = 0.390), described by Valdisser et al. (2016) and Müller et al. (2015), respectively. For the Mesoamerican germplasm, genetic diversity ($H_E$ =0.168, n = 111) was close to values obtained by Rodriguez

et al. (2016) for domesticated bean accessions ($H_E$ = 0.157, n = 100); however, they are lower compared to the studies of Cichy et al. (2015) ($H_E$ = 0.233, n = 21), who characterized accessions with different traits, such as geographic origin, breeding program, grain type, and growth habit. The Andean group presented less diversity when compared to the Mesoamerican ($H_E$ = 0.102, n = 64, $p \leq 0.05$), which was expected due to the prevalence of Mesoamerican introduction and domestication in Brazil (Burle et al. 2010, Cardoso et al. 2014), in addition to global historical events of domestication (Bitocchi et al. 2012). In the present study, all landraces originated from Brazil and only the breeding germplasm included introductions, whereas in the preceding studies (Cichy et al. 2015, Rodriguez et al. 2016) there was a global representativeness in both strata of germplasms supporting the highest estimates of diversity. From the 87 Brazilian landraces, lower estimates of $H_E$ were reported for the Andean (n = 33), compared to the Mesoamerican (n = 54) origin (1.9 x smaller) that predominates in Brazil (Burle et al. 2010). Small farmers (largest supplier of landraces) have a preference for common bean with small grain types. Moreover, the diversity of the Andean Brazilian landraces (n = 33, $H_E$ = 0.099) reduces only in the order of 1.3 x compared to the diversity estimated by Cichy et al. (2015) which analyzed accessions representative of ~30 countries using the BARCBean6K_3 bead chip (n = 201 landraces, $H_E$ = 0.128). Therefore, there is a strong indication that Brazil holds a center of secondary domestication with high diversity that deserves a deeper characterization of the accessions integrated into the Brazilian core collection.

**Genetic diversity and differentiation over the genome**

The nucleotide diversity presented a reduction in the order of 60.8% for the Andean ($\pi$ = 0.001781), compared to the Mesoamerican ($\pi$ = 0.004541) germplasm, as well as a reduction in the proportion of polymorphic SNPs (8.6 %), in accordance with previously reported values (Beebe et al. 2011, Cichy et al. 2015). Evolutionary events associated with the genetic bottleneck phenomenon in Andean populations, dated 165,000 years ago, caused widespread loss of genetic diversity, while the Mesoamerican germplasm was less affected (Bitocchi et al. 2013, Rossi et al. 2013, Schmutz et al. 2014). Inter-gene pool hybridization had a positive impact on Andean diversity (Gioia et al. 2013, Schmutz et al. 2014), whereas the Mesoamerican group was exposed to more frequent events of recombination, generating more diversity (Miklas et al. 2006). The Andean cultivars/lines (n = 31), representing 22 countries, produced an $\pi_{C/L}$ estimate of (0.002171) almost half the value for the Mesoamerican

group (n = 57, $\pi_{C/L}$= 0.004389), represented by 10 accessions of foreign origin and 49 Brazilian ones but were extensively submitted to breeding and directional selection for agronomical traits (Cardoso et al. 2014). As previously described, analysis of structure revealed the Mesoamerican accessions were grouped by grain type (K = 4; Müller et al. 2015, Valdisser et al. 2016), with a high proportion of admixture (55.85%) supported for long term genetic improvement and relationships in the breeding germplasm. Regarding the spatial distribution of the genetic diversity, no significant relationship between genetic and physical distances was identified (data not shown). However, by overlapping the thematic and diversity maps, target geographic areas were identified and important sites to collect landraces, considering both genetic diversity and adaptability under hydric restriction and high temperatures, were highlighted. This has important implications for conservation purposes and is potentially useful for genetic breeding programs.

**Loci under possible selection and annotation of the effects**

The process of domestication and artificial selection imposed by agriculture resulted in changes to allelic frequencies and allowed the identification of genomic regions under adaptive evolution using high-density SNP genotyping (Stapley et al. 2010, Li et al. 2013, Schmutz et al. 2014). In this study, high-resolution genetic analysis and a diverse set of domesticated accessions adapted to specific environments and subjected to natural and artificial selection allowed the identification of SNPs potentially related to these adaptive processes. SNP effects categorized as modifiers, where the prediction is difficult or does not have an evident effect on biological impact, was more abundant (77.8%) as was expected (Cingolani et al. 2012). The effect of SNPs on protein efficiency and loss-of-function under strong selective pressure was reported in a smaller proportion (7.97%), as expected, since these have a direct impact on gene function with adaptive interference during the course of selection. These high impact SNP effects provide important clues about the selective forces acting in germplasm adaptation. Alonso-Blanco et al. (2009) reported a loss of gene function conferring an adaptive advantage under domestication for the processes of germination, dormancy, and flowering. In addition, during the process of domestication, loss of function can be considered an important factor for rapid evolution (Olson 1999).

Genomic regions under possible selection were not homogeneous in the present study (predominant on chromossome 1, $F_{ST}$ = 0.86 and chromossome 9, $F_{ST}$ = 0.87), suggesting that distinct and broad genetic mechanisms were involved in the process of

common bean domestication. Schmutz et al. (2014) reported a greater proportion of loci under selection on chromosomes 1, 2, and 10 in the Andean group, and on chromosomes 2, 7, and 9 in the Mesoamerican group, as well as slow sharing in the regions under domestication (7.234 Mb), suggesting different genetic routes to domestication. Among the genes under possible positive selection, we identified enrichment in terms related to cell membrane transporters, receptors, sensors, gene recombination/mutation, and the complex network of intra- and extracellular signaling that could be attributed to adaptive changes, providing the ability to respond earlier to abiotic or biotic stimuli. Tolerance to multiple stresses is expected since the plants suffer from several forms of stress during their life cycle, where a range of molecular mechanisms act together through complex pathways with important mechanisms of crosstalk among them (Atkinson and Urwin 2012). Among the landrace and cultivar/line strata, a high number of outliers was reported in the accessions of Mesoamerican origin, which is consistent with the predominance of this germplasm in Brazil (Burle et al. 2010, Cardoso et al. 2014) and, consequently, the higher selective pressure imposed on this germplasm. These genes are potential targets for plant breeders because of their roles in plant adaptation under variable environmental conditions. The understanding the effect of these genes on the phenotypes will have a positive impact on crop improvement (Huq et al. 2016)

Outliers SNP associated with the same GO terms were reported in both gene pools. Among these, we highlighted integral components of membranes that could respond to plant demand to be more efficient in the process of water and nutrient transport, as well as the location of photoassimilates. Furthermore, several common transcripts related to the development of morpho-anatomical structures were reported, corroborating previous studies of QTLs involved in the domestication and diversification processes (Meyer and Purugganan 2013). Selective pressure on these genes is expected because in the process of domestication, several traits were privileged, for example, the trend for increases in wheat grain mass, which is strongly associated with endosperm development (Golan et al. 2015) and growth habit, as a trait under strong selection in common bean domestication (Repinski et al. 2012). Selective pressure on genes related to the redox status, plant development, and response to biotic and abiotic stresses was preferably identified in the Mesoamerican group, while processes of protein phosphorylation and ATP-binding predominated in the Andean germplasm. These genes play a fundamental role in the stimuli and signal processing of multiple stress responses, which are fundamental to plant adaptation in the evolution and domestication (Chen et al. 2015). In addition, transcription factors and other genes related to the regulation of gene expression were also under possible selection. Genes related to the same mechanisms

and associated with QTLs controlling domestication-related traits were reported by Doebley et al. (2006). Similar of those observed in maize (Rhode et al. 2011) and soybean (Li et al. 2013), transcription factors are abundant among genes under selection acting to regulate several process, such as grown habit, flowering, grain size, dormancy, and others (Swinnen et al. 2016). Lastly, genes related to secondary metabolites under selection that are known to respond to plant interactions with environmental changes, such as drought, radiation intensity, and pest attacks, were identified in this study (Kliebenstein 2009).

**Linkage disequilibrium**

A high proportion of alleles at low frequencies were observed within the gene pools, whereas for the whole set of accessions, most SNPs were present at high frequencies (≥ 0.3), reflecting the presence of fixed loci for alternative alleles between the gene pools. Without correcting for relatedness and structure, the LD presented elevate estimates of $r^2$, and after the correction, an increase in decay was observed, showing that the evolutionary and breeding history strongly affect the association among markers (Rossi et al. 2009, Mamidi et al. 2013). The LD decay observed in common bean extended over several bp (up to 88 kb), compared to allogamous species, such as the Eucalyptus (Silva-Junior and Grattapaglia 2015) and loblolly pine (Brown et al. 2004). This was expected due to the selfing nature of beans, which leads to increased amounts of LD. LD has being studied in several autogamous species, such as rice ($r^2$ values of ~123 kb and ~167 kb for *Indica* and *Japonica*, respectively; Huang et al. 2012) and soybean (values of ~27 kb, ~83 kb and 133 kb for wild, landrace and cultivars, respectively; Zhou et al. 2015). Compared to the above mentioned varieties, common bean presented higher LD extension for the Andean ($r^2_{SV}$ = ~395 kb) germplasm, whereas for the Mesoamerican groups of germplasm, LD was similar to that of *Indica* rice and cultivar of soybean ($r^2_{SV}$ = ~130 kb). These close values of LD among different cultivated species make sense because they are composed of groups of germplasm under more intense breeding. Among groups of germplasm within each species, the LD variation is determined by several factors, such as the demographic dynamics, recombination rates, and evolutionary mechanisms (Slatkin et al. 2008).

In this study, the cultivars/lines (Andean $LD_{C/L}$ = 4040 kbp, Meso $LD_{C/L}$ = 428 kbp) presented slower LD decay compared to the landraces (Andean $LD_L$ = 1722 kbp, Meso $LD_L$ = 389 kbp), consistent with previous studies (Valdisser et al. 2016). The more genetically diverse the germplasm, the more rapid the expected decay, which provides more opportunity

for selection, which is extremely important for common bean breeding (Li et al. 2011). The reduced diversity in the cultivated germplasm in Brazil probably is associated with the low maintenance and breeding base population size (Cardoso et al. 2014), in addition to the amount of recombination accumulated over the course of selection after breeding programs appeared in the 1930 (Cardoso et al. 2015), whereas the landraces have been disseminated by Brazil and domesticated since the sixteenth century (Vieira 1988). Furthermore, the variation in size of the haplotype blocks across the common bean genome (Table 5) revealed a considerable degree of LD variation, becoming more complex with reliability studies of association and genome selection for beans, as has also been reported for soybean (Hyten et al. 2007). In this way, the adoption of a general LD value is not recommended, as demonstrated for soybean (Zhou et al. 2015) and wheat (Würschum et al. 2013). For common bean, it is evident that the level of genetic diversity and LD decay are associated with the germplasm origin and process of domestication, which must be considered to choose the most appropriate strategy for analysis. The increase in SNP resolution, using an adequate effective population size, and considering the process of germplasm domestication, is shown here to be of great utility for association and genome selection studies in order to identify genes related to complex traits.

**SNP Panel**

DArTseq analysis over a diverse group of common bean germplasm allowed the identification of a panel composed of 560 SNPs, selected from the whole set of 6,286, with nearly 90% genome coverage. For breeding purposes, this panel of SNP, which allows identification of genetic intervals at low to moderate resolution, would be readily incorporated to routine genetic analysis of breeding programs. The benefits of marker-assisted breeding using this panel over a large set of SNPs are due to the increase in the efficiency of genome sampling at a lower cost. This panel certainly will be of great utility for germplasm characterization, linkage mapping, and assisted backcrossing, reaching the research demands with high technological impact for this crop.

**REFERENCES**

1. Alonso-Blanco C, Aarts MGM, Bentsink L, Keurentjes JJB, Reymond M, Vreugdenhil D, Koornneef M: What Has Natural Variation Taught Us about Plant Development, Physiology, and Adaptation.Plant Cell 2009, 21(7):1877-1896.
2. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997, 25(17):3389-3402.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al: Gene Ontology: tool for the unification of biology. Nature Genetics 2000, 25(1):25-29.
4. Atkinson NJ, Urwin PE: The interaction of plant biotic and abiotic stresses: from genes to the field. J Exp Bot 2012, 63:3523-3543.
5. Barrett JC, Fry B, Maller J, Daly MJ: Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005, 21(2):263-265.
6. Beebe RJ, Jarvi, A, Rao MI, et al. Genetic Improvement of Common Beans and the Challenges of Climate Change. In: Yadav SS, Redden JR, Hatfield LJ, Lotze-Campen H, Hall EA, editors. Crop Adaptation to Climate Change. Colombia: Blackwell Publishing Ltd. 2011. p. 356-369.
7. Bitocchi E, Nanni L, Bellucci E, et al: Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. Pnas 2012, 109:788–796.
8. Bitocchi E, Bellucci E, Giardini A, et al: Molecular analysis of the parallel domestication of the common bean (*Phaseolus vulgaris*) in Mesoamerica and the Andes. New Phytol 2013, 197:300-313.
9. Blair MW, Muñoz-Torres M, Giraldo MC, Pedraza F: Development and diversity assessment of Andean-derived, gene-based microsatellites for common bean (*Phaseolus vulgaris* L.). BMC Plant Bio 2009, 9: 100.
10. Blair MW, Davila AM, Reyes X, Avila T: Genetic Diversity of Bolivian Accessions of Phaseolus Species Evaluated with Fluorescent Microsatellite Markers. Crop Science 2012, 52(6):2619-2627.
11. Blair MW, Cortés AJ, Penmetsa RV, et al: A high-throughput SNP marker system for parental polymorphism screening, and diversity analysis in common bean (*Phaseolus vulgaris* L.). Theor Appl Genet 2013, 126:535-548.
12. Blair MW, Lorigados SM: Diversity of Common Bean Landraces, Breeding Lines, and Varieties from Cuba. Crop Science 2016, 56(1):322-330.

13. Briñez B, Blair MW, Kilian A, Carbonell SAM, Chiorato AF, Rubiano LB: A whole genome DArT assay to assess germplasm collection diversity in common beans. Molecular Breeding 2012, 30(1):181-193.

14. Broughton WJ, Hernandez G, Blair M, Beebe S, Gepts P, Vanderleyden J: Beans (Phaseolus spp.) - model food legumes. Plant and Soil 2003, 252(1):55-128.

15. Brown GR, Gill GP, Kuntz RJ, Langley CH, Neale DB: Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America 2004, 101(42):15255-15260.

16. Burle ML, Fonseca JR, Kami JA, Gepts P: Microsatellite diversity and genetic structure among common bean (*Phaseolus vulgaris* L.) landraces in Brazil, a secondary center of diversity. Theoretical and Applied Genetics 2010, 121(5):801-813.

17. Cardoso PCB, Brondani C, Menezes IPP, Valdisser PAMR, Borba TCO, Del Peloso MJ, Vianello RP: Discrimination of common bean cultivars using multiplexed microsatellite markers. Genetics and Molecular Research 2014, 13(1):1964-1978.

18. Chen J, Nolte V, Schlotterer C: Temperature Stress Mediates Decanalization and Dominance of Gene Expression in Drosophila melanogaster. PLOS Genetics 2015, 11(6): e1005315.

19. CIAT. Beans, Cassava, and Tropical Forages.2014. http://www.croptrust.org/wp-content/uploads/2014/12/CIAT.pdf. Accessed 14 December 2016.

20. Cichy KA, Porch TG, Beaver JS, Cregan P, Fourie D, Glahn RP, Grusak MA, Kamfwa K, Katuuramu DN, McClean Pet al: A *Phaseolus vulgaris* Diversity Panel for Andean Bean Improvement. Crop Science 2015, 55(5):2149-2160.

21. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w(1118); iso-2; iso-3. Fly 2012, 6(2):80-92.

22. Conesa A, Götz S, García-Gómez JM, et al: Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005, 21:3674–3676.

23. Cruz VMV, Kilian A, Dierig DA: Development of DArT Marker Platforms and Genetic Diversity Assessment of the US Collection of the New Oilseed Crop Lesquerella and Related Species. Plos One 2013, 8(5).

24. Desrousseaux D, Sandron F, Siberchicot A, Cierco-Ayrolles C and Mangin B (2013) LDcorSV: Linkage disequilibrium corrected by the structure and the relatedness. R package version 3.2.1.

25. Doebley JF, Gaut BS, Smith BD: The molecular genetics of crop domestication. Cell 2006, 127(7):1309-1321.

26. Dwivedi SL, Ceccarelli S, Blair MW , Upadhyaya HD, Are AK, Ortiz R: Landrace germplasm for improving yield and abiotic stress adaptation. Trends in Plant Science 2016, 21:31-42 .

27. Earl DA, Vonholdt BM: Structure Harvester: a website and program for visualizing structure output and implementing the Evanno method. Conservation Genetics Resources 2012, 4(2):359-361.

28. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Plos One 2011, 6(5).

29. Evanno G, Regnaut S, Goudet J: Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 2005, 14(8):2611-2620.

30. Excoffier L, Hofer T, Foll M: Detecting loci under selection in a hierarchically structured population. Heredity 2009, 103(4):285-298.

31. Excoffier L, Lischer HEL: Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 2010, 10(3):564-567.

32. FAO (2016). Faostat. http://www.fao.org/faostat/en/#data/QC. Accessed 14 October 2016.

33. Foll M, Gaggiotti O: A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. Genetics 2008, 180(2):977-993.

34. Füleky G. Cultivated plants, primarily as food resources. Eolss Publishers Company Limited, vol I. 2009.

35. Gill-Langarica HR, Muruaga-Martínez JS, Vargas-Vásquez MLP, et al: Genetic diversity analysis of common beans based on molecular markers. Genetics and Molecular Biology 2011, 34:595-605.

36. Gioia T, Logozzo G, Attene G, Bellucci E, Benedettelli S, Negri V, Papa R, Zeuli PS: Evidence for Introduction Bottleneck and Extensive Inter-Gene Pool (Mesoamerica x Andes) Hybridization in the European Common Bean (*Phaseolus vulgaris* L.) Germplasm. Plos One 2013, 8(10).

37. Golan G, Oksenberg A, Peleg Z: Genetic evidence for differential selection of grain and embryo weight during wheat evolution under domestication. Journal of Experimental Botany 2015, 66(19):5703-5711.

38. Goodstein DM, Shu SQ, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N et al: Phytozome: a comparative platform for green plant genomics. Nucleic Acids Research 2012, 40(D1):D1178-D1186.

39. Hahn V, Wurschum T: Molecular genetic characterization of Central European soybean breeding germplasm. Plant Breeding 2014, 133(6):748-755.

40. Hill WG, Weir BS: Variances and covariances of squared linkage disequilibria in finite populations. Theoretical Population Biology 1988, 33(1):54-78.

41. Huang XH, Kurata N, Wei XH, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu KY, Lu HY, Li WJ et al: A map of rice genome variation reveals the origin of cultivated rice. Nature 2012, 490(7421):497-+.

42. Hudson RR, Slatkin M, Maddison WP: Estimation of levels of gene flow from DNA-sequence data. Genetics 1992, 132(2):583-589.

43. Huq MA, Shahina A, Nou IS, et al: Identification of functional SNPs in genes and their effects on plant phenotypes. J Plant Biotechnol 2016, 43:1-11.

44. Hyten DL, Choi I-Y, Song Q: Highly Variable Patterns of Linkage Disequilibrium in Multiple Soybean Populations. Genetics 2007, 175: 1937-1944.

45. IBGE. Mapas. www.mapas.ibge.gov.br/. Accessed 18 June 2016.

46. Jaccoud D, Peng KM, Feinstein D, Kilian A: Diversity Arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Research 2001, 29(4).

47. Jakobsson M, Rosenberg NA: CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 2007, 23(14):1801-1806.

48. Jombart T, Ahmed I: adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 2011, 27(21):3070-3071.

49. Jombart T, Devillard S, Balloux F: Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 2010, 11.

50. Kilian A, Wenzl P, Huttner E, et al: Diversity Arrays Technology: a generic genome profiling technology on open platforms. Data Production and Analysis in Population Genomics 2012, 888: 67–89.

51. Kliebenstein D: Quantitative Genomics: Analyzing Intraspecific Variation Using Global Gene Expression Polymorphisms or eQTLs. Annual Review of Plant Biology 2009, 60:93-114.

52. Li X, Yan W, Agrama H, et al: Mapping QTLs for improving grain yield using the USDA rice mini-core collection. Planta 2011, 234: 347.

53. Li YH, Zhao SC, Ma JX, et al: Molecular footprints of domestication and improvement in soybean revealed by whole genome re-sequencing. BMC Genomics 2013, 14: 579.

54. Li H, Vikram P, Sing RP, et al: A high density GBS map of bread wheat and its application for dissecting complex disease resistance traits. BMC Genomics 2015, 16:216.

55. Mamidi S, Rossi M, Annam D, Moghaddam S, Lee R, Papa R, McClean P: Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. Functional Plant Biology 2011, 38(12):953-967.

56. Mamidi S, Rossi M, Moghaddam SM, Annam D, Lee R, Papa R, McClean PE: Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L. Heredity 2013, 110(3):267-276.

57. Manel S, Berthoud F, Bellemain E, Gaudeul M, Luikart G, Swenson JE, Waits LP, Taberlet P, Intrabiodiv C: A new individual-based spatial approach for identifying genetic discontinuities in natural populations. Molecular Ecology 2007, 16(10):2031-2043.

58. McCouch S, Baute GJ, Bradeen J, et al: Agriculture: Feeding the future. Nature 2013, 499:23-24.

59. Meyer RS, Purugganan MD: Evolution of crop species: genetics of domestication and diversification. Nature Reviews Genetics 2013, 14(12):840-852.

60. Meziadi C, Richard MMS, Derquennes A, Thareau V, Blanchet S, Gratias A, Pflieger S, Geffroy V: Development of molecular markers linked to disease resistance genes in common bean based on whole genome sequence. Plant Science 2016, 242:351-357.

61. Miklas PN, Kelly JD, Beebe SE, Blair MW: Common bean breeding for resistance against biotic and abiotic stresses: From classical to MAS breeding. Euphytica 2006, 147:105–131.

62. Müller BSD, Sakamoto T, de Menezes IPP, Prado GS, Martins WS, Brondani C, de Barros EG, Vianello RP: Analysis of BAC-end sequences in common bean (*Phaseolus vulgaris* L.) towards the development and characterization of long motifs SSRs. Plant Molecular Biology 2014, 86(4-5):455-470.

63. Müller BSF, Pappas GJ, Valdisser P, Coelho GRC, de Menezes IPP, Abreu AG, Borba TCO, Sakamoto T, Brondani C, Barros EG et al: An Operational SNP Panel Integrated to SSR Marker for the Assessment of Genetic Diversity and Population Structure of the Common Bean. Plant Molecular Biology Reporter 2015, 33(6):1697-1711.

64. Nei M, Li WH: Mathematical-model for studying genetic-variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America 1979, 76(10):5269-5273.

65. Nei M. Molecular Evolutionary Genetics. New York: Columbia University Press. 1987.

66. Oblessuc PR, Perseguini J, Baroni RM, Chiorato AF, Carbonell SAM, Mondego JMC, Vidal RO, Camargo LEA, Benchimol-Reis LL: Increasing the density of markers around a major QTL controlling resistance to angular leaf spot in common bean. Theoretical and Applied Genetics 2013, 126(10):2451-2465.

67. Olson MV: When less is more: Gene loss as an engine of evolutionary change. American Journal of Human Genetics 1999, 64(1):18-23.

68. Papa R, Gepts P: Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. Theoretical and Applied Genetics 2013, 106:239–250.

69. Papa R, Bellucci E, Rossi M, Leonardi S, Rau D, Gepts P, Nanni L, Attene G: Tagging the signatures of domestication in common bean (*Phaseolus vulgaris*) by means of pooled DNA samples. Annals of Botany 2007, 100(5):1039-1051.

70. Peakall R, Smouse PE: GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 2012, 28(19):2537-2539.

71. Perrier X, Jacquemoud-Collet JP (2006) Darwin software, http://darwin.cirad.fr/darwin. Accessed 22 September 2015.

72. Perseguini JMKC, Chioratto AF, Zucchi MI, et al. Genetic diversity in cultivated carioca common beans based on molecular marker analysis. Genet Mol Biol 2011, 34:88-102.

73. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ: PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. Molecular Biology and Evolution 2014, 31(7):1929-1936.

74. Porch TG, Beaver JS, Brick MA: Registration of Tepary Germplasm with Multiple-Stress Tolerance, TARS-Tep 22 and TARS-Tep 32. Journal of Plant Registrations 2013, 7(3):358-364.

75. Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 2000, 155(2):945-959.

76. Raman H, Raman R, Kilian A, Detering F, Carling J, Coombes N, Diffey S, Kadkol G, Edwards D, McCully M et al: Genome-Wide Delineation of Natural Variation for Pod Shatter Resistance in Brassica napus. Plos One 2014, 9(7):13.

77. R Development Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

78. Ren RS, Ray R, Li PF, Xu JH, Zhang M, Liu G, Yao XF, Kilian A, Yang XP: Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. Molecular Genetics and Genomics 2015, 290(4):1457-1470.

79. Repinski SL, Kwak M, Gepts P: The common bean growth habit gene PvTFL1y is a functional homolog of Arabidopsis TFL1. Theoretical and Applied Genetics 2012, 124(8):1539-1547.

80. Rhode H, Qin J, Cui Y, et al. Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. N Engl J Med 2011, 365:718-724.

81. Rodriguez M, Rau D, Bitocchi E, Bellucci E, Biagetti E, Carboni A, Gepts P, Nanni L, Papa R, Attene G: Landscape genetics, adaptive diversity and population structure in *Phaseolus vulgaris*. New Phytologist 2016, 209(4):1781-1794.

82. Rossi M, Bitocchi E, Bellucci E, Nanni L, Rau D, Attene G, Papa R: Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. Evolutionary Applications 2009, 2(4):504-522.

83. Sánchez-Sevilla JF, Horvath A, Botella MA, Gaston A, Folta K, Kilian A, Denoyes B, Amaya I: Diversity Arrays Technology (DArT) Marker Platforms for Diversity Analysis and Linkage Mapping in a Complex Crop, the Octoploid Cultivated Strawberry (Fragaria x ananassa). Plos One 2015, 10(12).

84. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu SQ, Song QJ, Chavarro C et al: A reference genome for common bean and genome-wide analysis of dual domestications. Nature Genetics 2014, 46(7):707-713.

85. Schröder S, Mamidi S, Lee R, McKain MR, McClean PE, Osorno JM: Optimization of genotyping by sequencing (GBS) data in common bean (*Phaseolus vulgaris* L.). Molecular Breeding 2016, 36(1).

86. Silva AR. Tools for Biometry and Applied Statistics in Agricultural Science. Package 'biotools'. Release 3.0., 2016.

87. Silva-Junior OB, Grattapaglia D: Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of *Eucalyptus grandis*. New Phytologist 2015, 208(3):830-845.

88. Slatkin M. Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. Nature 2008, 9:477-485.

89. Song QJ, Jia GF, Hyten DL, Jenkins J, Hwang EY, Schroeder SG, Osorno JM, Schmutz J, Jackson SA, McClean PE et al: SNP Assay Development for Linkage Map Construction, Anchoring Whole-Genome Sequence, and Other Genetic and Genomic Applications in Common Bean. G3-Genes Genomes Genetics 2015, 5(11):2285-2290.

90. Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AP, Slate J: Adaptation genomics: the next generation. Trends in Ecology & Evolution 2010, 25(12):705-712.

91. Swinnen G, Goossens A, Pauwels L: Lessons from Domestication: Targeting Cis-Regulatory Elements for Crop Improvement. Trends in Plant Science 2016, 21(6):506-515.

92. Tajima F: Statistical-method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989, 123(3):585-595.

93. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution 2011, 28(10):2731-2739.

94. Tiwari BK, Gowen A, McKenna B (2011) Pulse foods: Processing, quality and nutraceutical applications. In: Food, science and technology, international serie. San Diego, CA: EU: Academic Press, Elsevier

95. Valdisser P, Pappas GJ, de Menezes IPP, Müller BSF, Pereira WJ, Narciso MG, Brondani C, Souza T, Borba TCO, Vianello RP: SNP discovery in common bean by restriction-associated DNA (RAD) sequencing for genetic diversity and population structure analysis. Molecular Genetics and Genomics 2016, 291(3):1277-1291.

96. Vieira C. Phaseolus genetic resources and breeding in Brazil. In: Gepts P, editor. Genetic Resources of Phaseolus Beans. Kluwer, Dordrecht, Netherlands. 1988. p. 467-483.

97. Vlasova A, Capella-Gutierrez S, Rendon-Anaya M, Hernandez-Onate M, Minoche AE, Erb I, Camara F, Prieto-Barja P, Corvelo A, Sanseverino Wet al: Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene

duplications in establishing tissue and temporal specialization of genes. Genome Biology 2016, 17.

98. Wakeley J: The variance of pairwise nucleotide differences in two populations with migration. Theoretical Population Biology 1996, 49(1):39-57.
99. Watterson GA. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 1975, 7: 256–276
100. Weir BS, Cockerham CC: Estimating F-statistics for the analysis of population-structure. Evolution 1984, 38(6):1358-1370.
101. Willing EM, Hoffmann M, Klein JD, Weigel D, Dreyer C: Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics 2011, 27(16):2187-2193.
102. Womble WH. Differential Systematics. Science 1951, 114:315-322.
103. Würschum T, Langer SM, Longin CFH, Korzun V, Akhunov E, Ebmeyer E, Schachschneider R, Schacht J, Kazman E, Reif JC: Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. Theoretical and Applied Genetics 2013, 126(6):1477-1486.
104. Yang JA, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW et al: Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 2010, 42(7):565-U131.
105. Yang JA, Lee SH, Goddard ME, Visscher PM: GCTA: A Tool for Genome-wide Complex Trait Analysis. American Journal of Human Genetics 2011, 88(1):76-82.
106. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y et al: Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nature Biotechnology 2015, 33(4):408-U125.
107. Zou J, Raman H, Guo SM, Hu DD, Wei ZL, Luo ZL, Long Y, Shi WX, Fu Z, Du DZ et al: Constructing a dense genetic linkage map and mapping QTL for the traits of flower development in *Brassica carinata*. Theoretical and Applied Genetics 2014, 127(7):1593-1605.

**Supplementary Material 1 -** Identification of the common bean accessions used in the SNP analysis, the gene pool origin, type of germplasm, institution of origin, and the commercial type of grain.

**Supplementary Material 2** – SNP quality parameters for the whole set of accessions.

**Supplementary Material 3 -** Functional annotation of transcripts affected by SNPs with high or moderate predicted impact.

**Supplementary Material 4** – Enzymes associated with SNP sequences with high and moderate impact predicted.

**Supplementary Material 5 -** The KEGG pathway maps for SNPs with high or moderate predicted impact.

**Supplementary Material 6** – Tajima's D, number of segregating sites, Watterson's - $\theta_W$, nucleotide diversity, and diversity from Nei ($\pi$) statistics with loess smoothing of non-overlapping 100 Kb sliding windows.

**Supplementary Material 7** - Functional annotation of the transcripts affected by outlier SNPs.

**Supplementary Material 8** – The KEGG pathway associated with outlier SNPs.

**Supplementary Material 9** - Functional annotation of transcripts affected by outlier SNPs in the Mesoamerican gene pool.

**Supplementary Material 10** - Functional annotation of transcripts affected by outlier SNPs in the Andean accessions.

**Supplementary Material 11** – Positioning of the outlier SNPs on the *P. vulgaris* genome.

**Supplementary Material 12 -** Information of the selected 560 SNPs to compose a panel for genetic analysis of common bean.

CAPÍTULO II

# Análise de associação genômica ampla (GWAS) para caracteres de tolerância à seca em feijoeiro comum

# Genome-wide association study (GWAS) for drought tolerance in common bean

## RESUMO

A seca é um dos mais importantes estresses abióticos que afetam o desenvolvimento das plantas e reduzem a produtividade agrícola em todo o mundo. O déficit hídrico é o segundo maior causador da queda de produtividade em feijoeiro comum, afetando cerca de 60% das áreas de produção do grão, demandando grandes esforços para desenvolver cultivares com maior tolerância a esse estresse. A tolerância à seca é um caráter complexo que envolve várias modificações fisiológicas, bioquímicas, celulares e moleculares nas plantas. A análise de associação genômica ampla (GWAS) é um método que viabiliza identificar regiões genômicas associadas a caracteres de interesse, através da associação estatística entre os marcadores moleculares e o fenótipo. O objetivo deste estudo foi realizar GWAS para a identificação de locos associados a caracteres relacionados à produtividade de grão em condições de seca. Para isso, um conjunto de 343 acessos de feijoeiro comum de origem Mesoamericana foram genotipados utilizando 8789 SNPs derivados das tecnologias DArTseq e Capture-Seq. Esses mesmos acessos foram avaliados em campo, em ambientes com e sem deficiência hídrica, para os caracteres massa de 100 grãos e produtividade, por três anos consecutivos. As análises de associação foram realizadas usando o modelo linear misto implementado no software Tassel resultando na identificação de 93 e 139 SNPs associados com os caracteres avaliados, nas análises individuais e conjuntas, respectivamente. Para massa de 100 grãos foram identificados 213 SNP associados, enquanto para produtividade foram 19 SNPs. A variância fenotípica, explicada por cada SNP, variou de 4,46 a 10,25% para massa de 100 grãos e de 8 a 14,39% para produtividade. Através da anotação gênica foram identificados SNPs associados a genes com efeitos putativos como resposta à deficiência hídrica, tais como fatores de transcrição (MYB e bZIP), proteínas associadas ao metabolismo e transporte de membrana de ácido abscíssico (ABA). Quanto ao impacto predito para os efeitos dos SNPs, 80,13% foram do tipo modificador, 10,8% foram considerados de baixo impacto, 8% de impacto moderado e 1,08% de alto impacto. Os resultados obtidos revelaram genes com funções biológicas relacionadas ao processo de tolerância à seca sugerindo a identificação de importantes SNPs candidatos que poderão ser submetidos a um processo de validação e, posteriormente, implementados na Seleção Assistida por Marcadores (SAM). Adicionalmente, nesse estudo é disponibilizado um painel de associação que poderá ser utilizado para análise de outros caracteres de interesse para o melhoramento genético do feijoeiro comum.

## ABSTRACT

Drought is one of the most important abiotic stresses that affect plant development and reduce agricultural productivity worldwide. Water stress is the second major cause of the fall in productivity in common bean, affecting about 60% of the grain production areas, demanding great efforts to produce cultivars with greater tolerance to this abiotic stress. Drought tolerance is a complex trait involving various physiological, biochemical, cellular and molecular changes in plants. The genome-wide association studies (GWAS) is a method that makes it possible to identify genomic regions associated with the trait of interest, through the statistical association between the molecular markers and the phenotype. The objective of this study was to perform GWAS for the identification of loci associated to productivity traits in drought conditions. For this, a set of 343 common bean accessions of Mesoamerican origin were genotyped using 8,789 SNPs derived from DArTseq and Capture-Seq technologies. These same accessions were evaluated in the field, in environments with and without water deficiency, for the traits of 100-seed weight and yield, for three consecutive years. The association analyzes were performed using the mixed linear model implemented in the Tassel software and resulted in the identification of 93 and 139 SNPs associated with the evaluated traits in the individual and joint analyzes, respectively. For 100-seed weight, 213 SNPs were identified, while for yield, 19 SNPs were identified. The phenotypic variance, explained by each SNP, ranged from 4.46 to 10.25% for 100-seed weight and from 8.00 to 14.39% for yield. Through gene annotation, SNPs associated with genes with putative effects in response to water deficiency, such as transcription factors (MYB and bZIP), proteins associated with metabolism and membrane transport of abscisic acid (ABA) were identified. Regarding the impact predicted for the effects of SNPs, 80.13% were of the modifier type, 10.8% were considered low impact, 8% moderate impact and 1.08% with high impact. The results obtained revealed genes with biological functions related to the drought tolerance process suggesting important candidate SNPs that could be submitted to a validation process and, later, implementation in marker-assisted selection (MAS). Additionally, in this study an association panel is available that can be used to analyze different traits in common bean.

## INTRODUÇÃO

A seca é um dos mais importantes estresses abióticos que afetam o desenvolvimento das plantas e reduzem a produtividade agrícola em todo o mundo (Shao et al 2009). Segundo dados da FAO (Food and Agriculture Organization of the United Nations; 2015), atualmente, cerca de 40% das perdas de produção são causadas pela seca. Há previsões de que, devido às mudanças climáticas, ocorra o aumento da incidência e duração da seca nas principais áreas agrícolas, afetando negativamente os rendimentos das culturas e a segurança alimentar (McClean et al. 2011, Lauer et al. 2012, Lesk et al. 2016). Esse fato se torna ainda mais preocupante diante da previsão de escassez de chuva, principalmente nas regiões tropicais, onde estão localizados países em desenvolvimento com atuais níveis críticos de pobreza e desnutrição (Cavalieri et al. 2011). Diferentemente dos países desenvolvidos, os quais tem capital para investir em sistemas de irrigação para mitigar os estresses abióticos, os países em desenvolvimento carecem de infraestrutura adequada para minimizar os efeitos das alterações climáticas e os danos podem resultar em graves riscos para a segurança alimentar (McClean et al. 2011).

Dentre as leguminosas, o feijoeiro comum (*Phaseolus vulgaris* L.) destaca-se como um alimento de alto valor nutritivo e de grande importância econômica e social. Além de ser uma fonte rica em proteínas e fibras, sendo considerado o mais importante legume de grãos para o consumo humano direto no mundo (Broughton et al. 2003), o feijoeiro comum é uma cultura com alta capacidade de fixação biológica de nitrogênio, contribuindo para o aumento da sustentabilidade ambiental dos sistemas de cultivo (FAO 2016a). Trata-se de um alimento presente na dieta de mais de meio bilhão de pessoas na América Latina e na África, fornecendo até 15% do total de calorias diárias e 36% do total de proteínas diárias (Schmutz et al. 2014), sendo cultivado em 126 países, com uma área plantada anual de aproximadamente 30,6 milhões de hectares (FAO 2016b). Aproximadamente, 15% dessas áreas estão localizadas em regiões onde a seca é severa, como no Brasil, no litoral peruano, no norte do México e em regiões secas da África (Singh 2005). Para o feijoeiro comum, as doenças representam a maior causa de queda de produtividade sendo seguida pelo estresse hídrico (Singh 1995), o qual afeta cerca de 60% das áreas de plantio e é resultante tanto de períodos de estiagem, durante o ciclo de cultivo da cultura, quanto da irregularidade em precipitações (Graham e Ranalli 1997, McClean et al. 2011). O aumento da vulnerabilidade dessa cultura à seca também pode ser atribuido a presença de um sistema radicular superficial (Ramirez-Vallejo e Kelly 1998), no qual 50% das raízes são encontradas nos 30 cm de

profundidade do solo e menos de 10% atingem profundidade de 46 cm (Halterlein 1983), limitando a absorção de água contida nas camadas mais profundas do solo.

As plantas utilizam diferentes estratégias adaptativas frente ao estresse de déficit hídrico, tais como o escape ou ativação de mecanismos de tolerância e recuperação (Levitt 1972). Em feijoeiro comum, várias características associadas à tolerância à seca já foram identificadas em algumas variedades, incluindo o desenvolvimento de sistema radicular mais extenso possibilitando extrair água das camadas mais profundas do solo (Sponchiado et al. 1989, Frahm et al. 2004); (b) o acúmulo de biomassa e translocação de biomassa armazenada para a semente (Ramirez-Vallejo e Kelly 1998, Rao et al. 2004, Polania et al. 2016); (c) o ajuste de características fenológicas, como número de dias para florescimento e uso eficiente da água disponível para fotossíntese, crescimento e desenvolvimento (Acosta-Gallegos e Shibata 1989, Acosta-Gallegos e Adams 1991, Acosta-Gallegos e White 1995, Rosales-Serna et al. 2004); (d) e o ajuste de mecanismos fisiológicos relacionados à condutância estomática, área foliar e ajustamento osmótico (Beebe et al. 2010, Lanna et al. 2016).

A tolerância à seca é uma característica de herança complexa e de baixa herdabilidade (Schneider et al. 1997, Blair et al. 2012). Em termos práticos, a seleção para a tolerância à seca é complicada porque ela se manifesta de formas diferentes devido à duração e intensidade do estresse hídrico e pode ser potencializada por outros fatores, tais como baixa fertilidade e acidez do solo, doenças e altas temperaturas (Blum 2011). Fontes de alelos favoráveis para tolerância à seca podem estar disponíveis nas coleções silvestres e variedades tradicionais de feijoeiro comum presentes nos bancos de germoplasma, devido a sua alta diversidade genética e variabilidade fenotípica (Gepts et al. 2008, Vermeulen et al. 2012). Para resolver os problemas de seca intermitente e terminal, os programas de melhoramento estão aumentando os esforços para aprimorar os ganhos genéticos do feijoeiro comum em condições limitantes de água (Beebe et al. 2010). Embora linhagens/cultivares tolerantes à seca tenham sido desenvolvidas através de métodos tradicionais de melhoramento (Singh et al. 2001, 2007; Brick et al. 2008), existe um grande potencial para o desenvolvimento de novas cultivares através da seleção assistida por marcadores para acelerar o melhoramento para tolerância a este estresse (Frahm et al. 2004, Muñoz-Perea et al. 2006).

Não obstante os esforços mundiais para reduzir o efeito da seca na produtividade das espécies cultivadas, a precisa correlação entre os genes e a tolerância à seca permanece por ser demonstrada. Os marcadores moleculares são fortes aliados para analisar o controle genético de características complexas, como a tolerância à seca (Blair et al. 2012). Na maioria dos estudos, os mapas genéticos derivados de populações biparentais fornecem a base para o

mapeamento de locos de caracteres quantitativos (Quantitative Trait Loci, QTLs). Os primeiros mapas genéticos utilizados na identificação de caracteres associados à tolerância à seca em feijoeiro comum eram baseados em marcadores moleculares do tipo RAPD, AFLP, SSR (Scheineder et al. 1997, Blair et al. 2012, Asfaw e Blair 2012). Mais recentemente, utilizando-se de marcadores SNPs (Single Nucleotide Polymorphism), mapas genéticos com maior resolução estão sendo desenvolvidos e têm possibilitado a identificação de QTLs em condições de seca com maior resolução (Mukeshimana et al. 2014, Trapp et al. 2015, Villordo-Pineda et al. 2015).

Atualmente, o emprego de tecnologias de sequenciamento de nova geração (Next Generation Sequencing, NGS), aliado a métodos de redução de complexidade do genoma, tem possibilitado a identificação de milhares de SNPs amplamente distribuídos ao longo do genoma (Varshney et al. 2014). Recentemente, diferentes abordagens de genotipagem por sequenciamento em plantas estão sendo disponibilizadas, tais como RADseq (Willing et al. 2011), GbS (Elshire et al. 2011) e DArTseq (Cruz et al. 2013). A abordagem de redução da complexidade do genoma consiste no uso de enzimas de restrição e seleção dos fragmentos para o desenvolvimento das bibliotecas genômicas, seguido por NGS. Posteriormente, as sequências são alinhadas permitindo a identificação dos SNPs entre as amostras. A metodologia de DArTseq se diferencia das demais por utilizar duas enzimas de restrição, uma de corte frequente e outra de corte raro e sensível à metilação, fragmentando, preferencialmente, regiões hipometiladas que, geralmente, são enriquecidas de genes (Rabinowicz et al. 2005). Também aliada à tecnologia de NGS, a análise de sequenciamento de regiões genômicas alvo permite a identificação simultânea de milhares de SNPs presentes em genes alvos através da análise de captura, denominada de Capture-Seq (Neves et al. 2013).

O acesso crescente às diversas tecnologias de genotipagem em larga escala têm resultado em um grande volume de estudos focados na análise de associação genômica ampla (Genome-Wide Association Studies, GWAS). Os métodos estatísticos utilizados na análise de mapeamento associativo estimam a força de associação entre um marcador e a característica alvo mensurada. A relação é detectada quando um QTL, que está associado à característica em estudo, está em desequilíbrio de ligação com o loco do marcador. Em plantas, a metodologia usual envolve o uso de indivíduos de populações naturais ou coleções de germoplasmas que apresentam uma ampla variação fenotípica para a característica alvo (Abdurakhmonov e Abdukarimov 2008). Portanto, GWAS considera mais eventos de recombinação ao usar um painel de indivíduos com histórias de recombinações únicas, com

blocos de desequilíbrio de ligação (DL) menores do que os de uma população biparental, proporcionando um mapa com maior resolução (Moghaddam et al. 2016).

A metodologia de GWAS pode ser aplicada de modo efetivo no feijão por ser uma espécie autógama, possuir um genoma de referência de qualidade (Schmutz et al. 2014, Vlasova et al. 2016) e uma grande variabilidade genética armazenada em bancos de germoplasma. Recentemente, estudos de GWAS têm sido aplicados em feijoeiro comum e, devido a forte estruturação genética em dois pools gênicos, Andino e Mesoamericano (Mamidi et al. 2013, Schmutz et al. 2014), os estudos vêm sendo conduzidos, preferencialmente, dentro de cada pool gênico, separadamente. Estudos de associação para produtividade em feijão foram realizados por Schmutz et al. (2014) utilizando um painel de 271 cultivares Mesoamericanas avaliadas com 34799 SNPs; e por Moghaddam et al. (2016) usando um painel de 280 acessos Mesoamericanos e mais de 150000 SNPs. Nessas análises foram identificados QTLs associados a diversos caracteres agronômicos que afetam a produção em feijão, tais como dias para florescimento, dias para maturidade, hábito de crescimento, altura da copa, acamamento e peso de sementes. Outros estudos de GWAS utilizaram acessos de origem Andina e possibilitaram a identificação de regiões genômicas associadas ao tempo de cocção (Cichy et al. 2015), fixação biológica de nitrogênio (Kamfwa et al. 2015) e resistência à antracnose (Zuiderveen et al. 2016).

Os objetivos deste trabalho foram: 1) identificar SNPs polimórficos e com elevada representatividade do genoma de feijão através das tecnologias de DArTseq e Capture-Seq; 2) estimar os parâmetros e estrutura genética do germoplasma; 3) realizar a análise dos dados fenotípicos dos experimentos conduzidos na presença e ausência de deficiência hídrica; 4) realizar uma análise de associação genômica ampla (GWAS) para a identificação de locos associados à produtividade em condições de seca; 5) caracterizar os genes que estejam co-segregando significativamente com os SNPs associados à produtividade.

## MATERIAL E MÉTODOS

### Material Vegetal

O material vegetal consistiu num total de 343 acessos de feijoeiro comum, pertencentes ao pool gênico Mesoamericano e que integram a Coleção Nuclear de Feijão (CONFE) da Embrapa. Desses, 227 são variedades tradicionais brasileiras e 116 são cultivares/linhagens desenvolvidas por instituições de pesquisa do Brasil e exterior (Material

Suplementar 1). Os acessos da CONFE foram plantados em casa de vegetação e amostras de tecido vegetal foram coletadas a partir de plantas individuais. O DNA genômico total foi obtido utilizando o produto comercial Invisorb® Spin Plant Mini Kit (Stratec, Berlim), seguindo as orientações do fabricante. Amostras de DNA foram avaliadas quanto à integridade e quantificadas utilizando gel de agarose 1%. Posteriormente as amostras, incluindo duas replicatas técnicas, foram enviadas para genotipagem utilizando as tecnologias DArTseq e Capture-Seq.

**Genotipagem e Imputação**

**Genotipagem DArTseq:** A metodologia de genotipagem baseada na tecnologia DArT (DArTseq) foi desenvolvida pela empresa Australiana DArT Pty Ltd (Bruce, Austrália). Para o desenvolvimento dos marcadores DArTseq foi construída uma biblioteca de fragmentos de DNA a partir de 340 acessos Mesoamericanos, conforme detalhado por Sanchez-Sevilla et al. (2015). O método baseou-se na redução da complexidade do genoma utilizando as enzimas de restrição PstI-MseI. As amostras digeridas foram ligadas a adaptadores específicos aos primers da etapa de enriquecimento por PCR (reação em cadeia da polimerase), à *Illumina flowcell* e com *barcodes* para identificação posterior de cada amostra. Os fragmentos foram sequenciados em Illumina Hiseq2000. As sequências foram processadas usando o *pipeline* de propriedade da empresa DArT Pty Ltd. Os arquivos *fastqcall* resultantes foram usados em um *pipeline* secundário, também de propriedade da DArT Pty Ltd (DArTsoft-seq) contendo algoritmos para chamada dos SNPs.

**Genotipagem Capture-Seq:** A metodologia de captura de alvos específicos (Capture-Seq) seguida por NGS foi conduzida na empresa RAPiD Genomics (Gainesville, FL, EUA). Um conjunto de 5293 transcritos diferencialmente expressos, sob estresse de seca (Pereira et al., em elaboração), foi alvo de sequenciamento para uma amostra composta de 175 acessos Mesoamericanos de feijoeiro comum, selecionados como os mais divergentes geneticamente, a partir de análise com microssatélites (SSR, Simple Sequence Repeats) (Cascão et al. 2014). Para capturar estas regiões, as sondas foram desenhadas utilizando um genoma de referência de *P. vulgaris* ("Pvulgaris_218_v1.0.fasta"). As etapas iniciais de desenvolvimento das sondas foram realizadas de maneira a selecionar aquelas que não se sobrepõem e que passam pela filtragem em nível de sequências tais como, o conteúdo de GC (guanina/citosina) e o teor de homopolímero. Estas sondas foram ainda alinhadas contra o genoma de referência do

feijoeiro comum para verificar o seu estado de singularidade (*hit*) e remover as sondas que mapearam em genomas de organelas ou em elementos repetitivos. Para 4766 genes, houve uma ou mais sondas com um único *hit* no genoma e, uma sonda foi escolhida aleatoriamente entre essas. Para o restante dos genes, foram escolhidas sondas com dois ou, no máximo, três *hits* para certificar-se de que, para todos os genes havia, pelo menos, uma sonda desenhada. Finalmente, apenas uma sonda por gene foi selecionada, aleatoriamente, resultando em 5050 sondas em 5050 genes. Após o enriquecimento e NGS, os dados foram alinhados ao genoma de referência utilizado para o desenho das sondas e os SNPs foram identificados e genotipados com base nos resultados do alinhamento (Neves et al. 2013). Os dados brutos consistiram de sequências brutas e arquivos de dados filtrados pelos parâmetros de qualidade da empresa RAPiD Genomics.

**Imputação:** A imputação dos SNPs foi realizada através do software NGSEP v 3.0.1 (Duitama et al. 2014) usando o parâmetro c = 0,003 (estimativa do número médio de centiMorgans por kpb na regiões eucromáticas do genoma) e número de clusters igual a 20 (número máximo de grupos nos quais os haplótipos locais serão agrupados). A análise de imputação baseou-se no modelo HMM (Hidden Markov Model; Scheet e Stephens 2006), usando o desequilíbrio de ligação entre os SNPs em um haplótipo. Este software é indicado para populações endogâmicas em que a heterozigosidade é baixa. Para avaliar a acurácia do processo de imputação, 10% dos genótipos foram mascarados, aleatoriamente, e nova imputação foi realizada para posterior comparação entre os genótipos estimados com os genótipos atuais mascarados.

**Distribuição genômica dos SNPs**

As regiões genômicas que flanqueiam os SNPs identificados foram alinhadas contra o genoma de *P. vulgaris* v. 1.0 (Schmutz et al. 2014), usando o BLASTN com E-value ≤ 1,0E-25 (Altschul et al. 1997), para identificar as posições dos SNPs no genoma de referência. Para os SNPs que alinharam em mais de uma região no genoma, a posição do *hit* com menor e-value e maior porcentagem de identidade foram usados nas análises posteriores. Uma figura de distribuição dos SNPs ao longo do genoma foi gerada usando scripts específicos no programa R v 3.0.1 (R Core Team 2015).

**Avaliação dos componentes de produtividade**

Os experimentos de campo para avaliação dos componentes de produtividade, em ambientes com e sem deficiência hídrica, foram conduzidos no Sítio de Fenotipagem para Tolerância à Deficiência Hídrica situado na Estação Experimental da EMATER, localizada no município de Porangatu (49°06'W, 13°18'S, 396 m de altitude), estado de Goiás. Os experimentos foram conduzidos em três anos consecutivos (2014, 2015 e 2016), no período de entressafras (maio-agosto), quando a precipitação pluviométrica é praticamente nula permitindo total controle da água no solo.

Em 2014, foram avaliados 580 acessos e quatro testemunhas (BRS Estilo, BRS Esplendor, BRS Embaixador e Jalo Precoce) com delineamento de blocos aumentados de Federer (BAF) composto por 20 blocos (29 tratamentos e quatro testemunhas por bloco). Em 2015 e 2016 foram avaliados 324 acessos e quatro testemunhas (mesmas do ensaio de 2014) com delineamento de látice quadrado com duas repetições, sendo cada experimento composto por 18 blocos com 18 tratamentos. Todos os experimentos foram semeados em parcelas de uma linha de três metros de comprimento e espaçadas de 40 cm. A densidade de semeadura foi de 15 a 18 sementes por metro. A demanda das plantas por nitrogênio, fósforo e potássio foi suprida com a aplicação de 16, 120 e 64 kg $ha^{-1}$ de N, $P_2O_5$ e $K_2O$, respectivamente. A adubação de cobertura foi efetuada com 30 kg $ha^{-1}$ de N, aos 20 dias após a emergência. O controle de plantas daninhas foi efetuado com 250 g $ha^{-1}$ de fomesafen e 187 g $ha^{-1}$ de fluazifop-p-butyl, após a emergência.

Em cada ano, os experimentos foram conduzidos na presença e ausência de déficit hídrico. Os experimentos sem deficiência hídrica foram adequadamente irrigados durante todo o desenvolvimento da planta. Os períodos de irrigação foram controlados através do uso de tensiômetros e, lâminas de água de aproximadamente 25 mm, foram aplicadas quando o potencial da água no solo a 0,15 m de profundidade atingia - 0,035 Mpa (Silveira e Stone 1994). Para os experimentos com deficiência hídrica, a irrigação foi mantida em condições adequadas de água no solo até os 20 dias após a emergência da planta. A partir desse período, até o momento da colheita, foi imposto o déficit hídrico aplicando-se lâminas de água de aproximadamente 25 mm quando o potencial da água no solo a 0,15 m de profundidade atingia – 0,07 MPa.

Foram avaliados os seguintes caracteres nos ensaios de campo: a) Produtividade de grão (Prod) obtida através do peso de grãos por parcela, posteriormente transformados para kg $ha^{-1}$; b) Massa de 100 grãos que correspondeu à determinação do peso de 100 grãos, em gramas.

**Análise estatística dos dados fenotípicos**

Para obtenção de componentes de variância e médias ajustadas, foram realizadas análises individuais, por ambiente, e conjunta, envolvendo os ambientes para cada condição hídrica, para o conjunto de dados fenotípicos. Para análise individual, considerando o modelo geral com efeitos de repetição e bloco dentro de repetição, foi ajustado o seguinte modelo linear misto:

$$y_{ijkm} = \mu + r_j + b_{k/j} + g_{i/m} + \varepsilon_{ijk}$$

em que $y_{ijk}$ é o efeito do genótipo *i* na repetição *j* e no bloco *k*, sendo o genótipo pertencente ao grupo *m*; $\mu$ é a média geral; $r_j$ é o efeito aleatório da repetição *j*; $b_{k(j)}$ é o efeito aleatório do bloco *k* dentro da repetição *j*; $g_{i(m)}$ é o efeito do genótipo *i* (acesso, como de efeito aleatório, ou cultivar testemunha, como de efeito fixo), dentro do grupo *m*, de efeito fixo; e $\varepsilon_{ijk}$ é o erro experimental associado à observação *ijkm*.

Para análise conjunta foi ajustado o seguinte modelo linear misto:

$$y_{ijklm} = \mu + l_l + r_{j/l} + b_{k/j/l} + g_{i/m} + gl_{il} + \varepsilon_{ijk}$$

em que $y_{ijlk}$ é o efeito do genótipo *i* na repetição *j* no bloco *k* no ambiente *l*, sendo o genótipo pertencente ao grupo *m*; $\mu$ é a média geral; $l_l$ é o efeito fixo do local *l*; $r_{j/m}$ é o efeito aleatório da repetição *j* dentro do ambiente *l*; $b_{k/j/l}$ é o efeito aleatório do bloco *k* dentro da repetição *j* no ambiente *l*; $g_{i(m)}$ é o efeito do genótipo *i* (acesso, como de efeito aleatório, ou cultivar testemunha, como de efeito fixo), dentro do grupo *m*, de efeito fixo; $gl_{il}$ é o efeito do genótipo *i* e o ambiente *l*; e $\varepsilon_{ijk}$ é o erro experimental associado à observação *ijklm*.

As estimativas de componentes de variância foram obtidas pelo método REML (Restricted Maximum Likelihood), segundo Patterson e Thompson (1971), e a predição dos valores genéticos de cada acesso baseou-se no procedimento BLUP (Best Linear Unbiased Prediction), segundo Henderson (1984). De posse das estimativas de componentes de variância fenotípica, genética e residual, foi calculada a estimativa de herdabilidade, no sentido amplo, em nível de médias de acessos. Também foi calculada a estimativa de acurácia da seleção ($r_{g\hat{g}}$), correspondente à raiz quadrada da estimativa de herdabilidade, segundo Resende e Duarte (2007).

As análises para ajuste do modelo linear misto foi realizada com uso do pacote lme4 da plataforma R (R Core Team 2015). As médias ajustadas, valores BLUP, de acessos,

para os caracteres produtividade e massa de 100 grãos, foram utilizados como valores fenotípicos no estudo de GWAS. A partir dos dados médios obtidos, foi calculado o índice de susceptibilidade à seca (ISS), conforme proposto por Fischer e Maurer (1978), através da seguinte fórmula:

$$ISS = [1-(Yce/Yse)] / [1-(Mce/Mse)],$$

em que, Yce = produtividade com estresse hídrico, Yse = produtividade sem estresse hídrico, Mce = média de todos os genótipos com estresse hídrico, Mse = média de todos os genótipos sem estresse hídrico.

Com base nas médias preditas de massa de 100 grãos e produtividade, foram feitos gráficos do tipo Boxplot para os ambientes com e sem deficiência nos três anos avaliados, através do software R v 3.0.1 (R Core Team 2015).

**Estimativas de desequilíbrio de ligação (DL)**

As estimativas de DL par-a-par foram calculadas pela medida clássica da correlação ao quadrado das frequências alélicas de locos bialélicos ($r^2$), e corrigidas pelo viés devido ao parentesco e estrutura populacional ($r^2_{SV}$) utilizando o pacote LDcorSV (Desrousseaux et al. 2013). A matriz de parentesco (Genetic Relationship Matrix, GRM) foi estimada pelo software GCTA (Yang et al. 2011) usando o algoritmo proposto por Yang et al. (2010). O decaimento do DL para metade do valor máximo ($r^2 \sim 0,23$) foi explicado pelo modelo não linear proposto por Hill e Weir (1988), ajustado com a função *nls* do software R v 3.0.1 (R Core Team 2015).

**Diversidade e estruturação genética**

Os parâmetros de heterozigosidade observada e esperada (diversidade genética de Nei) foram estimados através da função *bootstrapHe*t do pacote PopGenKit v 1.0 (Paquette 2012) do software R. A estruturação genética foi determinada utilizando o método Bayesiano, implementado no software Structure versão 2.3.4 (Pritchard et al. 2000), usando apenas os SNPs que não estão em desequilíbrio de ligação ($r^2 < 0,5$), os quais foram selecionados através do comando *LD Prunning* no software Plink v 1.07 (Purcell et al. 2007). O Structure foi executado de acordo com o modelo admixture usando 100000 interações MCMC (Markov

Chain Monte Carlo), após um período de burn-in de 100000 interações, com um número de população (K) variando de 1 a 10, com 20 interações cada. A estatística *DeltaK*, proposta por Evanno et al. (2005), foi usada para determinar o valor mais provável para K no programa Structure Harvester v 0.6.93 (Earl e vonHoldt 2012). Um *heatmap* mostrando o parentesco entre os genótipos foi gerado usando o algoritmo VanRaden implementado no software GAPIT (Lipka et al. 2012).

**GWAS**

A análise de GWAS foi realizada pelo programa Tassel versão 5.2.31 (Bradbury et al. 2007), por meio do módulo *Mixed Linear Model* (MLM), considerando os SNPs e a estrutura populacional como fatores de efeito fixo, enquanto a matriz de parentesco foi considerada como fator de efeito aleatório. A matriz de parentesco (matriz K ou *kinship*) foi calculada através do algoritmo *Identity by State* (IBS) no software Tassel e a matriz de estrutura populacional (matriz Q) pelo Structure. Para a confirmação da significância das associações entre marcadores SNPs e caracteres fenotípicos foi utilizado o método FDR (false discovery rate), através do software Qvalue versão 1.0 (Storey 2002) do software R v 3.0.1 (R Core Team 2015). Um gráfico de Manhattan foi gerado para os *p*-valores através do pacote *qqman* v 0.1.2 (Turner 2014) no programa R v 3.0.1 (R Core Team 2015).

**Anotação gênica dos SNPs significativos**

A anotação funcional dos genes putativos afetados pelos SNPs significativamente associados à tolerância à seca, detectados nas análises GWAS, foi feita através da ferramenta Blast2GO v.3.2 (Conesa et al. 2005). Os transcritos identificados foram caracterizados utilizando-se os termos do Gene Ontology Consortium (Ashburner et al. 2000; Consortium 2015). A anotação dos efeitos preditos pelos SNPs significativos foi realizada através do software SnpEff v 4.2 (Cingolani et al. 2012) baseado no banco de dados Phytozome (Goodstein et al. 2012).

**RESULTADOS**

**Genotipagem**

Um total de 14407 marcadores SNPs foram obtidos a partir da genotipagem de 340 genótipos Mesoamericanos via DArTseq. As médias de homozigotos, heterozigotos e conteúdo de polimorfismo (PIC) foram de 0,98, 0,01 e 0,24, respectivamente. As médias de dados perdidos e frequência alélica mínima (MAF) foram de 7% e 16,23%, respectivamente. A reprodutibilidade dos dados foi elevada, com valores variando de 94,2 a 100%, com média de 99,80%.

Através da tecnologia de Capture-Seq foram avaliados 175 genótipos Mesoamericanos, gerando um conjunto de 10670 SNPs em 3304 sondas, com uma média de 3,2 SNPs por sonda. Utilizando uma análise mais rigorosa, foram declarados 3716 SNPs (34,8%) a partir de 1701 sondas (51,5%), com os parâmetros de qualidade: (1) valor mínimo de 10 para qualidade na escala phred; (2) profundidade média de leitura mínima de 15 X e máxima de 750 X; (3) profundidade de leitura dos genótipos para cada marcador maior que três; (4) MAF $\geq$ 5% e (5) dados faltantes menor que 40%.

A taxa de erro de genotipagem entre as replicatas técnicas, desconsiderando os dados faltantes, foi de 0,43% e 2,81%, para as tecnologias DArTseq e Capture-Seq, respectivamente. Ao todo foram gerados 18123 SNPs, dos quais 15694 SNPs (11989 SNPs-DArTseq e 3705 SNPs-Capture-Seq) foram imputados para 343 genótipos. Destes, 14322 apresentaram MAF $\geq$ 0,01 e 8789 com MAF $\geq$ 0,05. Para as análises subsequentes, foram considerados somente os SNPs com MAF $\geq$ 0,05. A acurácia da imputação foi estimada em 93%.

**Caracterização estrutural dos SNPs**

As sequências genômicas que flanqueiam os SNPs, obtidos por ambas as tecnologias, foram alinhadas no genoma de *P. vulgaris*. Do total de 14407 SNPs obtidos por DArTseq, 12639 (87,73%) apresentaram alinhamento no genoma, sendo 11371 (89,97%) com alinhamento único e 1267 (10,03%) com alinhamento múltiplo, variando de 2 a 188 alinhamentos, com 69 SNPs alinhados em *scaffolds*. Já os SNPs derivados de Capture-Seq (3716) foram todos alinhados no genoma, com 3705 alinhados nos 11 cromossomos de *P. vulgaris* e 11 alinhados em *scaffolds*. O número médio de SNPs por cromossomo (derivados de ambas as tecnologias DArTseq e Capture-Seq) foi de aproximadamente 1480, variando de 1163 no cromossomo 5 a 2003 no cromossomo 8 (Figura 1). Estimou-se, em média, um SNP a cada 31403 pb. Considerando-se somente os SNPs imputados e com MAF $\geq$ 0,05 (8789

SNPs), todos foram alinhados nos 11 cromossomos de feijoeiro comum, com uma média de 799 SNPs por cromossomo e um SNP a cada 58824 pb (Tabela 1).

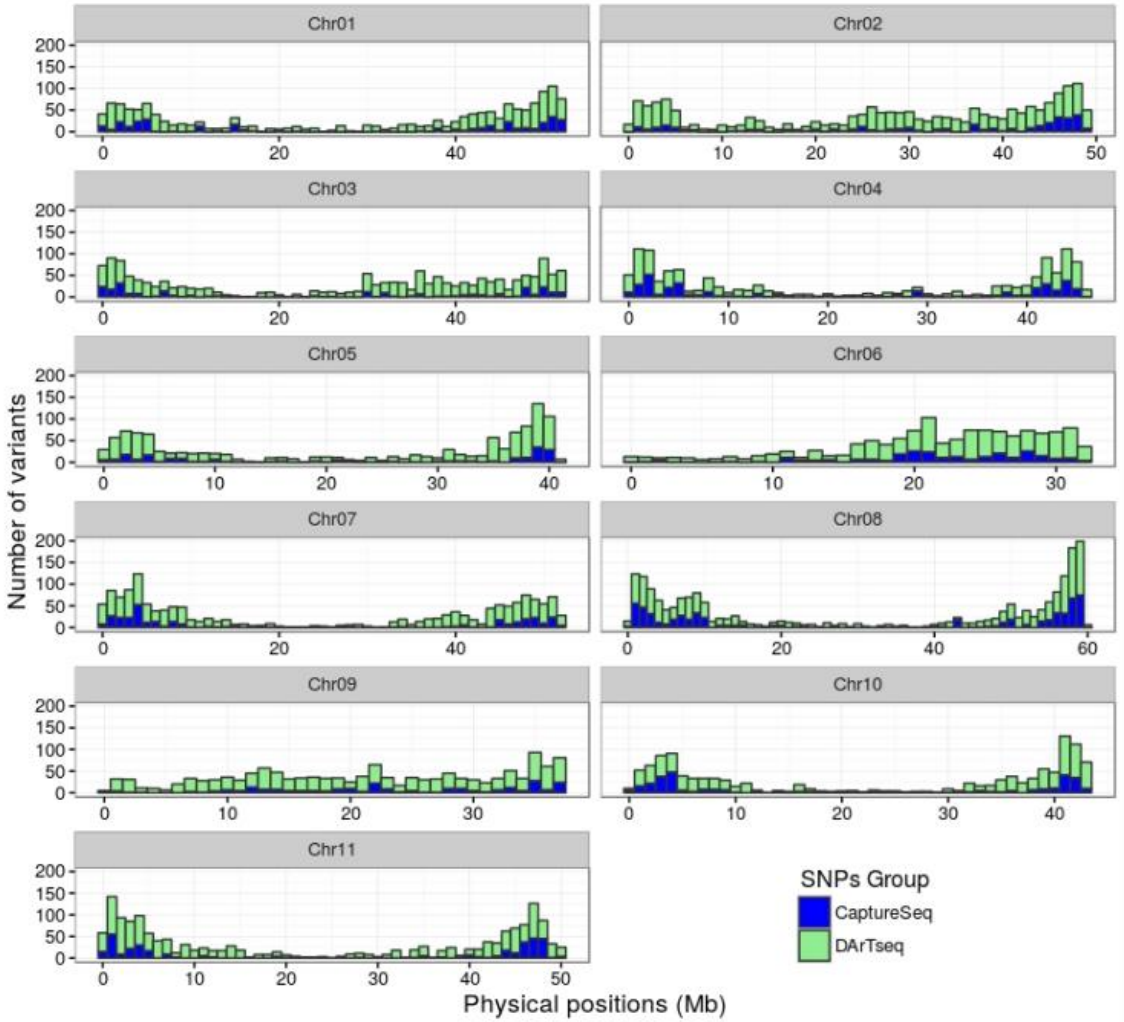



**Figura 1.** Distribuição dos 16355 SNPs, identificados pelas metodologias DArTseq e Capture-Seq, ao longo dos 11 cromossomos de *P. vulgaris*.

**Tabela 1.** Distribuição dos SNPs nos 11 cromossomos de *P. vulgaris*

| Cromossomo | Quantidade de SNPs total | | | Quantidade de SNPs pós-imputação com MAF ≥ 0,05 | | | Tamanho do cromossomo (kpb)* | Média de SNP (total) por Mpb | Média de SNP (pós-imputação com MAF ≥ 0,05) por Mpb |
|---|---|---|---|---|---|---|---|---|---|
| | DArTseq | Capture-Seq | Total | DArTseq | Capture-Seq | Total | | | |
| 1 | 1149 | 334 | 1483 | 533 | 292 | 825 | 52183,5 | 28,42 | 15,81 |
| 2 | 1505 | 360 | 1865 | 637 | 287 | 924 | 49033,7 | 38,04 | 18,84 |
| 3 | 1311 | 269 | 1580 | 571 | 243 | 814 | 52218,6 | 30,26 | 15,59 |
| 4 | 885 | 377 | 1262 | 508 | 306 | 814 | 45793,2 | 27,56 | 17,78 |
| 5 | 954 | 209 | 1163 | 406 | 172 | 578 | 40237,5 | 28,9 | 14,36 |
| 6 | 1015 | 270 | 1285 | 467 | 193 | 660 | 31973,2 | 40,19 | 20,64 |
| 7 | 1175 | 336 | 1511 | 478 | 271 | 749 | 51698,4 | 29,23 | 14,49 |
| 8 | 1371 | 632 | 2003 | 852 | 562 | 1414 | 59634,6 | 33,59 | 23,71 |
| 9 | 1100 | 249 | 1349 | 360 | 178 | 538 | 37399,6 | 36,07 | 14,39 |
| 10 | 891 | 307 | 1198 | 388 | 246 | 634 | 43213,2 | 27,72 | 14,67 |
| 11 | 1215 | 362 | 1577 | 519 | 320 | 839 | 50203,6 | 31,41 | 16,71 |
| Scaffolds | 68 | 11 | 79 | - | - | - | - | - | - |
| **Total** | **12639** | **3716** | **16355** | **5719** | **3070** | **8789** | **513589,1** | **31,84** | **17,11** |

* Schumtz et al. 2014

**Dados fenotípicos**

Os dados fenotípicos foram coletados em cinco ambientes, pois não houve produção de grão no experimento submetido à deficiência hídrica no ano de 2015, devido ao déficit hídrico excessivo e abortamento floral de todos os genótipos avaliados. Nas análises individuais, a produtividade média do experimento sem deficiência hídrica (SDH) em 2015 (1187 kg $ha^{-1}$) foi a menor quando comparada com os anos de 2014 (2391 kg $ha^{-1}$) e 2016 (1947 kg $ha^{-1}$). Nos experimentos com deficiência hídrica, a produtividade média em 2016 (798,8 kg $ha^{-1}$) foi superior a de 2014 (517,4 kg $ha^{-1}$). O experimento SDH de 2014 foi o que apresentou a maior média fenotípica, tanto para produtividade (2391 kg $ha^{-1}$) quanto para massa de 100 grãos (26,81 g), enquanto o experimento com deficiência hídrica (CDH) em 2014 foi o que apresentou médias fenotípicas menores, com produtividade média de 517,4 kg $ha^{-1}$ e massa de 100 grãos de 20,6 gramas (Figura 2). A herdabilidade ($h^2$) para massa de 100 grãos variou de 0,85 (CDH-2014) a 0,98 (SDH-2016), já para produtividade houve variação de 0,37 (CDH-2014) a 0,59 (SDH-2015). Portanto, o experimento CDH de 2014 apresentou as menores estimativas de herdabilidade para ambos os caracteres. As acurácias seletivas foram semelhantes entre todos os ambientes, para ambos os caracteres, apresentando altos valores para massa de 100 grãos (variando de 0,93 nos ambientes CDH-2014 e CDH-2016 a 0,99 no ambiente SDH-2016) e valores moderados para produtividade (variando de 0,71 no ambiente CDH-2016 a 0,76 no ambiente SDH-2015).

Na análise conjunta, as estimativas de produtividade média dos três anos de condução dos experimentos foram de 705,2 e 1870 kg $ha^{-1}$, nos ambientes com e sem deficiência hídrica, respectivamente. Portanto, os genótipos sofreram uma redução média de produtividade de 62,29%. Para massa de 100 grãos, os valores foram de 23,23 e 27,78 nos ambientes com e sem deficiência hídrica, respectivamente, com redução média de 16,38%. Informações sobre as produtividades médias e parâmetros genéticos de todos os experimentos estão descritos na Tabela 2.

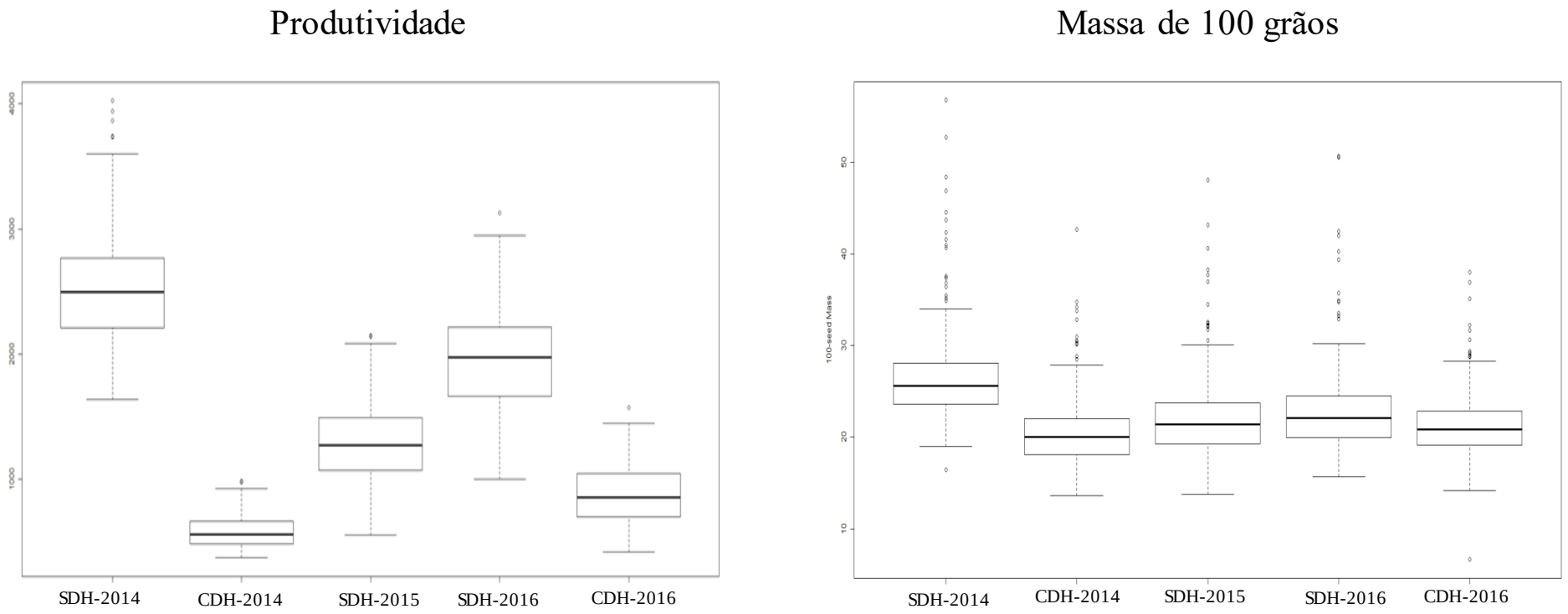



**Figura 2.** Boxplot das produtividades e massa de 100 grãos nos cinco experimentos conduzidos em Porangatu.

**Tabela 2.** Estimativas dos parâmetros genéticos nas análises fenotípicas individuais e conjuntas.

| Ambiente | Características | Nº de acessos | Média geral | Mínimo | Máximo | $h^2$ | $r_{gg}$ | CV |
|---|---|---|---|---|---|---|---|---|
| SDH-2014 | Produtividade | 533 | 2391 | 1638 | 4025 | 0,41 | 0,72 | 31,11 |
| | Massa de 100 grãos | 537 | 26,81 | 16,44 | 56,80 | 0,95 | 0,97 | 7,27 |
| CDH-2014 | Produtividade | 533 | 517,4 | 372 | 986 | 0,37 | 0,72 | 42,16 |
| | Massa de 100 grãos | 539 | 20,6 | 13,59 | 42,67 | 0,85 | 0,93 | 11,50 |
| SDH-2015 | Produtividade | 298 | 1187 | 557 | 2148 | 0,59 | 0,76 | 40,68 |
| | Massa de 100 grãos | 298 | 22,24 | 13,79 | 48,06 | 0,93 | 0,96 | 12,89 |
| SDH-2016 | Produtividade | 253 | 1947 | 999,64 | 3130 | 0,58 | 0,74 | 33,07 |
| | Massa de 100 grãos | 253 | 26,77 | 15,68 | 50,67 | 0,98 | 0,99 | 6,84 |
| CDH-2016 | Produtividade | 252 | 798,8 | 421,48 | 1575 | 0,57 | 0,71 | 42,37 |
| | Massa de 100 grãos | 254 | 21,31 | 6,63 | 37,99 | 0,91 | 0,93 | 11,38 |
| SDH conj | Produtividade | 531 | 1870 | 1528 | 2371 | 0,57 | 0,82 | 31,94 |
| | Massa de 100 grãos | 531 | 27,78 | 13,92 | 51,51 | 0,99 | 0,98 | 9,80 |
| CDH conj | Produtividade | 527 | 705,2 | 537 | 973 | 0,57 | 0,70 | 47,53 |
| | Massa de 100 grãos | 528 | 23,23 | 8,44 | 40,58 | 0,97 | 0,95 | 11,34 |

SDH: sem deficiência hídrica; CDH: com deficiência hídrica; conj: conjunta; Nº de acessos: número de acessos; $h^2$: herdabilidade; $r_{gg}$: acurácia seletiva; CV: coeficiente de variação.

De acordo com a análise conjunta dos valores médios das produtividades observadas em cada ambiente, nos três anos de condução dos experimentos, os genótipos foram classificados em quatro grupos (Tabela 3). Em condição de irrigação adequada, o grupo mais produtivo foi constituído por três cultivares/linhagens e sete variedades tradicionais, que no total produziram, em média, 2258 kg $ha^{-1}$. Sob essa mesma condição, o grupo menos produtivo foi composto por seis variedades tradicionais e quatro cultivares/linhagens, que produziram em média 1564 kg $ha^{-1}$, ou seja, 30,74% inferior ao do grupo anterior. Em condição de deficiência hídrica o grupo menos produtivo foi composto por seis variedades tradicionais e quatro cultivares/linhagens, que produziram em média 570 kg $ha^{-1}$. Sob essa mesma condição, o grupo mais produtivo compreendeu quatro cultivares/linhagens e seis variedades tradicionais, que produziram em média 893 kg $ha^{-1}$, ou seja, 63,8% superior à do grupo com menor produtividade. Três genótipos (Ouro Negro, CF200012 e CF800113) foram classificados no grupo mais produtivo dos tratamentos com e sem deficiência hídrica, com produção de 973, 873 e 871 kg $ha^{-1}$, respectivamente, no ambiente com deficiência hídrica e 2314, 2239 e 2262 kg $ha^{-1}$, respectivamente, no ambiente com irrigação adequada.

**Tabela 3.** Lista dos 10 genótipos com maior (10+) e menor (10-) produtividade, nas análises conjuntas, nos experimentos com e sem deficiência hídrica.

| Grupos | Sem deficiência hídrica (SDH) | | Com deficiência hídrica (CDH) | |
|---|---|---|---|---|
| | Genótipo | Produtividade | Genótipo | Produtividade |
| 10 + | CNF008845 | 2371,059 | Ouro Negro* | 973,0416 |
| | Ouro Negro* | 2314,025 | BRS Pontal | 905,3641 |
| | CF240056 | 2286,841 | CF871226 | 898,7706 |
| | CNF011036 | 2275,622 | CNF007143 | 894,4164 |
| | CF800113* | 2262,226 | CF250002 | 889,3753 |
| | CF200012* | 2239,854 | CNF007050 | 885,7935 |
| | CF840600 | 2225,91 | CF890223 | 880,9021 |
| | CF200048 | 2216,657 | CF200012* | 873,2352 |
| | CF870030 | 2208,872 | CF800113* | 871,1584 |
| | CF800110 | 2201,706 | CF860105 | 862,5147 |
| 10 - | CNF001611 | 1501,282 | CF840275 | 537,4872 |
| | CF200075 | 1528,191 | CNF007770 | 555,8042 |
| | CF840747 | 1545,479 | CF840650 | 556,6695 |
| | CNF006978 | 1555,991 | CNF007646 | 559,1432 |
| | CNF007770 | 1561,009 | CNF005887 | 561,7531 |
| | CF840275 | 1561,393 | CF890110 | 564,0479 |
| | CF840543 | 1561,518 | CF240032 | 567,7409 |
| | CF810080 | 1564,472 | CF240016 | 573,9301 |
| | CNF004121 | 1570,709 | CNF005482 | 575,1972 |
| | CF841226 | 1577,941 | CF800049 | 575,2152 |

* Genótipos que foram mais produtivos nos ambientes com e sem deficiência hídrica.

**Desequilíbrio de ligação (DL)**

Um total de 8789 SNPs, com uma média de 799 SNPs por cromossomo, com distâncias em pares de bases de até 59 Mb, foram utilizados nos cálculos de DL, resultando em cerca de 345 mil estimativas de $r^2$ par-a-par por cromossomo e mais de 3,78 milhões na escala genômica. O $r^2$ médio para cada cromossomo variou de 0,5043 (cromossomo 3) à 0,7874 (cromossomo 4), enquanto o $r^2_{sv}$ variou de 0,4687 (cromossomo 3) a 0,7676 (cromossomo 4) (Tabela 4).

O padrão de decaimento do desequilíbrio de ligação, em relação à distância entre os marcadores, foi investigado para cada um dos 11 cromossomos de *P. vulgaris* e, no geral, a distribuição do $r^2$ mostrou um rápido decaimento do DL conforme o aumento da distância física, com marcadores apresentando alto DL até a distância aproximada de 250 kb (Figura 3), variando de 163,39 kb no cromossomo 5 a 413,18 kb no cromossomo 3. Com a correção para estrutura e parentesco, a distância de decaimento do DL variou de 32,49 kb a 145,25 kb nos cromossomos 8 e 9, respectivamente. Para se estimar o tamanho médio dos blocos de ligação, considerou-se a distância física em que ocorreu o cruzamento da linha de regressão logarítmica na metade do valor máximo ($r^2 \sim 0,23$). Portanto, considerando todos os cromossomos, o tamanho dos blocos haplotípicos para o painel de associação deste estudo atingiu uma distância de 246,12 kb e 65,74 kb, sem e com correção para estrutura e parentesco, respectivamente.

**Tabela 4.** Desequilíbrio de ligação (DL) em cada um dos 11 cromossomos de *P. vulgaris*

| **Cromossomo** | **$r^2$ médio** | **dist. $r^2$*** | **$r^2$sv médio** | **dist. $r^2$sv*** |
|---|---|---|---|---|
| 1 | 0.61 | 306,27 | 0,55 | 89,89 |
| 2 | 0.57 | 349,22 | 0,60 | 104,15 |
| 3 | 0.50 | 413,18 | 0,47 | 127,84 |
| 4 | 0.79 | 188,55 | 0,77 | 48,67 |
| 5 | 0.72 | 163,39 | 0,68 | 59,49 |
| 6 | 0.60 | 275,93 | 0,52 | 100,68 |
| 7 | 0.62 | 264,06 | 0,49 | 92,65 |
| 8 | 0.67 | 195,18 | 0,62 | 32,49 |
| 9 | 0.57 | 367,61 | 0,53 | 145,25 |
| 10 | 0.56 | 246,02 | 0,57 | 54,87 |
| 11 | 0.66 | 172,95 | 0,62 | 42,00 |
| Total | 0.64 | 246,12 | 0,61 | 65,74 |

* Distância física, em kb, na qual o DL é aproximadamente igual a metade do valor máximo ($r^2$ ~ 0,23).

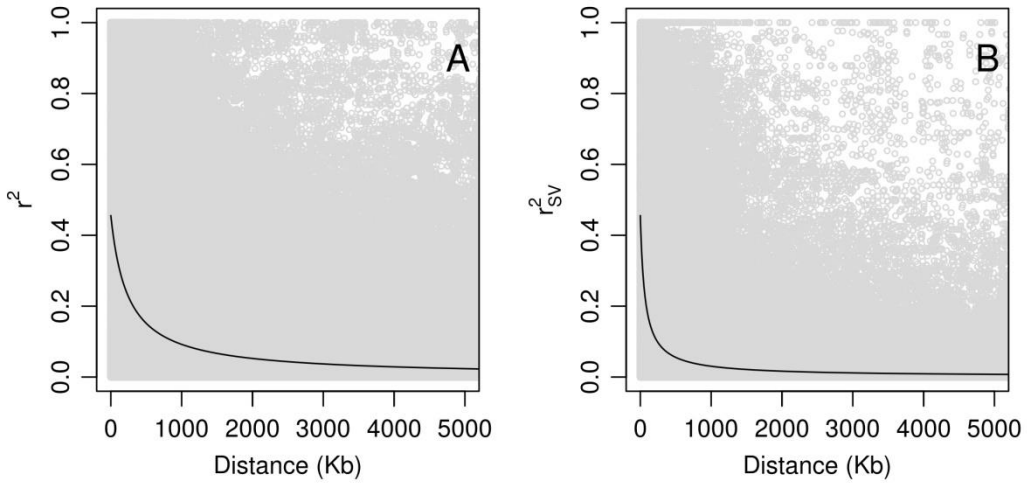



**Figura 3.** Padrão do decaimento do DL em relação à distância entre os marcadores SNPs até 5 Mb, considerando todos os cromossomos de *P. vulgaris*. (A) Curva de decaimento sem correção para estrutura e parentesco; (B) Curva de decaimento com correção para estrutura e parentesco.

**Diversidade e estruturação genética**

As estimativas médias de *Ho* e $H_E$ para o conjunto total de amostras foram de 0,022 (±0,002) e 0,285 (±0,010), respectivamente. A análise de estruturação mostrou que o K = 3 foi o mais provável pela estatística *deltaK* (Evanno et al 2005) (Figura 4), subdividindo os genótipos em três grupos: M1, M2 e M3. O grupo M1 consistiu no agrupamento de 87 acessos, sendo que 60% foram cultivares/linhagens e 64% dos grãos foram do tipo comercial preto. O grupo M2 foi o menor, com apenas nove acessos, sem predomínio por tipo de grão comercial, porém, a maioria (67%) apresentou tamanho de grão grande. O grupo M3 foi composto por 108 acessos, dos quais 95% foram variedades tradicionais e com ampla variação de tipos de grãos comerciais, com predomínio de grãos mulatinho e carioca (39%). Um total de 139 acessos (57%) foi considerado como contendo ancestria múltipla (*admixture*), com valor de q < 0,07. A matriz de parentesco (*kinship*) representada na Figura 5 mostra que a maioria dos genótipos tem distância genética alta entre eles, com exceção para alguns pontos, como por exemplo, no canto inferior esquerdo (vermelho), que tem um grupo com alta similaridade genética entre os acessos. Os genótipos pertencentes a este conjunto são os mesmos que compõem o agrupamento M2 identificado na análise do Structure.

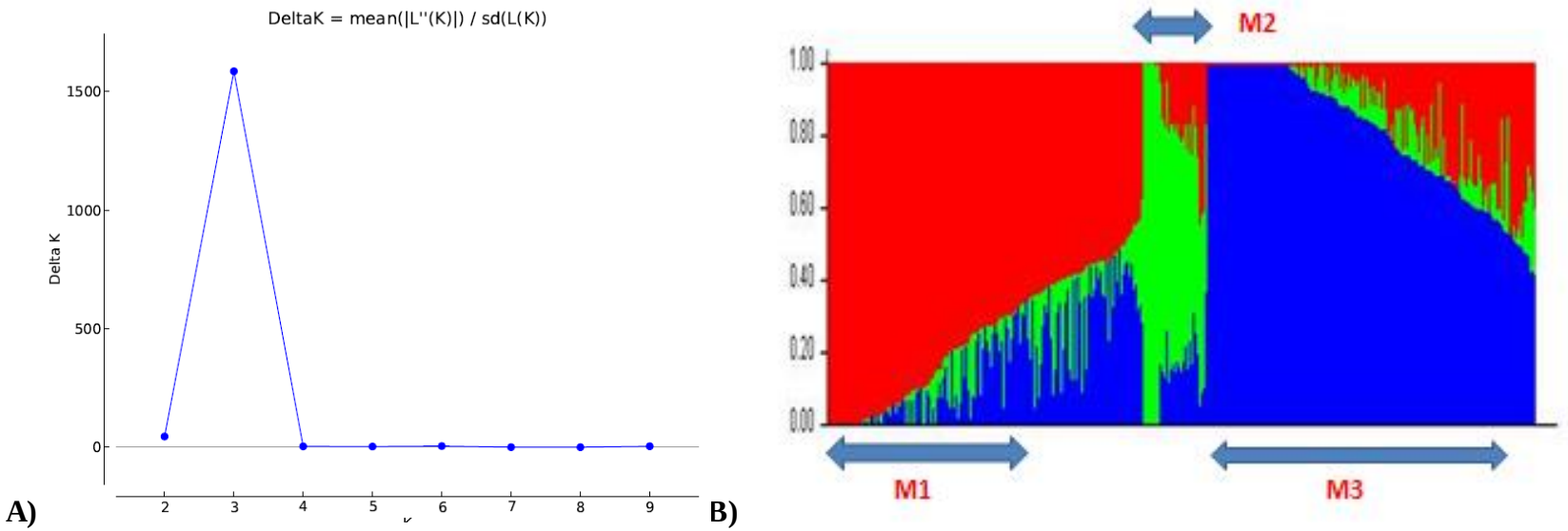



**Figura 4.** Estruturação genética dos 343 acessos Mesoamericanos: A) valores de delta K para análise do Structure; B) Divisão dos genótipos em três grupos. Os grupos incluem: M1 (vermelho): 87 acessos, sendo 64% dos grãos do tipo comercial preto; M2 (verde): nove acessos, sem predomínio por tipo de grão comercial, mas com 67% dos acessos com tamanho de grão grande; M3 (azul): 108 acessos, dos quais 95% foram variedades tradicionais, com predomínio de grãos mulatinho e carioca (39%).

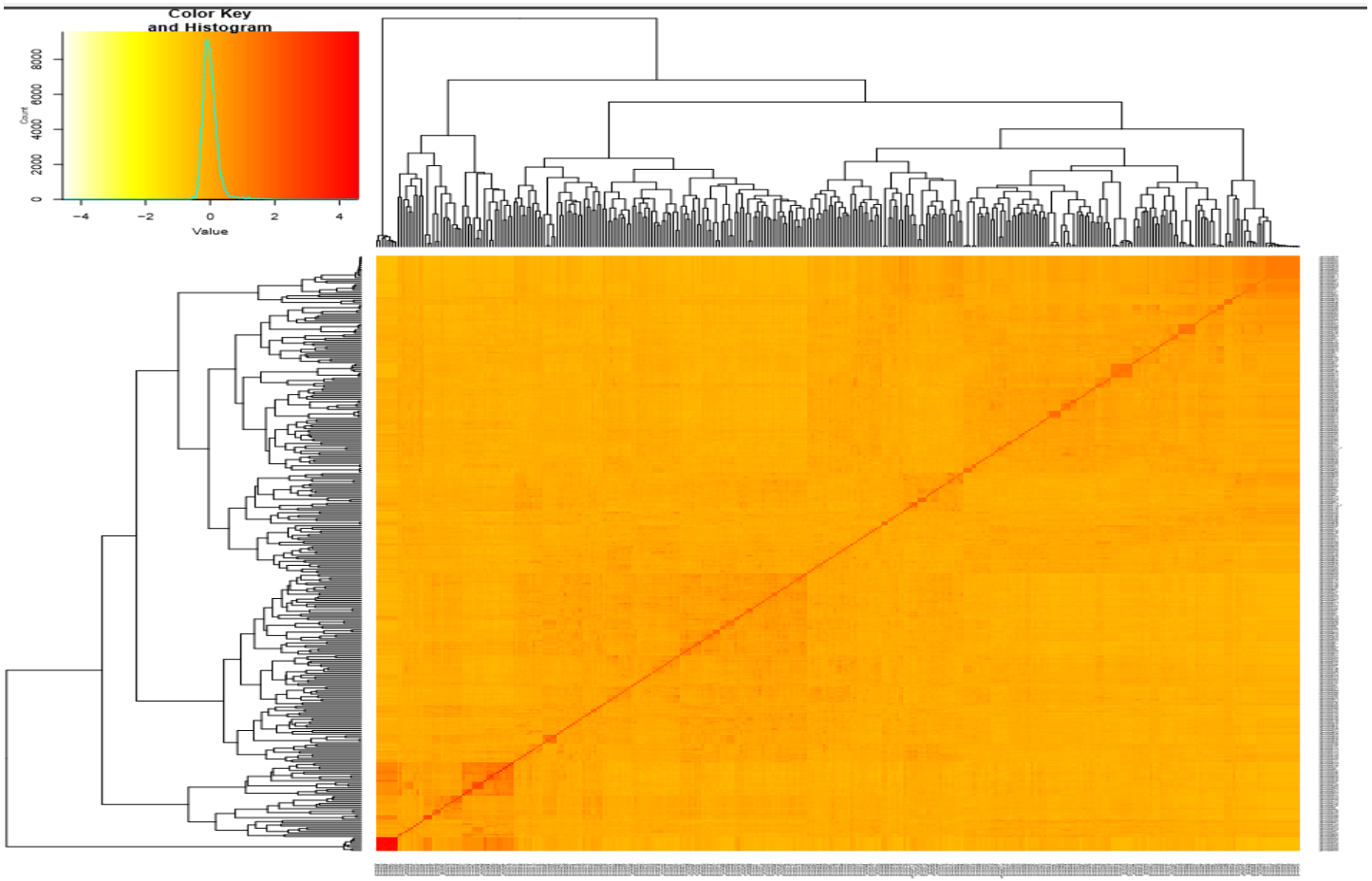



**Figura 5.** *Heatmap* das distâncias genéticas par-a-par entre os genótipos de acordo com o método de VanRaden. A legenda de cor no canto superior esquerdo mostra a coloração de acordo com os valores de distância par-a-par. O dendrograma e o parentesco são mostrados na parte superior e esquerda da figura.

**GWAS**

Nas análises individuais, após a correção dos *p*-valores para múltiplos testes de hipóteses (FDR), do total de 8789 SNPs, 93 SNPs foram significativamente associados a um dos caracteres em, pelo menos, um dos ambientes, sendo 59 e 34 SNPs gerados pelas tecnologias DArTseq e Capture-Seq, respectivamente (Tabela 5). Para o caráter massa de 100 grãos, foram detectados 74 e 6 SNPs significativamente associados nos experimentos SDH-2014 e CDH-2014, respectivamente. Estes marcadores estavam distribuídos em quase todos os cromossomos, com exceção dos cromossomos 9 e 11. Os SNPs detectados no ambiente SDH-2014 (Figura 6) presentes nos cromossomos 2 e 4 estão contidos em blocos haplotípicos, com tamanhos de 157 kb e 30 kb, respectivamente (dados não mostrados). Para o caráter produtividade foram detectados SNPs significativos nos ambientes com deficiência hídrica, sendo 12 SNPs significativos no ambiente CDH-2016 e um no ambiente CDH-2014, presentes nos cromossomos 1, 4, 7, 8 e 11. Não houve a identificação de SNPs significativos relacionados ao ISS em nenhum dos ambientes. Para o caráter massa de 100 grãos, dois SNPs foram detectados em dois ambientes simultaneamente: SDH-2014 e CDH-2014. A variância fenotípica, explicada por cada SNP, variou de 4,46 a 10,25% para massa de 100 grãos e de 8,00 a 14,39% para produtividade (Tabela 6). Os gráficos de Manhattan plot para produtividade nos ambientes CDH-16 e CDH-14 estão representados nas figuras 7A e 8A, respectivamente.

Utilizando as médias preditas pela análise conjunta, ou seja, a que considera a interação genótipo x ambiente, houve detecção de 133 SNPs significativos para massa de 100 grãos nos ambientes irrigados, sendo que 18 SNPs foram comuns aos detectados nas análises individuais, com dois SNPs comuns detectados em ambos os experimentos de 2014 (SDH-2014 e CDH-2014), 14 detectados somente no ambiente SDH-2014 e dois identificados somente no ambiente CDH-2014. Para produtividade foram detectados cinco SNPs significativos em ambiente com deficiência hídrica, sendo que um destes SNPs também foi detectado na análise individual no ambiente CDH-2016. A variância fenotípica, explicada por um único SNP, variou de 3,17 a 7,91% para massa de 100 grãos e de 5,62 a 7,29% para produtividade (Tabela 6).

**Tabela 5.** Total de SNPs significativos detectados nas análises individuais e conjunta

| Experimento | SNP DArTseq | SNP Capture-Seq | Total |
|---|---|---|---|
| Massa_SDH_2014 | 47 | 27 | 74 |
| Massa_CDH_2014 | 3 | 3 | 6 |
| Prod_CDH_2014 | 1 | 0 | 1 |
| Prod_CDH_2016 | 8 | 4 | 12 |
| Massa_SDH_conjunta | 72 | 61 | 133 |
| Prod_CDH_conjunta | 3 | 2 | 5 |

Massa: massa de 100 grãos; Prod: produtividade; SDH: sem deficiência hídrica; CDH: com deficiência hídrica.

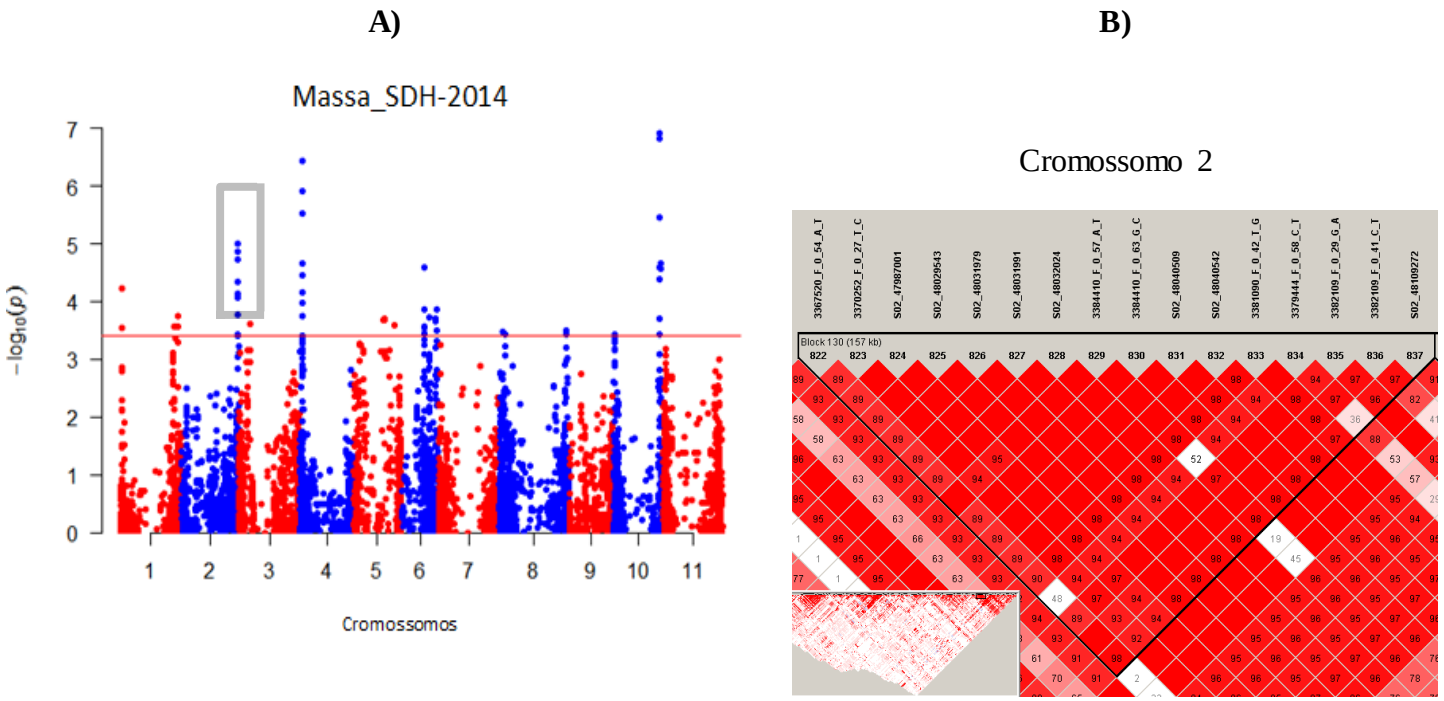



**Figura 6.** GWAS para o caráter massa de 100 grãos no experimento sem deficiência hídrica em 2014 (SDH-2014). (A) Manhattan plot mostrando os SNPs significativos (*p*-valor ≤ 3,95E-04); (B) Bloco haplotípico contendo os SNPs significativos detectados no Cromossomo 2 de *P. vulgaris* (47,98 a 48,10 Mb).

**Tabela 6.** Resultados da Análise de Associação Genômica Ampla (GWAS)

| Caract/Ambiente | Marcador | Crm. | Pos. (Mb) | Metodologia | *p*-valor | Efeito | Alelos | TVFE (%) |
|---|---|---|---|---|---|---|---|---|
| **P_CDH_conj** | S01_1681129* | 1 | 1,68 | Capture-Seq | 6,82E-06 | -84.37 | A/G | 6,63 |
| | 3383831_F_0_8_C_T | 2 | 4,31 | DArTseq | 3,29E-05 | 52.05 | C/T | 5,62 |
| | S02_4341198 | 2 | 4,34 | Capture-Seq | 3,29E-05 | 52.05 | G/T | 5,62 |
| | 3384331_F_0_44_T_A | 6 | 2,26 | DArTseq | 2,73E-06 | -46.61 | A/T | 7,22 |
| | 8196854_F_0_58_A_G | 6 | 2,27 | DArTseq | 3,27E-06 | -46.23 | A/G | 7,29 |
| **P_CDH_2014** | 8212543_F_0_56_C_A | 7 | 48,27 | DArTseq | 1,74E-06 | -115.85 | A/C | 9,44 |
| **P_CDH_2016** | 3372117_F_0_49_T_C | 1 | 1,10 | DArTseq | 1,09E-05 | 280.55 | C/T | 9,81 |
| | S01_1681129* | 1 | 1,68 | Capture-Seq | 1,26E-07 | -362.68 | A/G | 14,39 |
| | 3383663_F_0_34_A_G | 4 | 8,11 | DArTseq | 5,37E-05 | 239.25 | A/G | 8,26 |
| | 3370657_F_0_63_C_T | 4 | 8,11 | DArTseq | 5,37E-05 | -239.25 | C/T | 8,26 |
| | S04_39340571 | 4 | 39,34 | Capture-Seq | 1,30E-05 | -241.43 | C/T | 9,58 |
| | 3366987_F_0_61_T_C | 4 | 39,95 | DArTseq | 5,98E-06 | -217.30 | C/T | 10,47 |
| | 3378663_F_0_27_G_C | 4 | 39,95 | DArTseq | 1,58E-05 | -207.65 | C/G | 9,39 |
| | 8206842_F_0_19_G_A | 8 | 4,56 | DArTseq | 6,32E-05 | 248.29 | A/G | 8,01 |
| | S11_3061868 | 11 | 3,06 | Capture-Seq | 4,64E-05 | -280.01 | C/T | 8,32 |
| | 3371634_F_0_19_A_T | 11 | 3,10 | DArTseq | 4,64E-05 | -280.01 | A/T | 8,32 |
| | S11_3134976 | 11 | 3,13 | Capture-Seq | 4,21E-05 | 285.29 | A/G | 8,41 |
| | 3366930_F_0_66_G_T | 11 | 43,97 | DArTseq | 3,71E-05 | -277.01 | G/T | 8,54 |
| **M_SDH_conj.** | 8196517_F_0_39_C_T | 1 | 16,93 | DArTseq | 1,98E-04 | 3.51 | C/T | 3,92 |
| | S01_16989210 | 1 | 16,99 | Capture-Seq | 1,98E-04 | 3.51 | C/T | 3,92 |
| | 3377089_F_0_23_C_T | 1 | 19,48 | DArTseq | 2,02E-04 | 3.50 | C/T | 3,91 |
| | S01_19539286 | 1 | 19,54 | Capture-Seq | 1,98E-04 | 3.51 | C/T | 3,92 |
| | 3378493_F_0_34_A_C | 1 | 19,98 | DArTseq | 1,98E-04 | 3.51 | A/C | 3,92 |
| | 3365450_F_0_14_A_T | 2 | 3,11 | DArTseq | 3,91E-04 | -2.93 | A/T | 3,58 |
| | 3380754_F_0_38_A_G | 2 | 32,81 | DArTseq | 3,23E-04 | 2.18 | A/G | 3,65 |
| | 3379613_F_0_47_C_G | 2 | 32,83 | DArTseq | 3,61E-04 | 2.18 | C/G | 3,59 |
| | S02_36557526 | 2 | 36,56 | Capture-Seq | 2,72E-04 | -3.11 | A/G | 3,74 |
| | 3377490_F_0_18_G_A | 2 | 41,57 | DArTseq | 3,96E-04 | 2.20 | A/G | 3,54 |
| | 3381922_F_0_52_G_C | 3 | 6,64 | DArTseq | 7,20E-06 | 2.97 | C/G | 5,75 |
| | 3384231_F_0_60_C_G | 3 | 6,67 | DArTseq | 1,58E-05 | -2.81 | C/G | 5,32 |
| | 3378482_F_0_61_T_G | 3 | 6,69 | DArTseq | 3,74E-05 | 2.63 | G/T | 4,83 |
| | 16649566_F_0_5_G_A | 3 | 6,70 | DArTseq | 3,43E-05 | -2.65 | A/G | 4,89 |
| | S03_6773516 | 3 | 6,77 | Capture-Seq | 3,52E-04 | 2.15 | C/G | 3,60 |
| | S03_6773523 | 3 | 6,77 | Capture-Seq | 3,52E-04 | -2.15 | C/T | 3,60 |
| | S03_6773550 | 3 | 6,77 | Capture-Seq | 3,52E-04 | 2.15 | C/G | 3,60 |
| | S03_6773555 | 3 | 6,77 | Capture-Seq | 3,52E-04 | -2.15 | A/G | 3,60 |
| | S03_6773572 | 3 | 6,77 | Capture-Seq | 3,52E-04 | -2.15 | A/G | 3,60 |
| | S03_6773574 | 3 | 6,77 | Capture-Seq | 3,52E-04 | 2.15 | A/C | 3,60 |
| | 3368469_F_0_68_C_T | 3 | 6,91 | DArTseq | 4,43E-05 | -2.59 | C/T | 4,74 |
| | 3381777_F_0_15_C_T | 3 | 6,96 | DArTseq | 4,44E-05 | -2.59 | C/T | 4,73 |
| | 3378068_F_0_5_C_G | 3 | 6,96 | DArTseq | 4,09E-04 | -2.11 | C/G | 3,52 |
| | 3378952_F_0_62_C_A | 3 | 7,05 | DArTseq | 4,44E-05 | 2.59 | A/C | 4,73 |
| | S03_7065160 | 3 | 7,07 | Capture-Seq | 4,44E-05 | 2.59 | A/C | 4,73 |

| Caract/Ambiente | Marcador | Crm. | Pos. (Mb) | Metodologia | *p*-valor | Efeito | Alelos | TVFE (%) |
|---|---|---|---|---|---|---|---|---|
| | S03_7065209 | 3 | 7,07 | Capture-Seq | 4,44E-05 | -2.59 | C/T | 4,73 |
| | S03_7065258 | 3 | 7,07 | Capture-Seq | 4,44E-05 | 2.59 | A/C | 4,73 |
| | 3384344_F_0_48_A_G | 3 | 7,14 | DArTseq | 7,32E-05 | 2.54 | A/G | 4,63 |
| | S03_7455661 | 3 | 7,46 | Capture-Seq | 2,77E-04 | 2.14 | A/G | 3,73 |
| | S03_7455669 | 3 | 7,46 | Capture-Seq | 3,48E-05 | -2.56 | A/T | 4,87 |
| | S03_7676384 | 3 | 7,68 | Capture-Seq | 2,77E-04 | -2.14 | C/T | 3,73 |
| | S03_7676396 | 3 | 7,68 | Capture-Seq | 2,77E-04 | -2.14 | C/T | 3,73 |
| | S03_7676472 | 3 | 7,68 | Capture-Seq | 2,77E-04 | -2.14 | A/G | 3,73 |
| | S03_7676473 | 3 | 7,68 | Capture-Seq | 2,77E-04 | -2.14 | A/C | 3,73 |
| | 3377699_F_0_28_C_T | 3 | 7,75 | DArTseq | 7,27E-04 | -2.00 | C/T | 3,29 |
| | 8208053_F_0_34_C_T | 3 | 7,76 | DArTseq | 3,95E-05 | 2.55 | C/T | 4,80 |
| | 3384124_F_0_19_C_T | 3 | 7,82 | DArTseq | 1,80E-04 | -2.22 | C/T | 3,98 |
| | 3369267_F_0_57_G_A | 3 | 7,90 | DArTseq | 4,55E-04 | 2.70 | A/G | 3,47 |
| | 3381467_F_0_42_A_G | 3 | 7,90 | DArTseq | 3,47E-05 | -3.37 | A/G | 4,87 |
| | 3370408_F_0_67_C_A | 3 | 8,03 | DArTseq | 4,93E-06 | -3.68 | A/C | 5,96 |
| | S03_9172353 | 3 | 9,17 | Capture-Seq | 7,03E-04 | -1.98 | A/T | 3,23 |
| | 8668166_F_0_42_A_G | 3 | 9,94 | DArTseq | 3,12E-04 | -2.83 | A/G | 3,70 |
| | S03_44903283 | 3 | 44,90 | Capture-Seq | 3,73E-05 | -4.31 | A/T | 4,83 |
| | S03_44903298 | 3 | 44,90 | Capture-Seq | 3,73E-05 | 4.31 | A/G | 4,83 |
| | 3370726_F_0_18_C_T | 3 | 44,91 | DArTseq | 3,73E-05 | 4.31 | C/T | 4,83 |
| | 3366011_F_0_17_C_T | 3 | 44,95 | DArTseq | 3,15E-07 | -5.71 | C/T | 7,91 |
| | 3383946_F_0_26_G_A | 3 | 44,96 | DArTseq | 6,90E-05 | 4.11 | A/G | 4,59 |
| **M_SDH_conj** | 3377445_F_0_49_A_G | 3 | 45,41 | DArTseq | 4,56E-07 | -5.84 | A/G | 7,33 |
| | 3384388_F_0_54_T_C | 3 | 45,62 | DArTseq | 4,44E-06 | 5.47 | C/T | 6,26 |
| | S03_45676416 | 3 | 45,68 | Capture-Seq | 5,71E-05 | -4.38 | C/G | 4,60 |
| | 8207877_F_0_16_G_A | 3 | 46,78 | DArTseq | 4,92E-05 | -3.73 | A/G | 4,68 |
| | 3379142_F_0_15_G_T | 3 | 46,81 | DArTseq | 4,92E-05 | 3.73 | G/T | 4,68 |
| | 3381778_F_0_44_C_A | 3 | 46,82 | DArTseq | 5,53E-05 | 3.80 | A/C | 4,61 |
| | 3383288_F_0_47_C_T | 3 | 46,82 | DArTseq | 5,25E-05 | -3.85 | C/T | 4,64 |
| | S04_2116015* | 4 | 2,12 | Capture-Seq | 1,50E-04 | -2.39 | A/C | 4,07 |
| | S04_2116041* | 4 | 2,12 | Capture-Seq | 1,50E-04 | 2.39 | A/G | 4,07 |
| | 16649093_F_0_15_G_C | 4 | 43,33 | DArTseq | 5,97E-05 | -1.93 | C/G | 4,59 |
| | 3380531_F_0_38_C_T | 4 | 44,45 | DArTseq | 5,37E-04 | 2.61 | C/T | 3,38 |
| | 3366948_F_0_39_T_C* | 6 | 19,26 | DArTseq | 6,23E-04 | 2.94 | C/T | 3,30 |
| | 8198088_F_0_66_C_G | 6 | 19,38 | DArTseq | 5,24E-04 | -3.10 | C/G | 3,39 |
| | S06_19412197 | 6 | 19,41 | Capture-Seq | 5,24E-04 | -3.10 | A/G | 3,39 |
| | S06_19412202 | 6 | 19,41 | Capture-Seq | 5,24E-04 | -3.10 | C/T | 3,39 |
| | S06_19412224 | 6 | 19,41 | Capture-Seq | 5,24E-04 | 3.10 | C/G | 3,39 |
| | S06_19412227 | 6 | 19,41 | Capture-Seq | 5,24E-04 | 3.10 | C/T | 3,39 |
| | S06_19412240 | 6 | 19,41 | Capture-Seq | 5,24E-04 | 3.10 | C/T | 3,39 |
| | 3380824_F_0_8_T_A* | 6 | 19,60 | DArTseq | 3,81E-04 | 1.56 | A/T | 3,57 |
| | 3366197_F_0_7_G_C* | 6 | 19,73 | DArTseq | 2,86E-04 | 1.66 | C/G | 3,71 |
| | 3377397_F_0_11_A_T* | 6 | 19,73 | DArTseq | 2,86E-04 | -1.66 | A/T | 3,73 |
| | S06_27506464 | 6 | 27,51 | Capture-Seq | 1,14E-04 | -3.75 | A/T | 4,22 |
| | S06_27506487 | 6 | 27,51 | Capture-Seq | 1,14E-04 | -3.75 | A/G | 4,22 |

| Caract/Ambiente | Marcador | Crm. | Pos. (Mb) | Metodologia | *p*-valor | Efeito | Alelos | TVFE (%) |
|---|---|---|---|---|---|---|---|---|
| | S06_27506498 | 6 | 27,51 | Capture-Seq | 1,14E-04 | -3.75 | C/T | 4,22 |
| | S07_506504 | 7 | 0,51 | Capture-Seq | 2,43E-04 | 2.32 | A/G | 3,80 |
| | S07_506512 | 7 | 0,51 | Capture-Seq | 2,43E-04 | 2.32 | A/T | 3,80 |
| | S07_506516 | 7 | 0,51 | Capture-Seq | 2,43E-04 | -2.32 | A/T | 3,80 |
| | S07_506518 | 7 | 0,51 | Capture-Seq | 2,43E-04 | 2.32 | G/T | 3,80 |
| | S07_506526 | 7 | 0,51 | Capture-Seq | 2,43E-04 | -2.32 | A/G | 3,80 |
| | 8214713_F_0_5_C_A | 7 | 0,51 | DArTseq | 2,43E-04 | -2.32 | A/C | 3,80 |
| | S07_1121648* | 7 | 1,12 | Capture-Seq | 2,59E-06 | 3.44 | C/G | 6,33 |
| | S07_1121680* | 7 | 1,12 | Capture-Seq | 2,59E-06 | 3.44 | C/T | 6,33 |
| | S07_1139917 | 7 | 1,14 | Capture-Seq | 1,37E-06 | -3.34 | C/T | 6,69 |
| | 8195516_F_0_29_C_T | 7 | 1,16 | DArTseq | 1,37E-06 | 3.34 | C/T | 6,69 |
| | 3382397_F_0_28_A_C | 7 | 1,18 | DArTseq | 2,87E-05 | 2.99 | A/C | 4,98 |
| | 3380111_F_0_61_A_C | 7 | 1,21 | DArTseq | 2,36E-05 | 3.04 | A/C | 5,09 |
| | 3378610_F_0_36_G_A | 7 | 1,21 | DArTseq | 2,87E-05 | -2.99 | A/G | 4,98 |
| | 3380418_F_0_33_C_T | 8 | 1,05 | DArTseq | 4,02E-04 | 3.13 | C/T | 3,58 |
| | 8216048_F_0_53_G_T | 8 | 1,08 | DArTseq | 7,16E-04 | -3.55 | G/T | 3,87 |
| | S08_1138073 | 8 | 1,14 | Capture-Seq | 5,55E-04 | 2.24 | A/T | 3,36 |
| | S08_1138136 | 8 | 1,14 | Capture-Seq | 5,55E-04 | 2.24 | A/G | 3,36 |
| | 3381234_F_0_53_G_A | 8 | 1,23 | DArTseq | 2,05E-04 | 1.84 | A/G | 3,91 |
| | S08_1400776 | 8 | 1,40 | Capture-Seq | 2,23E-04 | -2.95 | G/T | 3,85 |
| | S08_1478954 | 8 | 1,48 | Capture-Seq | 2,41E-04 | 2.90 | A/G | 3,83 |
| | S08_1478991 | 8 | 1,48 | Capture-Seq | 2,46E-04 | -2.90 | C/T | 3,80 |
| **M_SDH_conj** | 3382122_F_0_48_T_C | 8 | 1,57 | DArTseq | 3,77E-04 | -2.72 | C/T | 3,59 |
| | S08_1744472 | 8 | 1,74 | Capture-Seq | 9,01E-05 | 2.97 | G/T | 4,34 |
| | 3381158_F_0_52_G_A | 8 | 1,88 | DArTseq | 5,06E-04 | -2.72 | A/G | 3,41 |
| | S08_1886060 | 8 | 1,89 | Capture-Seq | 5,42E-04 | 2.70 | A/G | 3,37 |
| | 3381059_F_0_50_G_A | 8 | 2,57 | DArTseq | 6,89E-04 | 1.74 | A/G | 3,25 |
| | 3381821_F_0_62_C_T | 8 | 4,35 | DArTseq | 5,40E-04 | 3.36 | C/T | 3,38 |
| | 3384309_F_0_37_G_A | 8 | 4,37 | DArTseq | 2,06E-04 | -3.53 | A/G | 3,89 |
| | 3368476_F_0_44_A_C | 8 | 4,41 | DArTseq | 2,05E-04 | -3.53 | A/C | 3,90 |
| | 3381402_F_0_44_A_T | 8 | 4,42 | DArTseq | 1,24E-06 | 4.05 | A/T | 6,75 |
| | 3381889_F_0_17_T_G | 8 | 48,57 | DArTseq | 2,28E-04 | 2.86 | G/T | 3,97 |
| | S08_48616865 | 8 | 48,62 | Capture-Seq | 6,55E-04 | -2.61 | A/G | 3,27 |
| | S09_8696600 | 9 | 8,70 | Capture-Seq | 6,91E-04 | -3.74 | C/T | 3,24 |
| | S09_8764673 | 9 | 8,76 | Capture-Seq | 6,90E-04 | 3.74 | A/G | 3,24 |
| | S09_8895621 | 9 | 8,90 | Capture-Seq | 6,90E-04 | 3.74 | G/T | 3,24 |
| | 3379011_F_0_24_C_T | 9 | 35,21 | DArTseq | 4,42E-04 | -2.34 | C/T | 3,50 |
| | 3377566_F_0_61_A_T | 9 | 35,26 | DArTseq | 6,53E-05 | -2.64 | A/T | 4,55 |
| | 3381021_F_0_34_C_A* | 10 | 39,95 | DArTseq | 2,08E-06 | 3.59 | A/C | 6,45 |
| | 3379372_F_0_50_G_T* | 10 | 39,96 | DArTseq | 6,75E-05 | -3.04 | G/T | 4,50 |
| | S10_39972372* | 10 | 39,97 | Capture-Seq | 3,91E-04 | 2.27 | C/T | 3,55 |
| | S10_39972399* | 10 | 39,97 | Capture-Seq | 4,99E-05 | -2.72 | C/T | 4,67 |
| | 3382850_F_0_52_A_G* | 10 | 40,09 | DArTseq | 4,82E-05 | -3.19 | A/G | 4,70 |
| | 3377812_F_0_9_T_A* | 10 | 40,17 | DArTseq | 3,61E-05 | 3.11 | A/T | 4,87 |
| | 3377213_F_0_23_T_C* | 10 | 40,32 | DArTseq | 4,34E-04 | 2.77 | C/T | 3,50 |

| Caract/Ambiente | Marcador | Crm. | Pos. (Mb) | Metodologia | *p*-valor | Efeito | Alelos | TVFE (%) |
|---|---|---|---|---|---|---|---|---|
| M_SDH_conj | S10_40547057* | 10 | 40,55 | Capture-Seq | 5,11E-05 | 3.32 | A/T | 4,66 |
| | 3379386_F_0_66_T_A | 10 | 41,03 | DArTseq | 1,79E-04 | 3.05 | A/T | 3,97 |
| | S10_41183075 | 10 | 41,18 | Capture-Seq | 5,19E-04 | 1.49 | A/G | 3,40 |
| | S10_41183110 | 10 | 41,18 | Capture-Seq | 5,13E-04 | -1.50 | C/T | 3,40 |
| | S10_41183114 | 10 | 41,18 | Capture-Seq | 5,19E-04 | -1.49 | A/G | 3,39 |
| | S10_41183115 | 10 | 41,18 | Capture-Seq | 5,19E-04 | -1.49 | A/G | 3,39 |
| | S10_41183129 | 10 | 41,18 | Capture-Seq | 5,19E-04 | 1.49 | A/G | 3,40 |
| | S10_41183134 | 10 | 41,18 | Capture-Seq | 5,13E-04 | -1.50 | C/G | 3,40 |
| | S10_41183147 | 10 | 41,18 | Capture-Seq | 5,13E-04 | -1.50 | C/T | 3,40 |
| | 8212518_F_0_5_G_C | 11 | 1,44 | DArTseq | 9,50E-05 | -2.83 | C/G | 4,32 |
| | 8207515_F_0_37_T_G | 11 | 1,50 | DArTseq | 1,74E-04 | 2.10 | G/T | 3,98 |
| | 3365978_F_0_16_G_C | 11 | 1,51 | DArTseq | 1,11E-04 | -2.16 | C/G | 4,23 |
| | 3381765_F_0_66_T_C | 11 | 1,53 | DArTseq | 5,68E-05 | 3.37 | C/T | 4,62 |
| | 3377359_F_0_50_G_A | 11 | 1,56 | DArTseq | 3,19E-04 | 2.93 | A/G | 3,66 |
| | 3378511_F_0_64_C_A | 11 | 1,66 | DArTseq | 7,85E-04 | 2.55 | A/C | 3,17 |
| | 3382969_F_0_65_A_C | 11 | 1,67 | DArTseq | 7,85E-04 | -2.55 | A/C | 3,17 |
| | 3368731_F_0_17_A_G | 11 | 1,70 | DArTseq | 6,41E-04 | -2.49 | A/G | 3,28 |
| | 3380677_F_0_38_A_T | 11 | 46,66 | DArTseq | 6,34E-04 | -1.61 | A/T | 3,30 |
| M_SDH_2014 | 3381496_F_0_66_A_G | 1 | 1,10 | DArTseq | 5,91E-05 | -3.85 | A/G | 5,77 |
| | 3383779_F_0_47_T_A | 1 | 1,14 | DArTseq | 2,80E-04 | 3.06 | A/T | 4,70 |
| | 8206455_F_0_42_G_A | 1 | 45,48 | DArTseq | 2,72E-04 | -3.77 | A/G | 4,79 |
| | 3368286_F_0_32_A_T | 1 | 48,97 | DArTseq | 1,81E-04 | 2.46 | A/T | 5,00 |
| | 8208534_F_0_58_C_T | 1 | 48,98 | DArTseq | 2,67E-04 | -2.44 | C/T | 4,79 |
| | 3367520_F_0_5_T_C | 2 | 47,95 | DArTseq | 3,92E-04 | -3.49 | C/T | 4,47 |
| | 3370252_F_0_27_T_C | 2 | 47,98 | DArTseq | 7,55E-05 | 4.00 | C/T | 5,73 |
| | S02_47987001 | 2 | 47,99 | Capture-Seq | 7,50E-05 | 4.00 | C/T | 5,60 |
| | S02_48029543 | 2 | 48,03 | Capture-Seq | 7,77E-05 | -3.99 | G/T | 5,58 |
| | S02_48031979 | 2 | 48,03 | Capture-Seq | 7,77E-05 | -3.99 | G/T | 5,58 |
| | S02_48031991 | 2 | 48,03 | Capture-Seq | 7,77E-05 | -3.99 | A/T | 5,58 |
| | S02_48032024 | 2 | 48,03 | Capture-Seq | 7,77E-05 | -3.99 | C/T | 5,58 |
| | 3384410_F_0_57_A_T | 2 | 48,04 | DArTseq | 1,71E-04 | 3.56 | A/T | 5,03 |
| | 3384410_F_0_63_G_C | 2 | 48,04 | DArTseq | 7,77E-05 | -3.99 | C/G | 5,58 |
| | S02_48040509 | 2 | 48,04 | Capture-Seq | 7,77E-05 | 3.99 | C/T | 5,58 |
| | S02_48040542 | 2 | 48,04 | Capture-Seq | 7,77E-05 | 3.99 | A/T | 5,58 |
| | 3381090_F_0_42_T_G | 2 | 48,06 | DArTseq | 1,88E-05 | -4.36 | G/T | 6,59 |
| | 3382109_F_0_29_G_A | 2 | 48,10 | DArTseq | 8,36E-05 | -3.98 | A/G | 5,54 |
| | 3379952_F_0_34_A_G | 2 | 48,22 | DArTseq | 1,38E-05 | 4.89 | A/G | 7,61 |
| | 3375189_F_0_26_A_G | 2 | 48,29 | DArTseq | 4,53E-05 | -4.43 | A/G | 5,96 |
| | 3378476_F_0_13_G_A | 2 | 48,31 | DArTseq | 9,80E-06 | -4.94 | A/G | 7,17 |
| | 3381963_F_0_12_T_C | 2 | 48,31 | DArTseq | 3,75E-04 | -3.74 | C/T | 4,52 |
| | S02_48376751 | 2 | 48,38 | Capture-Seq | 4,53E-05 | 4.43 | G/T | 5,96 |
| | 3377152_F_0_36_G_A | 3 | 9,85 | DArTseq | 2,45E-04 | 3.24 | A/G | 4,82 |
| | S04_1964918 | 4 | 1,96 | Capture-Seq | 3,95E-04 | 4.55 | A/T | 4,46 |
| | 3377630_F_0_63_T_A | 4 | 2,03 | DArTseq | 1,83E-04 | 4.23 | A/T | 5,00 |

| Caract/Ambiente | Marcador | Crm. | Pos. (Mb) | Metodologia | *p*-valor | Efeito | Alelos | TVFE (%) |
|---|---|---|---|---|---|---|---|---|
| | 3372264_F_0_14_T_C | 4 | 2,07 | DArTseq | 3,71E-07 | 3.59 | C/T | 9,58 |
| | S04_2116015 | 4 | 2,12 | Capture-Seq | 3,50E-05 | -3.38 | A/C | 6,14 |
| | S04_2116041 | 4 | 2,12 | Capture-Seq | 3,50E-05 | 3.38 | A/G | 6,14 |
| | S04_2119449 | 4 | 2,12 | Capture-Seq | 6,96E-05 | -2.88 | C/T | 5,66 |
| | S04_2290298 | 4 | 2,29 | Capture-Seq | 2,20E-05 | 2.80 | G/T | 6,47 |
| | S04_2297372 | 4 | 2,30 | Capture-Seq | 3,06E-06 | 3.24 | A/G | 7,87 |
| | S04_2297375 | 4 | 2,30 | Capture-Seq | 3,06E-06 | 3.24 | C/T | 7,87 |
| | S04_2297390 | 4 | 2,30 | Capture-Seq | 3,06E-06 | 3.24 | A/G | 7,87 |
| | S04_2297405 | 4 | 2,30 | Capture-Seq | 3,06E-06 | 3.24 | G/T | 7,87 |
| | S04_2297457 | 4 | 2,30 | Capture-Seq | 3,06E-06 | 3.24 | A/G | 7,87 |
| | 3369473_F_0_19_G_A | 4 | 2,30 | DArTseq | 1,21E-06 | 3.35 | A/G | 8,55 |
| | 8668701_F_0_48_T_C | 4 | 2,38 | DArTseq | 1,03E-04 | 2.92 | C/T | 5,49 |
| | 3383197_F_0_7_G_T | 5 | 25,16 | DArTseq | 2,11E-04 | -3.43 | G/T | 4,89 |
| | 3370982_F_0_64_A_G | 5 | 25,93 | DArTseq | 1,94E-04 | 3.46 | A/G | 5,03 |
| | S05_26033353 | 5 | 26,03 | Capture-Seq | 2,11E-04 | -3.43 | A/G | 4,89 |
| | 3381117_F_0_20_G_A | 5 | 35,18 | DArTseq | 2,53E-04 | -5.18 | A/G | 4,77 |
| | 3378505_F_0_19_C_T | 6 | 19,24 | DArTseq | 3,57E-04 | -4.59 | C/T | 4,55 |
| | 3366948_F_0_39_T_C | 6 | 19,26 | DArTseq | 2,77E-04 | 4.66 | C/T | 4,70 |
| | 3377542_F_0_21_T_A | 6 | 19,56 | DArTseq | 2,86E-04 | 2.37 | A/T | 4,77 |
| | 3380824_F_0_8_T_A | 6 | 19,60 | DArTseq | 2,59E-05 | 2.79 | A/T | 6,37 |
| | S06_19606501 | 6 | 19,61 | Capture-Seq | 1,38E-04 | -2.55 | A/C | 5,18 |
| | S06_19609666 | 6 | 19,61 | Capture-Seq | 1,38E-04 | 2.55 | C/T | 5,18 |
| **M_SDH_2014** | 3366197_F_0_7_G_C | 6 | 19,73 | DArTseq | 3,70E-04 | 2.47 | C/G | 4,51 |
| | 3377397_F_0_11_A_T | 6 | 19,73 | DArTseq | 3,67E-04 | -2.48 | A/T | 4,51 |
| | 3381348_F_0_51_T_C | 6 | 23,76 | DArTseq | 1,87E-04 | 2.98 | C/T | 4,98 |
| | 3372209_F_0_35_G_A | 6 | 28,87 | DArTseq | 2,18E-04 | 4.50 | A/G | 4,87 |
| | 3376862_F_0_7_T_A | 6 | 29,62 | DArTseq | 1,97E-04 | -4.13 | A/T | 4,94 |
| | S06_30279544 | 6 | 30,28 | Capture-Seq | 3,11E-04 | -2.90 | A/C | 4,62 |
| | 3369518_F_0_54_C_T | 6 | 30,41 | DArTseq | 1,35E-04 | -3.13 | C/T | 5,20 |
| | 3378457_F_0_47_G_A | 6 | 30,43 | DArTseq | 1,38E-04 | 3.23 | A/G | 5,18 |
| | 3381811_F_0_43_A_T | 8 | 3,16 | DArTseq | 3,30E-04 | 4.46 | A/T | 4,66 |
| | 3377179_F_0_35_G_A | 8 | 5,17 | DArTseq | 3,63E-04 | 3.12 | A/G | 4,54 |
| | S08_57021618 | 8 | 57,02 | Capture-Seq | 3,15E-04 | 5.70 | G/T | 4,62 |
| | 3380693_F_0_23_A_G | 8 | 57,02 | DArTseq | 3,13E-04 | -5.71 | A/G | 4,63 |
| | S08_57026206 | 8 | 57,03 | Capture-Seq | 3,15E-04 | -5.70 | C/G | 4,62 |
| | 3382224_F_0_48_G_T | 8 | 57,31 | DArTseq | 3,45E-04 | -5.93 | G/T | 4,55 |
| | 3377156_F_0_20_T_G | 8 | 57,40 | DArTseq | 3,77E-04 | 5.88 | G/T | 4,49 |
| | 3381999_F_0_19_C_T | 8 | 57,41 | DArTseq | 3,45E-04 | -5.93 | C/T | 4,55 |
| | 8207393_F_0_59_C_T | 10 | 1,33 | DArTseq | 3,64E-04 | -4.51 | C/T | 5,53 |
| | 3381021_F_0_34_C_A* | 10 | 39,95 | DArTseq | 1,51E-07 | 5.97 | A/C | 10,10 |
| | 3379372_F_0_50_G_T | 10 | 39,96 | DArTseq | 1,22E-07 | -6.20 | G/T | 10,25 |
| | S10_39972372 | 10 | 39,97 | Capture-Seq | 4,07E-05 | 3.98 | C/T | 6,04 |
| | S10_39972399 | 10 | 39,97 | Capture-Seq | 2,01E-04 | -3.82 | C/T | 4,93 |
| | 3381321_F_0_53_G_A | 10 | 40,00 | DArTseq | 3,75E-04 | 3.40 | A/G | 4,50 |
| | 3382850_F_0_52_A_G | 10 | 40,09 | DArTseq | 2,50E-05 | -5.10 | A/G | 6,40 |

| Caract/Ambiente | Marcador | Crm. | Pos. (Mb) | Metodologia | *p*-valor | Efeito | Alelos | TVFE (%) |
|---|---|---|---|---|---|---|---|---|
| **M_SDH_2014** | 3377812_F_0_9_T_A* | 10 | 40,17 | DArTseq | 3,44E-06 | 5.18 | A/T | 7,82 |
| | 3377213_F_0_23_T_C | 10 | 40,32 | DArTseq | 2,77E-05 | 4.94 | C/T | 6,31 |
| | S10_40547057 | 10 | 40,55 | Capture-Seq | 2,20E-05 | 5.21 | A/T | 6,48 |
| **M_CDH_2014** | 3378581_F_0_36_C_A | 6 | 29,69 | DArTseq | 1,29E-05 | 3.86 | A/C | 7,51 |
| | S06_29713507 | 6 | 29,71 | Capture-Seq | 3,50E-06 | -4.29 | C/T | 8,57 |
| | S07_1121648 | 7 | 1,12 | Capture-Seq | 7,29E-07 | 4.21 | C/G | 9,79 |
| | S07_1121680 | 7 | 1,12 | Capture-Seq | 7,29E-07 | 4.21 | C/T | 9,79 |
| | 3381021_F_0_34_C_A | 10 | 39,95 | DArTseq | 1,66E-05 | 3.63 | A/C | 7,33 |
| | 3377812_F_0_9_T_A* | 10 | 40,17 | DArTseq | 1,87E-05 | 3.45 | A/T | 7,23 |

* SNPs identificados em mais de um ambiente

Caract: característica; Crm: cromossomo; Pos.: posição; TVFE: total da variância fenotípica explicada pelo marcador; P: produtividade; M: massa de 100 grãos; SDH: sem deficiência hídrica; CDH: com deficiência hídrica; conj: análise conjunta

**Anotação gênica**

Considerando as análises individuais e conjuntas, 231 SNPs significativos foram identificados para massa de 100 grãos e produtividade. Desses, 17 SNPs foram detectados em mais de um ambiente e 195 foram identificados apenas em um ambiente, totalizando 212 SNPs para a anotação funcional. Dos 212 SNPs, 180 estão dentro de genes e 32 estão próximos a genes, com distâncias variando de 0,15 a 21,29 kb. Um total de 158 transcritos codificados por 156 genes foram anotados, dos quais, a grande maioria codifica para proteínas putativas que estão presentes em componentes integrais de membrana, sistema endomembranoso e membrana plasmática. Quanto à função celular, a maioria tem atividades de hidrolase, oxirredutase e ligação a DNA, participando, principalmente, dos processos de transporte de transmembrana, óxido-redução e organização de organela.

A anotação gênica permitiu a identificação de marcadores SNPs dentro ou próximos de importantes genes com efeitos putativos associados às respostas à deficiência hídrica, para os quais, os acessos que apresentaram determinado alelo foram, significativamente, mais produtivos. Por exemplo, um dos SNPs identificados para produtividade no ambiente CDH-2016 (3366987_F_0_61_T_C) está dentro do gene Phvul.004G124700. Este gene codifica uma proteína putativa "*cytochrome P450 704C1-like*", a qual tem papel crucial no metabolismo do hormônio ácido abscísico (ABA), o qual, por sua vez, está envolvido em vários processos críticos no crescimento e no desenvolvimento normal da planta, bem como em respostas adaptativas a estresses ambientais (Saito et al. 2004). Os acessos com alelo T foram, significativamente (*p*-valor = 4,603E-04), mais produtivos do que os que tinham o alelo C (Figura 7B). No mesmo ambiente, também foram detectados os SNPs 3371634_F_0_19_A_T e S11_3061868, ambos no cromossomo 11. O primeiro está localizado a 150 pb do gene Phvul.011G035200, que está associado com a proteína putativa "*E3 ubiquitin-ligase UPL3*". Esta proteína está envolvida nos processos de ubiquitinação, que consiste na degradação eficiente de proteínas anormais resultantes, dentre outros fatores, de estresses ou danos oxidativos e atua como regulador positivo nas respostas à seca ABA-dependente (Ryu et al. 2010, Lee e Kim 2011). Os acessos que possuem o alelo T foram, em média, significativamente (p = 1,015E-02) mais produtivos do que os indivíduos que possuem o alelo A (Figura 7C). Já o segundo SNP, o qual está localizado dentro do gene Phvul.011G034900, codifica para a proteína putativa "myb-related 308". Os acessos que possuem o alelo T foram, em média, significativamente (p = 1,01E-02), mais produtivos do que os indivíduos que possuem o alelo C (Figura 7D). Esta proteína tem papel essencial no

crescimento, desenvolvimento e resposta a estresses ambientais regulando a síntese de ABA (Xiong et al. 2014). Estes dois últimos SNPs estão a uma distância de 40 kpb e apresentam o mesmo padrão de variação de produtividade representado nos boxplots, sugerindo que eles estejam em desequilíbrio de ligação e que ambos podem estar contribuindo para o aumento da tolerância à seca.

O "SNP 8212543_F_0_56_C_A", o único significativamente relacionado à produtividade no ambiente CDH-2014, está localizado dentro do gene Phvul.007G243200, que codifica o transcrito Phvul.007G243200.1, o qual está associado com a proteína putativa "*ABC transporter G family member 24-like*". Os acessos que possuem o alelo C foram, em média, significativamente (p = 3,95E-05) mais produtivos do que os indivíduos que possuem o alelo A (Figura 8). Esta proteína atua no transporte de ABA e está envolvida na via de sinalização de ABA intercelular (Kuromori e Shinozaki 2010).

Para massa de 100 grãos, o SNP "3377812_F_0_9_T_A" foi detectado nos ambientes SDH-14 e CDH-14 e na análise conjunta dos experimentos irrigados. Este SNP possui dois efeitos preditos no íntron e dois efeitos no sítio aceptor de splicing do gene Phvul.010G131400, o qual codifica para o fator de transcrição "*MYB*" relacionado ao subgrupo "*TRF-like6*". Os genes TRF-like são expressos em maior abundância nas sementes (Du et al. 2013). Os indivíduos que possuem o alelo A apresentaram, significativamente, maior peso para 100 grãos do que os que possuem o alelo T (p = 1,911E-06) (Figura 9).

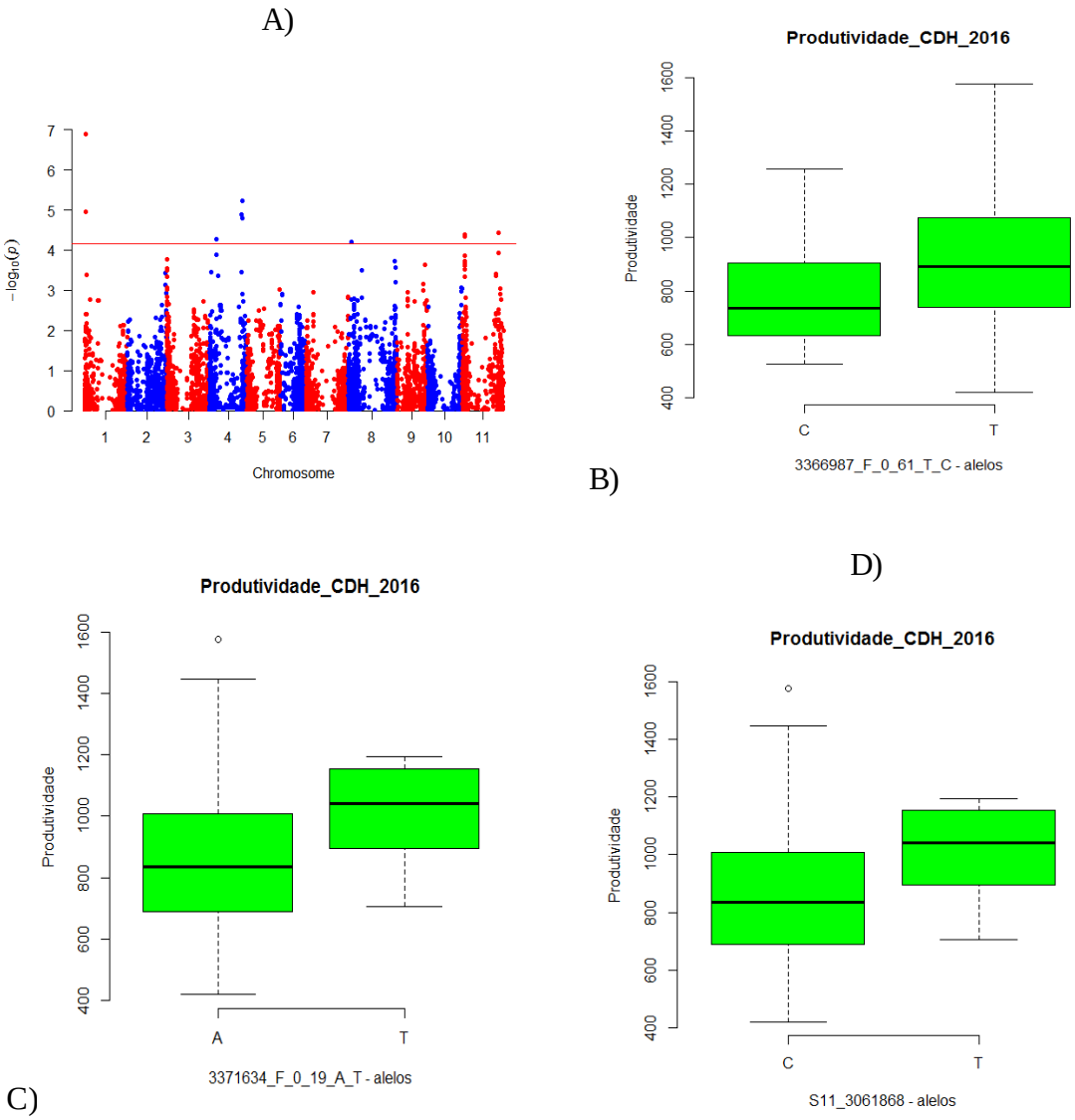
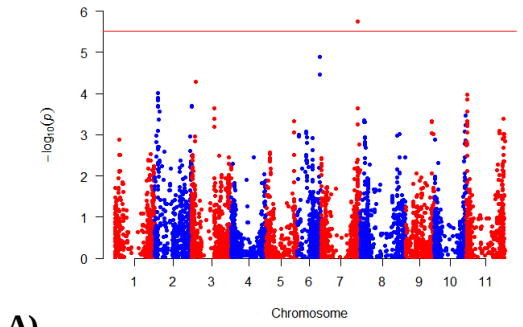
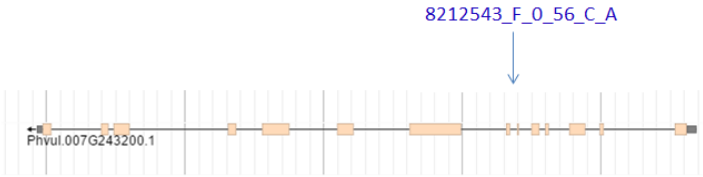



**Figura 7.** GWAS para produtividade no experimento com deficiência hídrica em 2016 (CDH-2016). A) Manhattan plot mostrando os SNPs significativos (*p*-valor ≤ 6,32E-05); B) Boxplot para o SNP "3366987_F_0_61_T_C", mostrando que os acessos com alelo T tiveram produtividade, significativamente, maior do que os que possuem o alelo C, pelo teste de Wilcoxon (p = 4,603E-04); C) Boxplot para o SNP "3371634_F_0_19_A_T", mostrando que os acessos com alelo T tiveram, em média, produtividade, significativamente, maior do que os que possuem o alelo A (p = 1,015E-02); D) Boxplot para o SNP "S11_3061868", mostrando que os acessos com alelo T tiveram, em média, produtividade, significativamente, maior do que os que possuem o alelo C (p = 1,01E-02).




A)

B)




C)

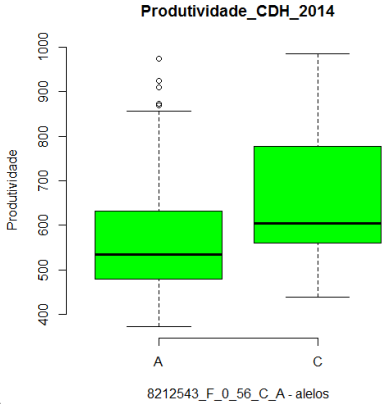
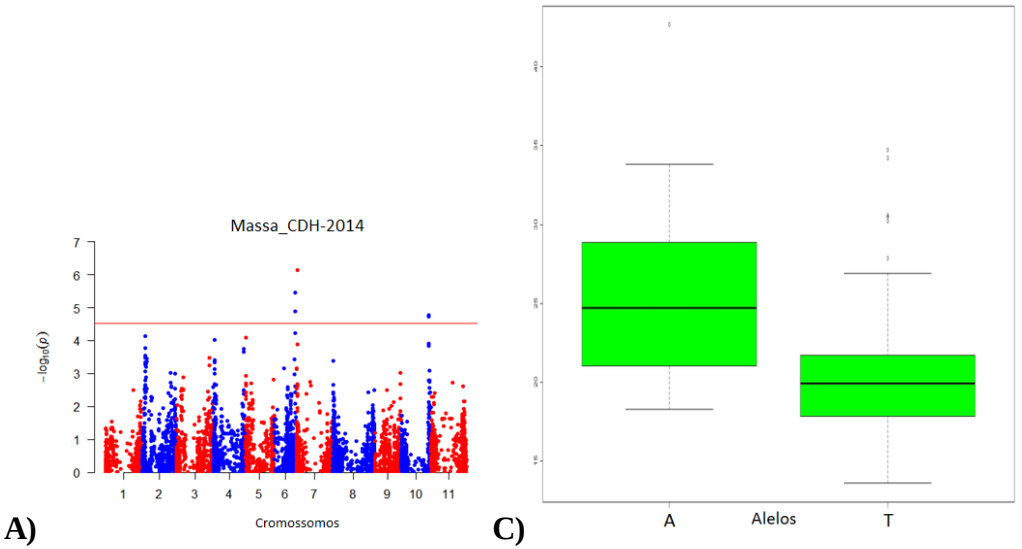



**Figura 8**. GWAS para produtividade no experimento com deficiência hídrica (CDH-2014). A) Manhattan plot (*p*-valor ≤ 1,74E-06); B) Posição do marcador 8212543_F_0_56_C_A (seta azul), detectado pelo GWAS no cromossomo 07, localizado no íntron do transcrito Phvul.007G243200 .1; C) Boxplot para massa de 100 grãos com base nos alelos do marcador 8212543_F_0_56_C_A. A diferença entre os grupos que contém cada a lelo foi analisada pelo teste de Wilcoxon (p = 1,911E-06).



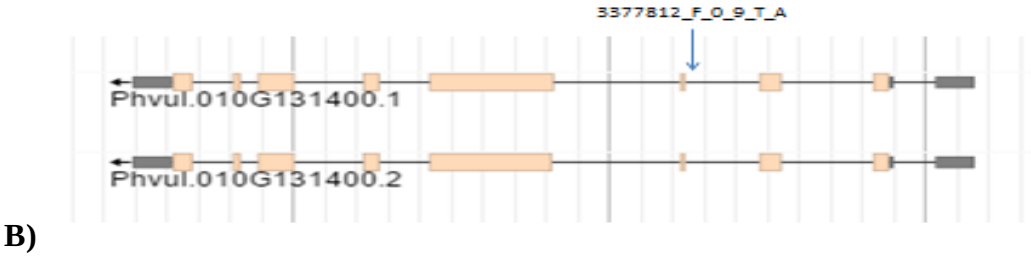




**Figura 9.** GWAS para massa de 100 grãos nos experimentos com deficiência hídrica (2014). A) Manhattan plot mostrando os SNPs significativos (*p*-valor ≤ 1,87E-05); B) Estrutura exon-íntron do gene Phvul.010G121400 e a posição do marcador 3377812_F_0_9_T_A (seta azul), detectado pelo GWAS no cromossomo 10, localizado na região intrônica/sítio de splice; C) Boxplots para massa de 100 grãos no experimento com deficiência hídrica em 2014 com base nos alelos do marcador 3377812_F_0_9_T_A . A diferença entre os grupos que contém cada alelo foi analisada pelo teste de Wilcoxon (p = 1,911E-06).

Considerando somente os experimentos com deficiência hídrica, foram detectados 24 SNPs significativos, dos quais um foi detectado em mais de um ambiente. Destes, 18 estão dentro de genes e cinco estão próximos, com distância variando entre 150 a 1885 pb (Tabela 7). Estes genes codificam para proteínas putativas com importância na tolerância à seca, tais como transportadores de membrana (*sec61*), enzimas com função de óxido-redução (*dihydroflavonol-4-reductase DFR1*, *cytochrome P450 704C1-like*), fatores de transcrição de ligação ao DNA (*polyadenylate-binding RBP47C-like*, *pre-rRNA-processing ESF2*, *myb transcription fator*, *bZIP transcription factor*), metabolismo de carboidratos (*glucan endo-1,3-beta-glucosidase 9*), canal de transporte de íons (*mechanosensitive ion channel 6-like*), transportador de ABA (*ABC transporter G family member 24-like*) e processo de ubitiquinação (*E3 ubiquitin- ligase UPL3*).

Foram preditos 463 efeitos para os 212 SNPs significativos através da análise no SnpEff. Quanto ao impacto predito para esses efeitos, 371 (80,13%) foram do tipo modificador, 50 (10,8%) foram considerados de baixo impacto, 37 (8%) de impacto moderado e cinco (1,08%) com alto impacto. Os efeitos foram observados, predominantemente, em sequências flanqueadoras de genes (5 kb) ou dentro de genes (93,09%), dos quais 18,58% foram observados dentro de exons e 18,14% dentro de regiões intrônicas (Figura 10).

**Tabela 7.** Anotação gênica dos SNPs significativos detectados nos ambientes com deficiência hídrica

| Caract/Ambiente | Marcador | Crom. | Dist. (kb) | Gene | Transcrito anotado | Descrição |
|---|---|---|---|---|---|---|
| Massa/CDH-2014 | 3378581_F_0_36_C_A | 6 | 1885 | Phvul.006G188800 | Phvul.006G188800.1 | TMV resistance N-like |
| | S06_29713507 | 6 | 0 | Phvul.006G189000 | Phvul.006G0189000.1 | enoyl- delta isomerase peroxisomal-like |
| | S07_1121648 | 7 | 1498 | Phvul.007G016600 | Phvul.007G016600.1 | mechanosensitive ion channel6-like |
| | S07_1121680 | 7 | 1466 | Phvul.007G016600 | Phvul.007G016600.1 | mechanosensitive ion channel6-like |
| | 3381021_F_0_34_C_A | 10 | 0 | Phvul.010G129500 | Phvul.010G129500.1 | mitochondrial intermembrane space import and assembly 40 |
| | 3377812_F_0_9_T_A | 10 | 0 | Phvul.010G131400 | Phvul.010G131400.1 | myb transcription factor [Medicago truncatula] |
| Prod/CDH-conj | S01_1681129* | 1 | 0 | Phvul.001G019600 | Phvul.001G019600.1 | polyadenylate-binding RBP47C-like |
| | 3383831_F_0_8_C_T | 2 | 0 | Phvul.002G046700 | Phvul.002G046700.1 | vacuole membrane KMS1-like |
| | S02_4341198 | 2 | 0 | Phvul.002G046900 | Phvul.002G046900.1 | plastocyanin-like domain [Medicago truncatula] |
| | 3384331_F_0_44_T_A | 6 | 217 | Phvul.006G005800 | Phvul.006G005800.1 | Breas cancer susceptibility 2 homolog B-like isoform X2 |
| | 8196854_F_0_58_A_G | 6 | 0 | Phvul.006G005900 | Phvul.006G005900.1 | breast cancer type 2 susceptibility [Medicago truncatula] |
| Prod/CDH-2014 | 8212543_F_0_56_C_A | 7 | 0 | Phvul.007G243200 | Phvul.007G243200.1 | ABC transporter G family member 24-like |

| Caract/Ambiente | Marcador | Crom. | Dist. (kb) | Gene | Transcrito anotado | Descrição |
|---|---|---|---|---|---|---|
| Prod/CDH-2016 | 3372117_F_0_49_T_C | 1 | 0 | Phvul.001G012800 | Phvul.001G012800.1 | dihydroflavonol-4-reductase DFR1 [Glycine max] |
| | S01_1681129* | 1 | 0 | Phvul.001G019600 | Phvul.001G019600.1 | polyadenylate-binding RBP47C-like |
| | 3383663_F_0_34_A_G | 4 | 0 | Phvul.004G060300 | Phvul.004G060300.1 | pre-rRNA-processing ESF2 [Vigna angularis] |
| | 3370657_F_0_63_C_T | 4 | 0 | Phvul.004G060400 | Phvul.004G060400.1 | Pentatricopeptide repeat-containing mitochondrial |
| | S04_39340571 | 4 | 0 | Phvul.004G122100 | Phvul.004G122100.1 | hypothetical protein PHAVU_004G122100g [Phaseolus vulgaris] |
| | 3366987_F_0_61_T_C | 4 | 0 | Phvul.004G124700 | Phvul.004G124700.1 | cytochrome P450 704C1-like |
| | 3378663_F_0_27_G_C | 4 | 0 | Phvul.004G124700 | Phvul.004G124700.1 | cytochrome P450 704C1-like |
| | 8206842_F_0_19_G_A | 8 | 0 | Phvul.008G051500 | Phvul.008G051500.1 | sister chromatid cohesion 1 4 |
| | S11_3061868 | 11 | 0 | Phvul.011G034900 | Phvul.011G034900.1 | myb-related 308 |
| | 3371634_F_0_19_A_T | 11 | 150 | Phvul.011G035200 | Phvul.011G035200.1 | E3 ubiquitin-ligase UPL3 |
| | S11_3134976 | 11 | 0 | Phvul.011G035700 | Phvul.011G035700.1 | bZIP transcription factor bZIP121 isoform X1 [Glycine max] |
| | 3366930_F_0_66_G_T | 11 | 0 | Phvul.011G169200 | Phvul.011G169200.1 | transport Sec61 subunit beta-like |

*SNPs detectados em mais de um ambiente

Caract: característica; Dist. (kb): distância do SNP em relação ao gene em kb

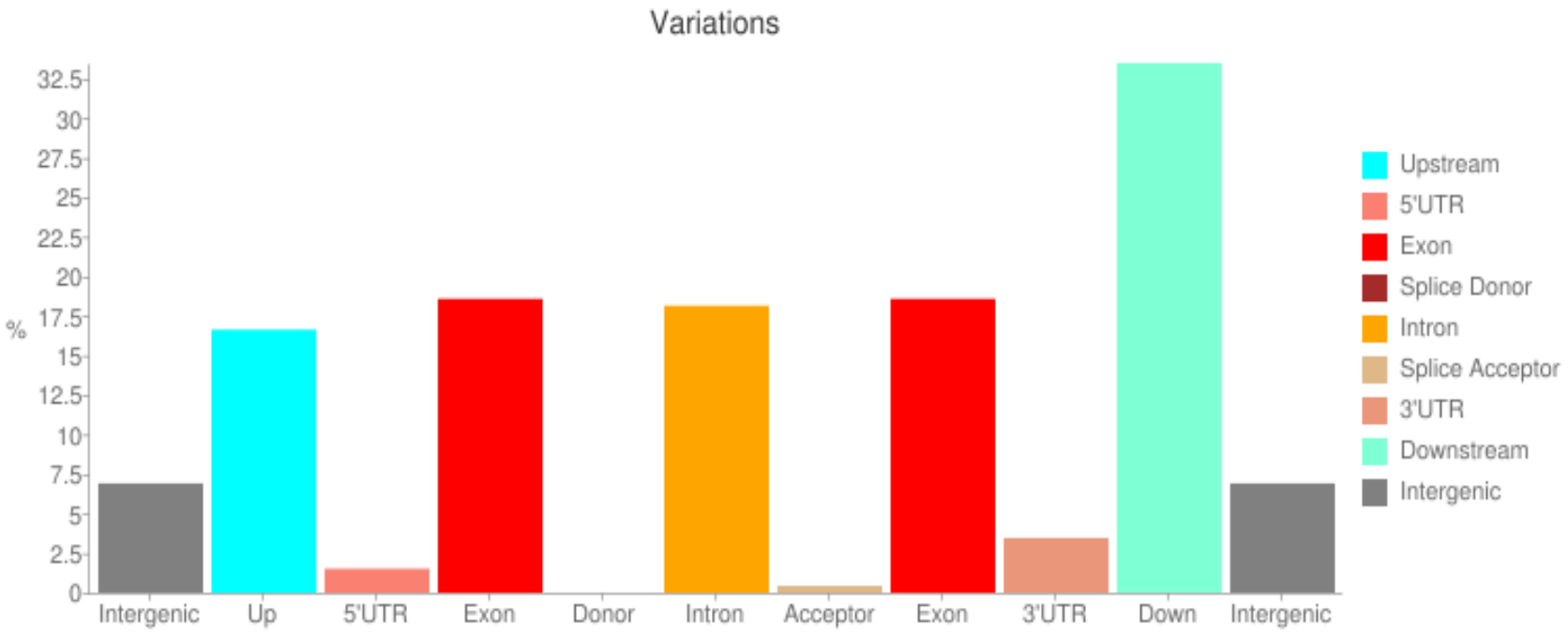



**Figura 10.** Número de efeitos dos SNPs por regiões no genoma

## DISCUSSÃO

O melhoramento para tolerância à seca é um grande desafio devido às variações na intensidade e duração dos episódios de seca, além da natureza quantitativa e baixa herdabilidade do caráter (Schneider et al. 1997, Blum 2011, Blair et al. 2012).

Os estudos de GWAS representam uma abordagem com grande potencial para a caracterização da arquitetura de características complexas e dependem, basicamente, de marcadores moleculares abundantes e distribuídos em todo genoma, um tamanho de amostra populacional suficientemente grande e geneticamente diverso para detectar um maior número de associações, cujos fenótipos sejam bem definidos e confiáveis. Nesse estudo, para detectar regiões genômicas associadas à tolerância à seca em feijoeiro comum, foi utilizado um conjunto de 343 acessos da CONFE, todos pertencentes ao pool gênico Mesoamericano, representado por germoplasma tradicional e cultivado, com grandes variações nas cores, tamanho e forma dos grãos (Rangel et al. 2013), além de serem diversos geneticamente (Cascão et al. 2014).

O ambiente e a época escolhidos para instalação dos experimentos de campo se mostraram adequados para avaliação dos caracteres associados à tolerância das plantas à deficiência hídrica, pois segundo Heinemann et al. (2007), a precipitação pluvial em Porangatu/GO no período de maio a agosto apresenta uma variação de 0 a 13 mm ao mês, com temperaturas máximas variando de 35,4 a 37,5 ºC. Portanto, nesta época de plantio, as plantas são submetidas ao estresse de alta temperatura e escassez de chuvas, o que permitiu o controle preciso da disponibilidade de água, através da irrigação nos momentos indicados pelas leituras no tensiômetro. Os dois caracteres fenotípicos avaliados nesse estudo, massa de 100 grãos e produtividade, apresentam alta e baixa herdabilidade, respectivamente. Segundo Rezende & Duarte (2007), para os caracteres de produtividade, em geral com alta influência ambiental e baixo coeficiente de determinação genética, os números de repetições usualmente empregados (entre dois e quatro) não permitem atingir a meta de 90% para acurácia seletiva, mesmo que avaliados em diferentes locais e anos e com baixos coeficientes de variação. Por isso, os autores recomendam, ao menos, seis repetições para avaliação destes caracteres. Apesar dos experimentos apresentarem somente duas repetições, valores de acurácia acima de 70% foram obtidos, o qual é considerado alto segundo Rezende e Duarte (2007), o que indica alta precisão experimental para avaliação dos caracteres agronômicos.

Os marcadores SNPs obtidos pelas tecnologias DArTseq e Capture-Seq se mostraram adequados para explorar o desequilíbrio de ligação ao longo do genoma de

feijoeiro comum e possibilitaram a identificação de SNPs tanto em DL, quanto diretamente associados aos genes de importância no caráter tolerância à seca. O fato de estas tecnologias identificarem polimorfismos em regiões ricas em genes (DArTseq) e em genes alvos (Capture-Seq), as tornam um excelente recurso nos ensaios de GWAS, e pode ter favorecido na identificação de um número considerável de polimorfismos associados aos fenótipos em estudo, principalmente para massa de 100 grãos, que é um caráter com alta herdabilidade (Schneider et al. 1997, Blair et al. 2012). O aproveitamento dos SNPs gerados pela tecnologia DArTseq foi alto e comparável com outros estudos que utilizaram a mesma tecnologia, atingindo altos valores de reprodutibilidade, com média de 99,8%, similar à média obtida para *Physaria* (99,7%) por Cruz et al. (2013). O valor médio de *Call rate* (93%) foi superior ao relatado para melancia (91,3%) por Ren et al. (2015), mas inferior ao relatado para *Physaria* (98,78%) por Cruz et al. (2013). A taxa de erros de replicatas (0,43%) foi inferior ao encontrado para café (4%) Garavito et al. (2016). Na análise por Capture-Seq, do total de 5050 genes alvos para o desenho das sondas, 3304 (65,42%) foram capturadas, valor superior ao relatado para *Pinus taeda* L. (57%) por Neves et al. (2013). Das sondas capturadas, 1701 (33,68%) contendo SNPs com MAF $\geq$ 0,05 foram selecionadas, obtendo, portanto, um aproveitamento de 33,68%, em relação ao número inicial de genes, valor este superior ao descrito por Neves et al. (2013) que foi de 28% para a espécie *Pinus taeda* L.. O método HMM, utilizado neste estudo para imputação dos dados faltantes, é o mais comum e tem mostrado melhor poder e resolução nos estudos de GWAS (Xavier et al. 2016). A taxa de erro para imputação obtida (1,55%) foi similar às obtidas por outros softwares que utilizam o mesmo método de imputação (Howie et al. 2009, Xavier et al. 2016).

Entre os vários fatores que afetam a eficiência do GWAS, a densidade de marcadores é um fator crucial. Em geral, o aumento da densidade de marcadores potencializa a resolução do mapeamento, aumentando assim a precisão para detecção de QTLs (Li et al. 2015). Este estudo gerou uma densidade média de marcadores SNPs, que mesmo após filtragem pela MAF e imputação de dados faltantes (8789 SNPs), apresentou uma ampla representatividade do genoma do feijão, com uma distribuição de 1 SNP/59 kpb. Este valor foi superior ao relatado por estudos prévios em feijoeiro comum, de 1SNP/86 kpb por Valdisser et al. (em elaboração) e 1 SNP/500 kpb (Valdisser et al. 2016).

A extensão do DL é outro fator determinante da eficiência do GWAS (Lipka et al. 2015). O decaimento do DL com a distância física entre os SNPs foi de 250 kb para $r^2 \sim 0{,}23$, o qual é inferior ao relatado para soja por Zhou et al. (2015, 420 kb) e superior ao relatado para Oryza sativa (123 kb e 167 kb em *indica* e *japonica*, respectivamente) por Huang et al.

(2010). A efetividade de um estudo de associação também depende da intensidade do DL entre o marcador molecular e a variante causal de determinado fenótipo (Myles et al. 2009). A estimativa de DL sem correções tem menor confiabilidade porque inclui DL entre marcadores a longas distâncias, devido a estruturação e parentesco genético entre os acessos. Usando a correção, os valores de DL são altos somente para os marcadores geneticamente ligados (Cericola et al. 2014). Neste estudo, a estimativa final para a extensão de DL foi de aproximadamente 66 kb. Este nível de DL é ideal para o método GWAS, uma vez que permite uma cobertura eficiente do genoma com base num número relativamente moderado de marcadores. Os valores de DL obtidos foram inferiores aos relatados por Valdisser et al. (em elaboração), estimados para 111 acessos Mesoamericanos, que foram de ~312 Kb e ~130 Kb, sem e com correção para estrutura e parentesco, respectivamente. Portanto, para as estimativas com correção do viés, o valor relatado no estudo atual foi, aproximadamente, 50% menor, o que pode ser explicado pelo maior tamanho amostral e consequente, acúmulo de maior número de eventos de recombinação, reduzindo o desequilíbrio de ligação.

A análise criteriosa de anotação dos SNPs associados aos caracteres agronômicos na presença e ausência da deficiência hídrica revelou uma extensa rede de termos associados a diversos mecanismos fisiológicos previamente identificados como responsivos ao déficit hídrico. A análise funcional do SNP "S01_1681129", que foi detectado associado à produtividade no ambiente CDH-2016 e na análise conjunta para os ambientes com deficiência hídrica, revelou que o mesmo está localizado dentro do gene Phvul.001G019600, o qual codifica para uma proteína de ligação ao RNA ("*polyadenylate-binding RBP47C-like*"). Essas proteínas de ligação ao RNA (RBP) têm papel essencial na regulação gênica pós-transcricional, exercendo importante função no crescimento, desenvolvimento e resposta a estresse, permitindo a adaptação das plantas as várias condições ambientais (Lorkovic 2009, Lee e Kang 2016). Em arroz, a superexpressão de RBP em plantas transgênicas mostram que estas tiveram maior taxa de recuperação e produtividade em comparação as plantas não transgênicas sob condições de deficiência hídrica (Yang et al. 2014). A função da RBP47C ainda não é conhecida, mas um estudo recente em Arabidposis, conduzido por Narayanan et al. (2014), usando um mutante contendo um gene defeituoso que codifica RBP47, demonstrou que esta proteína tem função na síntese de carotenóides. Sabe-se que os carotenóides tem vários papéis no desenvolvimento das sementes, principalmente induzindo a produção de ABA (Wurtzel et al. 2001). O fitohormônio ABA atua no fechamento dos estômatos para evitar a perda de água (Taiz e Zeiger 2006, Zhang et al. 2006). Durante a seca os níveis de ABA aumentam drasticamente nas plantas (Zeevaart et al. 1990), já que sua produção é

importante para iniciar respostas precoces de sobrevivência ao estresse hídrico (Zhang et al. 2006).

Outros SNPs associados à produtividade e identificados no ambiente CDH-2016 foram os 3366987_F_0_61_T_C e 3371634_F_0_19_A_T. O primeiro está dentro de um gene cujo transcrito está associado à proteína putativa "*cytochrome P450 704C1-like*". Vários estudos tem relatado o papel desta proteína no controle da produção de ABA em plantas (Kushiro et al. 2004, Saito et al. 2004). Uma das enzimas que atuam no catabolismo do ABA, a 8'-hidroxilase é uma proteína P450 (Ueno et al. 2005). Kitahata et al. 2005, demonstraram que a inibição desta enzima pode aumentar o contéudo de ABA nas plantas e aumentar a resistência a deficiência hídrica. Os genótipos que apresentaram o alelo T para este marcador apresentaram média de produtividade, significativamente, maior do que os acessos que tem o alelo C. Portanto, supõe-se que um polimorfismo que atue reduzindo ou inibindo a expressão desta enzima, seja vantajoso para a planta em momentos de déficit hídrico. Segundo Nepomuceno et al. (2000) a desativação da expressão de determinados genes também pode estar associado ao aumento da tolerância à seca. Já o segundo SNP, está a 150 pb do gene que codifica para "*E3 ubiquitin- ligase UPL3*". Esta proteína está envolvida nos processos de ubiquitinação, que consiste em um sistema de degradação de proteínas utilizado pelos organismos eucariotos para degradar, eficientemente, as proteínas celulares prejudiciais e controlar todo o conjunto de componentes reguladores. Em plantas, a adaptação em resposta a vários estresses abióticos pode ser conseguida através da ubiquitinação e da degradação resultante de componentes específicos para a sinalização do estresse (Lee e Kim 2011). Vários estudos demonstraram que as ligases E3 estão envolvidas em vias dependentes ou independentes de ABA em resposta ao estresse de seca (Zhang et al. 2007, Zhang et al. 2008).

O único SNP (8212543_F_0_56_C_A), associado à produtividade, que foi detectado no ambiente CDH-2014, está presente no gene que codifica para a proteína putativa "*ABC transporter G family member 24-like*". Através de análises genéticas e bioquímicas, muitos estudos tem relatado o papel destes transportadores na exportação de ABA através da membrana plasmática e sua função nas vias de sinalização intercelular, sugerindo, fortemente, que a presença do mecanismo de transporte atue sobre o controle ativo do ABA entre as células vegetais, provocando respostas multicelulares aos estresses ambientais (Kuromori et al. 2010, Kuromori e Shinozaki 2010). Segundo Kang et al. (2010), um transportador de ABA é muito relevante durante as condições de estresse, já que nesta situação o pH extracelular é aumentado, fazendo com que o ABA se dissocie na sua forma carregada, dificultando sua difusão passiva através da bicamada lipídica, o que é contraditório com a necessidade de

liberar rapidamente o hormônio do stress para a célula com o objetivo de se obter uma resposta rápida (Wilkinson e Davies 1997). Por isso, a localização intracelular destes transportadores de ABA destaca a importância da captação de ABA na célula para que ocorram processos de sinalização celular e demonstra o potencial de um transportador, o qual poderia fornecer ABA de uma forma rápida e regulada para iniciar respostas imediatas e controladas às várias condições de estresse, como a seca, percebidas pelo hormônio.

A sinalização do estresse de seca pode ser dividida em dois tipos de vias baseadas na dependência ou independência do hormônio ABA. A via dependente de ABA é mediada por dois sistemas diferentes (Agarwal et al. 2006). No primeiro sistema, os fatores de transcrição bZIP se ligam à elementos *cis-acting* ABRE presentes em genes responsivos à seca, tais como RD29B, AIL1 e RAB18, ativando-os (Fujita et al. 2005, Nakashima et al. 2009). O segundo sistema é controlado por fatores de transcrição MYC/MYB. Estas proteínas ligam-se a elementos *cis-acting*, tais como MYCRS e MYBRS (sequência de reconhecimento MYC e MYB, respectivamente) e ativam a transcrição do gene responsivo à seca RD22 (Abe et al. 1997). Estes genes responsivos à seca atuam na síntese e transdução de sinais para aumento do ABA nas plantas e consequente, tolerância à deficiência hídrica. Alguns SNPs detectados neste estudo (S11_3061868, S11_3134976), associados com massa de 100 grãos e produtividade, estão presentes em genes que codificam proteínas homólogas aos fatores de transcrição bZIP e MYB.

Nos ambientes com irrigação adequada houve detecção de SNPs somente associados com a massa de 100 grãos. Estes SNPs estão dentro ou próximo de importantes genes associados ao desenvolvimento das sementes, tais como metabolismo de aminoácidos, lipídeos e carboidratos, biossíntese de flavonóides e carotenóides, fosforilação oxidativa (Lepiniec et al. 2006, Frey et al. 2006, Fatihi et al. 2013), com destaque para o SNP "3380824_F_0_8_T_A". Este SNP, detectado no ambiente SDH-2014 e na análise conjunta para os ambientes irrigados está próximo ao gene (~ 3Kb) Phvul.006G077200 que codifica para a proteína putativa "*expansin*". As proteínas "*expansins*" tem função conhecida na extensão da parede celular durante o crescimento da célula. Xu et al. (2013) demonstrou que plantas transgênicas de Arabidopsis e algodão que superexpressaram a proteína "*expansins*" tiveram aumento no tamanho das sementes. Mais recentemente, em estudo com feijão, Schmutz et al. (2014) também detectaram através de GWAS genes candidatos associados a massa de grãos com funções putativas de "*expansin*".

Os resultados deste estudo estão de acordo com estudos prévios de caracteres associados à seca (Schneider et al. 1997; Blair et al. 2012), nos quais o número de QTLs

associados à produtividade foi baixo e para massa de grãos foi alto. Neste estudo houve detecção de SNPs associados à produtividade em ambientes com deficiência hídrica nos cromossomos 1, 2, 4, 6, 7, 8 e 11. Em feijão, QTLs associados à produtividade foram relatados nos cromossomos 1, 2, 5 e 10 (Trapp et al. 2015) e nos cromossomos 3 e 9 por diversos autores (Blair et al. 2006, 2012; Checa e Blair 2012; Wright e Kelly 2011; Mukeshimana et al. 2014). Certamente, essa variação dos QTLs identificados em diferentes cromossomos está, provavelmente, relacionada à alta influência ambiental nesta característica, às diferentes condições de estresse aplicadas e às variações no *background genético* do germoplasma avaliado, ocasionando baixa estabilidade na detecção de QTLs. Já para a característica massa de 100 grãos foram identificados QTLs em todos os cromossomos, com o maior número presente no cromossomo 3 (45 SNPs). Estudos anteriores revelaram QTLs nos cromossomos 1, 2, 8 (Trapp et al. 2015), 3 e 7 (Mukeshimana et al. 2014) e 9 (Blair et al. 2006, Mkwaila et al. 2011). O maior número de QTLs associados ao peso de grãos identificados no estudo atual pode ser resultante da maior variabilidade genética existente entre os acessos, enquanto na maioria dos estudos foram utilizadas populações biparentais.

Um estudo de GWAS para tolerância à seca em feijoeiro comum foi conduzido por Hoyos-Villegas et al. (2016) em um painel contendo 96 genótipos Mesoamericanos, avaliados para diferentes caracteres, tais como acamamento, peso de 100 grãos, biomassa de florescimento e produtividade. Considerando o total de 27 SNPs significativos identificados, dois SNPs foram associados à massa de 100 grãos, detectados em ambientes com e sem deficiência hídrica, e estão localizados no cromossomo 9. As posições destes SNPs não coincidiram com os SNPs detectados em nossos estudos. Diferentemente do nosso estudo, Hoyos-Villegas et al. (2016) não encontraram SNPs associados à produtividade. Provavelmente, o tamanho do painel de associação (343 acessos) favoreceu um mapeamento com maior resolução e detecção de QTLs associados à produtividade, além de maior número de QTLs associados à massa de 100 grãos.

A proporção da variação fenotípica, explicada pelos SNPs significativos detectados nos ambientes com deficiência hídrica, variou de 5,62% a 14,39%, o que é consistente com a complexidade genômica desta característica, a qual é controlada por vários genes com efeitos menores. Valores similares foram relatados por Mukeshimana et al. (2014, $R^2$ = 8,3 a 20,2%), Asfaw e Blair (2012; $R^2$ = 9 a 14%), Blair et al. (2012; $R^2$ = 11 a 24%) e Trapp et al. (2016; $R^2$ = 16%) através da análise de populações biparentais avaliadas sob condições de deficiência hídrica em feijoeiro comum.

Para a seleção de genótipos superiores em condições de deficiência hídrica, também se deve considerar sua produtividade em condições adequadas de irrigação. Segundo Jongdee et al. (2006), cultivares com boa produtividade podem ser desenvolvidas em condições de deficiência hídrica e, ao mesmo tempo, responder bem às condições favoráveis de umidade do solo, desde que sejam avaliadas em ambos os ambientes. O genótipo CF840047, uma variedade tradicional pertencente ao grupo comercial mulatinho, foi a que teve menor redução de produtividade (45,87%), comparando os ambientes com e sem deficiência hídrica, apresentando produção de 1579 e 855 $Kgha^{-1}$, respectivamente. Um dos genótipos que se destacou neste estudo foi a cultivar Ouro Negro (grupo comercial preto), pois apresentou maior produtividade nos ambientes com déficit hídrico e a segunda mais produtiva nos ambientes sob irrigação adequada, apesar das perdas de produtividade estimadas chegarem a 58%. Este genótipo apresentou vários dos alelos com desempenho superior para os marcadores encontrados dentro ou muito próximos de genes com funções preditas importantes para tolerância à seca. Isto demonstra a importância de se combinar mais alelos favoráveis em uma cultivar para que o rendimento da produtividade seja aumentado, objetivo este que pode ser potencializado pelo uso da seleção assistida por marcadores (SAM).

A seleção assistida por marcadores, baseada nos alelos identificados em estudos de GWAS, necessita de uma forte associação entre as regiões genômicas avaliadas e o caráter de interesse. Caso a associação não seja tão forte, torna-se preferível optar pela seleção fenotípica. Por isso é importante desenvolver análises GWAS de alta qualidade, com bons experimentos de fenotipagem e ampla cobertura de SNPs (Elshire et al. 2011). A identificação das variantes alélicas associadas aos QTLs é potencializada quando um grupo diverso de germoplasma é utilizado favorecendo a identificação de alelos que, em conjunto, possam ser empregados na busca por genótipos com desempenho destacado para tolerância à seca. Dessa forma, os resultados da análise de GWAS podem ser convertidos em uma plataforma de genotipagem para estudos moleculares em feijão. Estudos de validação de SNPs associados à tolerância à seca, detectados através de GWAS, foram realizados em arroz (Pantalião et al. 2016) e grão de bico (Kujur et al. 2015), e mostraram grande potencial para uso dos marcadores validados na seleção de genótipos com performance superior para a característica avaliada. Portanto, nossos resultados podem ser usados como ponto de partida para o desenvolvimento de marcadores úteis aos programas de melhoramento genético do feijoeiro comum para tolerância à seca. Após etapas de validação, os alelos dos SNPs associados ao fenótipo de tolerância à seca podem, então, ser usados para selecionar acessos de feijão a

partir de bancos de germoplasma ou para monitorar o alelo presente em procedimentos de introgressão em retrocruzamentos, sem a necessidade de realizar ensaios de campo para avaliar tolerância à seca nestes materiais.

## REFERÊNCIAS BIBLIOGRÁFICAS


Abdurakhmonov IY, Abdukarimov A (2008) Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int J Plant Genomics. doi: 10.1155/2008/574927

Abe H, Yamaguchi-Shinozaki K, Urao T, Iwasaki T, Shinozaki K (1997) Role of MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. Plant Cell 9**:**1859–1868.

Acosta-Gallego JA, White W (1995) Phenological plasticity as an adaptation by common bean rainfed environment. Crop Science 35:199-204.

Acosta-Gallegos JA, Adams MW (1991) Plant traits and yield stability of dry bean (*Phaseolus vulgaris* L:) cultivars under drought stress.J.Agr.Sci.117:213–219.

Agarwal PK, Agarwal P, Reddy MK, Sopory SK (2006) Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. Plant Cell Reports 25:1263-1274.

Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. doi: 10.1093/nar/25.17.3389

Amir Hossain M, Lee Y, Cho J-I, et al (2010) The bZIP transcription factor OsABF1 is an ABA responsive element binding factor that enhances abiotic stress signaling in rice. Plant Mol Biol 72:557–566. doi: 10.1007/s11103-009-9592-9

Asfaw A, Blair MW (2012) Quantitative trait loci for rooting pattern traits of common beans grown under drought stress versus non-stress conditions. Mol Breed 30:681–695. doi: 10.1007/s11032-011-9654-y

Ashburner M, Ball CA, Blake JA, et al (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25:25–29. doi: 10.1038/75556

Beebe SE, Rao IM, Blair MW, Acosta-Gallegos JA (2010) Phenotyping common beans for adaptation to drought. in: J. M. Ribaut, and P. Monneveux (eds), Drought Phenotyping in Crops: From Theory to Practice, 311—334. Generation Challenge Program Special Issue on Phenotyping.

Blair MW, Galeano CH, Tovar E, et al (2012) Development of a Mesoamerican intra-

genepool genetic map for quantitative trait loci detection in a drought tolerant × susceptible common bean (Phaseolus vulgaris L.) cross. Mol Breed 29:71–88. doi: 10.1007/s11032-010-9527-9

Blair MW, Iriarte G, Beebe S (2006) QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean x wild common bean (Phaseolus vulgaris L.) cross. Theor Appl Genet 112:1149–1163. doi: 10.1007/s00122-006-0217-2

Blum A (2011) Drought resistance - is it really a complex trait? Functional Plant Biology 38:753-757.

Bradbury PJ, Zhang Z, Kroon DE, et al (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. doi: 10.1093/bioinformatics/btm308

Brick MA, Ogg JB, Singh SP, et al (2008) Registration of Drought-Tolerant , Bean Germplasm Line CO46348. J Plant Regist 2:120–124. doi: 10.3198/jpr2007.06.0359crg

Broughton WJ, Hernández G, Blair M, et al (2003) Beans (Phaseolus spp.) - Model food legumes. Plant Soil 252:55–128. doi: 10.1023/A:1024146710611

Cascão LM, Cieslak JF, Oliveira JP de, et al (2014) Caracterização da estrutura populacional da Coleção Nuclear de Feijão da Embrapa por marcadores SSR. Santo Antônio de Goiás: Embrapa Arroz e Feijão. 8º Seminário Jovens Talentos: Coletânea de Resumos, p 45.

Cavalieri A, Merchant A, Van Volkenburgh E (2011) Why not beans? Funct Plant Biol 38:984. doi: 10.1071/FPv38n12_FO

Cericola F, Portis E, Lanteri S, et al (2014) Linkage disequilibrium and genome-wide association analysis for anthocyanin pigmentation and fruit color in eggplant. BMC Genomics 15:896.

Checa OE, Blair M (2012) Inheritance of Yield-Related Traits in Climbing Beans (Phaseolus vulgaris L.). Crop Science 52:1998–2013.

Cichy KA, Wiesinger JA, Mendoza FA (2015) Genetic diversity and genome-wide association analysis of cooking time in dry bean (Phaseolus vulgaris L.). Theor Appl Genet 128:1555–1567. doi: 10.1007/s00122-015-2531-z

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly 6(2):80-92.

Conesa A, Götz S, García-Gómez JM, et al (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–

3676. doi: 10.1093/bioinformatics/bti610

Consortium TGO (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43:D1049–D1056. doi: 10.1093/nar/gku1179

Cruz VM V, Kilian A, Dierig DA (2013) Development of DArT Marker Platforms and Genetic Diversity Assessment of the U.S. Collection of the New Oilseed Crop Lesquerella and Related Species. PLoS One 8:1–13. doi: 10.1371/journal.pone.0064062

Desrousseaux D, Sandron F, Siberchicot A, et al (2013) LDcorSV: Linkage disequilibrium corrected by the structure and the relatedness.

Du H, Wang Y Bin, Xie Y, et al (2013) Genome-wide identification and evolutionary and expression analyses of MYB-related genes in land plants. DNA Res 20:437–448. doi: 10.1093/dnares/dst021

Duitama J, Quintero JC, Cruz DF, et al (2014) An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. Nucleic Acids Res 42:1–13. doi: 10.1093/nar/gkt1381

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361. doi: 10.1007/s12686-011-9548-7

Elshire RJ, Glaubitz JC, Sun Q, et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:1–10. doi: 10.1371/journal.pone.0019379

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol Ecol 14:2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

FAO (2015) The impact of natural hazards and disasters on agriculture and food security and nutrition. Disponível em http://www.fao.org/3/a-i4434e.pdf. Acesso em 22/12/2016.

FAO (2016a). 2016 International Year of Pulses. Disponível em http://www.fao.org/pulses-2016/en/. Acesso em 23/12/2016.

FAO (2016b). Faostat. Disponível em:www.fao.org/faostat/en/#data/QC. Acesso: 10 de dezembro de 2016.

Fatihi A, Zbierzak AM, Dörmann P (2013) Alterations in seed development gene expression affect size and oil content of Arabidopsis seeds. Plant Physiol 163:973–85. doi: 10.1104/pp.113.226761

Fischer RA, Maurer R (1978) Drought resistance in spring wheat cultivars: I. Grain yield responses. Aust. J. Agric. Res. 29:897-912.

Frahm MA, Rosas JC, Mayek-Pérez N, et al (2004) Breeding beans for resistance to terminal drought in the lowland tropics. Euphytica 136:223–232. doi: 10.1023/B:EUPH.0000030678.12073.a9

Frey A, Boutin JP, Sotta B, et al (2006) Regulation of carotenoid and ABA accumulation during the development and germination of Nicotiana plumbaginifolia seeds. Planta 224:622–632. doi: 10.1007/s00425-006-0231-2

Fujita Y, Fujita M, Satoh R, Maruyama K, Parvez MM, Seki M, et al (2005) AREB1 is a transcription activator of novel ABRE-dependent ABA signaling that enhances drought stress tolerance in *Arabidopsis*. Plant Cell 17:3470–3488.

Gallegos JAA, Shibata JK (1989) Effect of Water Stress on Growth and Yield of Indeterminate Dry-Bean (Phaseolus vulgaris) Cultivars. Science (80- ) 20:81–93.

Gepts P, Aragão FJL, Barros E, Blair MW, Brondani R, Broughton W, Galasso I, Hernández G, Kami J, Lariguet P, McClean P, Melotto M, Miklas P, Pauls P, Pedrosa-Harand A, Porch T, Sánchez F, Sparvoli F, Yu K (2008) Genomics of Phaseolus beans, a major source of dietary protein and micronutrients in the tropics. Genomics of Tropical Crop Plants.P. Moore and R. Ming. Berlin, Springer.: 113-143.

Goodstein DM, Shu S, Howson R, et al (2012) Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res 40:1178–1186. doi: 10.1093/nar/gkr944

Halterlein AJ (1983). Bean. In: Teare, I.D. and Peet, M.M. (eds). Crop-Water Relations. John Wiley, New York.

Heinemann AB, da Silva SC, Junior SL, et al (2007) Características climáticas dos municípios de Santo Antônio de Goiás (GO), Porangatu (GO), Janaúba (MG), Sete Lagoas ( MG ), Parnaíba ( PI ) e Teresina (PI), Brasil. Santo Antônio de Goiás: Embrapa Arroz e Feijão. Documentos, 214, 36p.

Henderson CR (1984) Applications of Linear Models in Animal Breeding Models. Univ Guelph 384. doi: 10.1002/9780470316856.ch7

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33:54–78. doi: 10.1016/0040-5809(88)90004-4

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. doi: 10.1371/journal.pgen.1000529

Huang X, Wei X, Sang T, et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967. doi: 10.1038/ng.695

Jongdee B, Pantuwan G, Fukai S, Fischer K (2006) Improving drought tolerance in rainfed

lowland rice: An example from Thailand. Agric Water Manag 80:225–240. doi: 10.1016/j.agwat.2005.07.015

Kamfwa K, Cichy KA, Kelly JD (2015) Genome-wide association analysis of symbiotic nitrogen fixation in common bean. Theor Appl Genet 128:1999–2017. doi: 10.1007/s00122-015-2562-5

Kang J, Hwang J-U, Lee M, et al (2010) PDR-type ABC transporter mediates cellular uptake of the phytohormone abscisic acid. Proc Natl Acad Sci 107:2355–2360. doi: 10.1073/pnas.0909222107

Kang JY, Choi HI, Im MY, Kim SY (2002) Arabidopsis basic leucine zipper proteins that mediate stress- responsive abscisic acid signaling. Plant Cell 14:343–357. doi: 10.1105/tpc.010362.tase

Katiyar A, Smita S, Lenka SK, et al (2012) Genome-wide classification and expression analysis of MYB transcription factor families in rice and Arabidopsis. BMC Genomics 13:544. doi: 10.1186/1471-2164-13-544

Kitahata N, Nakano T, Kuchitsu K, et al (2005) Biotin-labeled abscisic acid as a probe for investigating abscisic acid binding sites on plasma membranes of barley aleurone protoplasts. Bioorganic Med Chem 13:3351–3358. doi: 10.1016/j.bmc.2005.03.017

Kujur A, Bajaj D, Upadhyaya HD, et al (2015) A genome-wide SNP scan accelerates trait-regulatory genomic loci identification in chickpea. Sci Rep 5:11166. doi: 10.1038/srep11166

Kuromori T, Miyaji T, Yabuuchi H, et al (2010) ABC transporter AtABCG25 is involved in abscisic acid transport and responses. Proc Natl Acad Sci 107:2361–2366. doi: 10.1073/pnas.0912516107

Kuromori T, Shinozaki K (2010) ABA transport factors found in Arabidopsis ABC transporters. Plant Signal Behav 5:1124–1126. doi: 10.4161/psb.5.9.12566

Kushiro T, Okamoto M, Nakabayashi K, et al (2004) The Arabidopsis cytochrome P450 CYP707A encodes ABA 8'-hydroxylases: key enzymes in ABA catabolism. EMBO J 23:1647–56. doi: 10.1038/sj.emboj.7600121

Lanna AC, Mitsuzono ST, Gledson T, et al (2016) Physiological characterization of common bean (Phaseolus vulgaris L.) genotypes , water- stress induced with contrasting response towards drought. AJCS 10(1):1–6.

Lauer JG, Bijl CG, Grusak MA, et al (2012) The scientific grand challenges of the 21st century for the Crop Science Society of America. Crop Sci 52:1003–1010. doi:10.2135/cropsci2011.12.0668

Lee JH, Kim WT (2011) Regulation of abiotic stress signal transduction by E3 ubiquitin ligases in arabidopsis. Mol Cells 31:201–208. doi: 10.1007/s10059-011-0031-9

Lee K, Kang H (2016) Emerging Roles of RNA-Binding Proteins in Plant Growth, Development, and Stress Responses. Mol Cells 39:179–185. doi: 10.14348/molcells.2016.2359

Lepiniec L, Debeaujon I, Routaboul J-M, et al (2006) Genetics and Biochemistry of Seed Flavonoids. Annu Rev Plant Biol 57:405–430. doi: 10.1146/annurev.arplant.57.032905.105252

Lesk C, Rowhani P, Ramankutty N (2016) Influence of extreme weather disasters on global crop production. Nature 529:84–87. doi: 10.1038/nature16467

Levitt J (1972) Responses of Plants to Environmental Stresses. New York, NY: Academic Press, 698.

Lipka AE, Tian F, Wang Q, et al (2012) GAPIT: Genome association and prediction integrated tool. Bioinformatics 28:2397–2399. doi: 10.1093/bioinformatics/bts444

Lipka AE, Kandianis CB, Hudson ME, et al (2015) From association to prediction: Statistical methods for the dissection and selection of complex traits in plants. Curr Opin Plant Biol 24:110–118. doi: 10.1016/j.pbi.2015.02.010

Lorković ZJ (2009) Role of plant RNA-binding proteins in development, stress response and genome organization. Trends Plant Sci 14:229–236. doi: 10.1016/j.tplants.2009.01.007

Mamidi S, Rossi M, Moghaddam SM, et al (2013) Demographic factors shaped diversity in the two gene pools of wild common bean *Phaseolus vulgaris* L. Heredity (Edinb) 110:267–76. doi: 10.1038/hdy.2012.82

McClean PE, Burridge J, Beebe S, et al (2011) Crop improvement in the era of climate change: An integrated, multi-disciplinary approach for common bean (Phaseolus vulgaris). Funct Plant Biol 38:927–933. doi: 10.1071/FP11102

Moghaddam SM, Mamidi S, Osorno JM, et al (2016) Genome-Wide Association Study Identifies Candidate Loci Underlying Agronomic Traits in a Middle American Diversity Panel of Common Bean. Plant Genome 9:1-21. doi: 10.3835/plantgenome2016.02.0012

Mukeshimana G, Butare L, Cregan PB, et al (2014) Quantitative trait loci associated with drought tolerance in common bean. Crop Sci 54:923–938. doi: 10.2135/cropsci2013.06.0427

Muñoz-Perea CG, Terán H, Allen RG, Wright JL, Westermann DT, Singh SP (2006) Selection for drought resistance in dry bean landraces and cultivars. Crop Sci. 46:2111–2120.

Myles S, Peiffer J, Brown PJ, et al (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21:2194–2202. doi: 10.1105/tpc.109.068437

Nakashima K, Fujita Y, Kanamori N, et al (2009) Three *Arabidopsis*SnRK2 protein kinases, SRK2D/SnRK2.2, SRK2E/SnRK2.6/OST1 and SRK2I/SnRK2.3, involved in ABA signaling are essential for the control of seed development and dormancy. Plant Cell Physiol 50: 1345–1363.

Narayanan K (2014) Screening and Characterization of *Arabidopsis thaliana* mutants with altered carotenoid profile. Tese (Doutorado). University of Saskatchewan, Saskatoon. 182 p.

Nepomuceno AL, Stewart JM, Oostehuis DM, et al (2000) Isolation of a cotton NADP(H) oxidasehomologue induced by drought stress. Pesquisa Agropecuária Brasileira. 35:1407-1416.

Neves LG, Davis JM, Barbazuk WB, Kirst M (2013) Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J 75:146–156. doi: 10.1111/tpj.12193

Pantalião GF, Narciso M, Guimarães C, et al (2016) Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. Genetica 144:651–664. doi: 10.1007/s10709-016-9932-z

Paquette SR (2012) PopGenKit: useful functions for (batch) file conversion and data resampling in microsatellite databases.

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58:545-554.

Polania J, Rao IM, Cajiao C, et al (2016) Physiological traits associated with drought resistance in Andean and Mesoamerican genotypes of common bean (Phaseolus vulgaris L.). Euphytica 210:17–29. doi: 10.1007/s10681-016-1691-5

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959. doi: 10.1111/j.1471-8286.2007.01758.x

Purcell S, Neale B, Todd-Brown K, et al (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575. doi: 10.1086/519795

R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rabinowicz PD, Citek R, Budiman MA, et al (2005) Differential methylation of genes and

repeats in land plants. Genome Res 15:1431–1440. doi: 10.1101/gr.4100405

Ramirez-Vallejo P, Kelly JD (1998) Traits related to drought resistance in common bean. Euphytica 99:127–136. doi: 10.1023/A:1018353200015

Rangel PHN, Oliveira JP de, Costa JGC da, et al (2013) Banco ativo de germoplasma de arroz e feijão: passado, presente e futuro. Santo Antônio de Goiás: Embrapa Arroz e Feijão. Documentos 288, 59 p.

Rao IM, Beebe S, Polania J, Ricaurte J, Cajiao C, Garcia R (2004) Evaluation of drought resistance and associated traits in advanced lines, in Annual Report 2004. Project IP-1: Bean Improvement for the Tropics (Cali, Colombia: CIAT; ), 5–13.

Ren R, Ray R, Li P, et al (2015) Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. Mol Genet Genomics 290:1457–1470. doi: 10.1007/s00438-015-0997-7

Resende MDV de, Duarte JB (2007) Precisão e controle de qualidade em experimentos de avaliação de cultivares. Pesqui Agropecuária Trop 37:182–194. doi: 10.5216/pat.v37i3.1867

Rosales-Serna R, Kohashi-Shibata J, Acosta-Gallegos JA, et al (2004) Biomass distribution, maturity acceleration and yield in drought-stressed common bean cultivars. F Crop Res 85:203–211. doi: 10.1016/S0378-4290(03)00161-8

Ryu MY, Cho SK, Kim WT (2010) The Arabidopsis C3H2C3-type RING E3 ubiquitin ligase AtAIRP1 is a positive regulator of an abscisic acid-dependent response to drought stress. Plant Physiol 154:1983–97. doi: 10.1104/pp.110.164749

Sánchez-Sevilla JF, Horvath A, Botella MA, et al (2015) Diversity arrays technology (DArT) marker platforms for diversity analysis and linkage mapping in a complex crop, the octoploid cultivated strawberry (*Fragaria* x *ananassa*). PLoS One 10(12): e0144960. doi: 10.1371/journal.pone.0144960

Saito S, Hirai N, Matsumoto C, et al (2004) Arabidopsis CYP707As Encode (+)-Abscisic Acid 8'-Hydroxylase, a Key Enzyme in the Oxidative Catabolism of Abscisic Acid. Plant Physiol 134:1439–1449. doi: 10.1104/pp.103.037614.1

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644. doi: 10.1086/502802

Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–13. doi:

10.1038/ng.3008

Schneider KA, Rosales-Serna R, Ibarra-Perez F, et al (1997) Improving common bean performance under drought stress. Crop Science 37:43-50.

Shao HB, Chu LY, Jaleel CA, et al (2009) Understanding water deficit stress-induced changes in the basic metabolism of higher plants-biotechnologically and sustainably improving agriculture and the ecoenvironment in arid regions of the globe. Crit Rev Biotechnol 29:131-151.

Silveira PM, Stone LF (1994) Manejo da irrigação do feijoeiro: uso do tensiômetro e avaliação do desempenho do pivô central. Brasília: EMBRAPA, Circular Técnica 27, 46p.

Singh SP (1995) Selection for water stress tolerance in inter-racial populations of common bean. Crop Sci. 35:118–124. doi:10.2135/cropsci1995.0011183X003500010022x

Singh SP (2001) Broadening the genetic base of common bean cultivars: a review. Crop Sci. 41:1659–1675

Singh SP (2005) Common Bean (Phaseolus vulgaris L.). In: Singh RJ, Jauhar PP, editors. Genetic Resources, Chromosome Engineering, and Crop Improvement, Grain Legumes. London: CRC Press.

Singh SP (2007) Drought resistance in the race Durango dry bean landraces and cultivars. Agron. J. 99:1219–1225

Sponchiado BN, White JW, Castillo JA, Jones PG (1989) Root growth of four common bean cultivars in relation to drought tolerance in environments with contrasting soil types. Exp. Agric. 25:249-257.

Storey J (2002) A Direct Approach to False Discovery Rates on JSTOR. Wiley Online Libr 64:479–498. doi: 10.1111/1467-9868.00346

Trapp JJ, Urrea C a., Cregan PB, Miklas PN (2015) Quantitative Trait Loci for Yield under Multiple Stress and Drought Conditions in a Dry Bean Population. Crop Sci 55:1596. doi: 10.2135/cropsci2014.11.0792

Turner SD (2014)qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. biorXiv. 10.1101/005165.

Ueno K, Yoneyama H, Saito S, et al (2005) A lead compound for the development of ABA 8'-hydroxylase inhibitors. Bioorganic & Medicinal Chemistry Letters 15:5226-5229.

Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the Promising Fruits of Genomics: Applying Genome Sequencing Technologies to Crop Breeding. PLoS Biol 12:1–8. doi: 10.1371/journal.pbio.1001883

Vermeulen SJ, Campbell BM, Ingram JS (2012) Climate change and food systems. Annu Rev Environ Resour 37:195–222. doi: 10.1146/annurev-environ-020411-130608

Villordo-Pineda E, González-Chavira MM, Giraldo-Carbajo P, et al (2015) Identification of novel drought-tolerant-associated SNPs in common bean (*Phaseolus vulgaris*). Front Plant Sci 6:1–9. doi: 10.3389/fpls.2015.00546

Vlasova A, Capella-Gutiérrez S, Rendón-Anaya M, et al (2016) Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. Genome Biol 17:1–18. doi: 10.1186/s13059-016-0883-6

Wilkinson S, Davies WJ (1997) Xylem Sap pH Increase: A Drought Signal Received at the Apoplastic Face of the Guard Cell That Involves the Suppression of Saturable Abscisic Acid Uptake by the Epidermal Symplast. Plant Physiol 113:559–573.

Willing EM, Hoffmann M, Klein JD, et al (2011) Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics 27:2187–2193. doi: 10.1093/bioinformatics/btr346

Wright EM, Kelly JD (2011) Mapping QTL for seed yield and canning quality following processing of black bean (*Phaseolus vulgaris* L.). Euphytica 179:471–484. doi: 10.1007/s10681-011-0369-2

Wurtzel ET, Luo RB, Yatou O (2001) A simple approach to identify the first rice mutants blocked in carotenoid biosynthesis. J. Exp. Bot. 52:161–166.

Xavier A, Muir WM, Rainey KM (2016) Impact of imputation methods on the amount of genetic variation captured by a single-nucleotide polymorphism panel in soybeans. BMC Bioinformatics 17:55. doi: 10.1186/s12859-016-0899-7

Xiong H, Li J, Liu P, et al (2014) Overexpression of OsMYB48-1, a novel MYB-related transcription factor, enhances drought and salinity tolerance in rice. PLoS One 9:1–13. doi: 10.1371/journal.pone.0092913

Xu B, Gou JY, Li FG, et al (2013) A cotton BURP domain protein interacts with ??-expansin and their co-expression promotes plant growth and fruit production. Mol Plant 6:945–958. doi: 10.1093/mp/sss112

Yang DH, Kwak KJ, Kim MK, et al (2014) Expression of Arabidopsis glycine-rich RNA-binding protein AtGRP2 or AtGRP7 improves grain yield of rice (Oryza sativa) under drought stress conditions. Plant Sci 214:106–112. doi: 10.1016/j.plantsci.2013.10.006

Yang J, Benyamin B, Lund MS, et al (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Gen 42:565–569. doi: 10.1038/ng.608.Common

Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82. doi: 10.1016/j.ajhg.2010.11.011

Zeevaart JAD, Gage DA, Creelman RA (1990) Recent studies on the metabolism of abscisic acid. In: Pharis RP and Rood SB eds., Plant growth substances 1988, pp. 233-240. Springer-Verlag, Berlin, Germany.

Zhang J, Jia W, Yang J, Ismail AM (2006) Role of ABA in integrating plant responses to drought and salt stresses. F Crop Res 97:111–119. doi: 10.1016/j.fcr.2005.08.018

Zhang Y, Yang C, Li Y, Zheng N, Chen H, Zhao Q, Gao T, Guo H, Xie Q(2007) SDIR1 is a RING finger E3 ligase that positively regulates stress-responsive abscisic acid signaling in Arabidopsis. Plant Cell. 19:1912–1929.

Zhang Y, Feng S, Chen F, Chen H, Wang J, McCall C, Xiong Y, Deng XW (2008) Arabidopsis DDB1-CUL4 ASSOCIATED FACTOR1 forms a nuclear E3 ubiquitin ligase with DDB1 and CUL4 that is involved in multiple plant develop-mental processes. Plant Cell. 20:1437–1455.

Zhou Z, Jiang Y, Wang Z, et al (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33:408–14. doi: 10.1038/nbt.3096

Zimmermann IM, Heim MA, Weisshaar B, Uhrig JF (2004) Comprehensive identification of Arabidopsis thaliana MYB transcription factors interacting with R/B-like BHLH proteins. Plant J 40:22–34. doi: 10.1111/j.1365-313X.2004.02183.x

Zou M, Guan Y, Ren H, et al (2008) A bZIP transcription factor, OsABI5, is involved in rice fertility and stress tolerance. Plant Mol Biol 66:675–683. doi: 10.1007/s11103-008-9298-4

Zuiderveen GH, Padder BA, Kamfwa K, et al (2016) Genome-Wide association study of anthracnose resistance in andean beans (Phaseolus vulgaris). PLoS One 11:1–17. doi: 10.1371/journal.pone.0156391

**Material Suplementar 1 –** Identificação dos acessos de feijoeiro comum utilizados na análise de GWAS, pool gênico, tipo de germoplasma, instituição de origem e tipo de grão comercial.

# CONCLUSÕES GERAIS

As metodologias de genotipagem (DArTseq e Capture-Seq) utilizadas neste estudo permitiram a descoberta de milhares de SNPs com ampla distribuição no genoma de *P. vulgaris*, proporcionando elevada eficiência na caracterização do germoplasma estudado e na identificação de SNPs associados à tolerância à seca, com grande potencial para serem usados na rotina de seleção assistida por marcadores.

Os marcadores gerados pela metodologia DArTseq, descritos no primeiro capítulo, possibilitaram uma ampla caracterização molecular dos acessos mais diversos da CONFE, gerando conhecimento para utilização da diversidade genética existente e promovendo maior eficiência no uso deste germoplasma. Também foi possível selecionar um conjunto de SNPs para compor um painel de média densidade, capaz de amostrar os haplótipos neste germoplasma brasileiro e com potencial para ser utilizado, de modo rotineiro, para estudos diversos, tais como caracterização de bancos de germoplasma, análises de ligação e associação.

O segundo capítulo descreveu o desenvolvimento de um painel de associação contendo 343 acessos Mesoamericanos genotipados para 8789 SNPs, o qual poderá ser utilizado para análise de mapeamento associativo para diferentes caracteres de interesse dos programas de melhoramento genético de feijoeiro comum. Esse estudo também viabilizou a identificação de genótipos com desempenho superior para tolerância à seca, com a finalidade de serem utilizados como genitores em cruzamentos visando ampliar a base genética de linhagens do programa de melhoramento, além de propiciar a identificação de marcadores ligados às regiões genômicas associadas com a tolerância à seca, com grande potencial de serem incorporados à rotina de seleção assistida por marcadores, após etapas de validação da eficiência dos mesmos, ou utilizados como parte de um conjunto de marcadores altamente informativos para estudos de seleção genômica.

# ANEXOS

UNICAMP
COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br
IB

## DECLARAÇÃO

Em observância ao **§4º do Artigo 1º da Informação CCPG-UNICAMP/002/13**, de 14/08/2013, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Dissertação de Mestrado, intitulada "***Genômica Populacional e Análise de Associação Genômica Ampla (GWAS) para tolerância à seca e produtividade em feijoeiro comum (Phaseolus vulgaris)***", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: Paula Arielle M. R. Valdisser
Nome do(a) aluno(a): Paula Arielle Mendes Ribeiro Valdisser

Assinatura: MIZucchi
Nome do(a) orientador(a): Maria Imaculada Zucchi

Data: 22/04/2017

**Declaração**

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **GENÔMICA POPULACIONAL E ANÁLISE DE ASSOCIAÇÃO GENÔMICA AMPLA (GWAS) PARA TOLERÂNCIA À SECA E PRODUTIVIDADE EM FEIJOEIRO COMUM (Phaseolus vulgaris L.)**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 18/04/2017

Assinatura : Paula Arielle M. K. Valdisser
Nome do(a) autor(a): **Paula Arielle Mendes Ribeiro Valdisser**
RG n.° 4300905

Assinatura : MBZucchi
Nome do(a) orientador(a): **Maria Imaculada Zucchi**
RG n.°