Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Diversidade genética e associação genômica ampla em banco de germoplasma de
arroz (*Oryza sativa*)

Diane Simon Rozzetto

Tese apresentada para obtenção do título de Doutora em
Ciências. Área de concentração: Genética e
Melhoramento de Plantas

Piracicaba
2019

Diane Simon Rozzetto
Engenheira Agrônoma

Diversidade genética e associação genômica ampla em banco de germoplasma de arroz
(*Oryza sativa*)
versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **JOSÉ BALDIN PINHEIRO**

Tese apresentada para obtenção do título de Doutora em
Ciências. Área de concentração: Genética e
Melhoramento de Plantas

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Rozzetto, Diane Simon

Diversidade genética e associação genômica ampla em banco de Germoplasma de arroz (*Oryza sativa*) /Diane Simon Rozzetto- - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2019.

59 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. *Oryza sativa* 2. Marcadores SNP 3. Banco de germoplasma 4. Mapeamento associativo 5. Análise de associação I. Título

DEDICATÓRIA

À minha família

Dedico

AGRADECIMENTOS

À Deus pela vida.

Ao Prof. Dr. José Baldin Pinheiro pela orientação, exemplo profissional e amizade. À Prof^a. Dr^a Maria Imaculada Zucchi pelo apoio e amizade.

Aos Pesquisadores do Projeto Arroz Irrigado da EPAGRI, em especial ao Dr Rubens Marschalek a quem congratulo não apenas pela parceria, apoio e orientação mas por ter sido exemplo profissional e um dos grandes incentivadores à minha dedicação a pesquisa. Aos assistentes de pesquisa Samuel Batista dos Santos e Geovani Porto que não mediram esforços possibilitando excelência na condução dos experimentos, e mais do que isso, pela amizade e confiança.

À Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ/USP) e ao programa de Pós-graduação em Genética e Melhoramento de Plantas do Departamento de Genética pela oportunidade de realizar o curso de Doutorado;

Aos colegas e amigos do Laboratório de Diversidade Genética e Melhoramento, ao Tiago, Máisa, Vanessa, João Paulo, Emanuel e Maiara pelo auxílio nos experimentos, análises, discussões e especialmente pela convivência, amizade, troca de ideias.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão da bolsa de estudos e apoio financeiro.

À Professora Lucía Gutiérrez e à Universidade de Wisconsin - Madison, que possibilitaram incontestável contribuição a este projeto de doutoramento, enriquecendo o desenvolvimento da pesquisa por meio de internacionalização e oportunidade de realizar intercâmbio no exterior. À Juliana Leite e à família Hausler (Bill, Susan, Lisa Anne, Charlie e Mia) que me acolheu como parte dela, proporcionando-me mais que uma casa, um lar. Que me deu suporte para permanecer nos EUA durante o intercâmbio, todos os dias me lembro com carinho e profundas saudades. “My dear American Family, I Love You”.

Em especial aos meus pais a quem empenho grande honra, por me ensinarem o que é o amor e por terem sempre me proporcionado o seu melhor; ao meu irmão a quem muito admiro e a toda a minha família com quem sempre pude contar e que compreendeu minha ausência. Ao meu noivo Pedro Henrique, por todo o apoio, confiança e incentivo. À todas as pessoas que de alguma forma participaram desta caminhada e que estarão sempre em meu coração.

Muito obrigada!

EPÍGRAFE

“Porque dEle, por Ele, e para Ele, são todas as coisas”

Romanos 11:36

SUMÁRIO

RESUMO.....	7
ABSTRACT	8
1. INTRODUÇÃO.....	9
REFERÊNCIAS.....	10
2. DIVERSIDADE GENÉTICA E ESTRUTURA POPULACIONAL EM BANCO DE GERMOPLASMA DE ARROZ UTILIZANDO MARCADORES DART-SEQ	13
Resumo.....	13
Abstract.....	13
2.1. INTRODUÇÃO	14
2.2. MATERIAL E MÉTODOS.....	15
2.2.1. Material vegetal:.....	16
2.2.2. Extração de DNA:	16
2.2.3. Genotipagem:	16
2.2.4. Filtragem dos dados genotípicos:	16
2.2.5. Diversidade genética da população:	16
2.2.6. Estrutura genética da população:.....	17
2.2.7. Análise de agrupamento:	17
2.3. RESULTADOS E DISCUSSÃO.....	18
2.3.1. Marcadores	18
2.3.2. Diversidade genética	18
2.3.3. Estruturação genética	18
2.3.4. Agrupamento.....	22
2.4. CONCLUSÃO.....	24
REFERÊNCIAS.....	25
3. ASSOCIAÇÃO GENÔMICA AMPLA PARA CARACTERES DE INTERESSE AGRONÔMICO EM ARROZ UTILIZANDO MARCADORES DART-SEQ.....	27
Resumo.....	28
Abstract.....	28
3.1. INTRODUÇÃO	28
3.2. MATERIAL E MÉTODOS	29
3.2.1. Material vegetal:.....	29
3.2.2. Fenotipagem:	29
3.2.3. Extração de DNA:	30
3.2.4. Genotipagem DArTseq:	30
3.2.5. Filtragem dos dados	31
3.2.6. Estrutura da população, matriz de parentesco e desequilíbrio de ligação:	31
3.2.7. Mapeamento associativo:	31
3.2.8. Anotação de genes:.....	31
3.3. RESULTADOS E DISCUSSÃO.....	32
3.3.1. Matriz de parentesco, estrutura populacional e desequilíbrio de ligação	33
Matriz de Parentesco	33
Estrutura Populacional	34
Desequilíbrio de Ligação	35
3.3.2. Mapeamento associativo	36
3.3.3. Anotação de genes categorização funcional	40
3.4. CONCLUSÃO.....	50
REFERÊNCIAS.....	50
ANEXO.....	53

RESUMO

Diversidade genética e associação genômica ampla em banco de germoplasma de arroz (*Oryza sativa*)

O Departamento de Genética da Escola Superior de Agricultura “Luiz de Queiroz” – ESALQ/USP possui um banco de germoplasma de arroz com aproximadamente 460 acessos, oriundos de diferentes instituições de pesquisa do Brasil e do mundo. Desses, cerca de 190 acessos pertenciam ao Japão, 142 são de origem Filipina e os demais são variedades crioulas e cultivares brasileiras. Tais acessos estão passando por uma extensa pesquisa científica afim de obter informações a respeito de sua origem, diversidade e estruturação genética. Este trabalho teve como objetivo estudar a diversidade genética e fazer uma análise de associação genômica ampla (GWAS) de características diversas dos acessos, a fim de conhecer a diversidade e estrutura genética dos mesmos e as regiões genômicas responsáveis pelo controle de características que poderiam ser utilizadas em programas de melhoramento. A genotipagem foi realizada utilizando marcadores SNPs (Single Nucleotide Polimorfism) gerados por meio da tecnologia DArTseq (Diversity Arrays Sequence) de genotipagem por sequenciamento (GBS). A Diversidade genética da população foi analisada por meio do pacote “hierfstat”, implementado no Software R, onde foram utilizados 269 acessos (os acessos de origem Filipina e Brasileira). Para as análises de associação foram utilizados os 142 acessos filipinos. A instalação dos experimentos de caracterização fenotípica foi realizada em áreas de cultivo de sequeiro, na área experimental do Departamento de Genética da ESALQ/USP em dois anos agrícolas. O delineamento experimental foi de blocos aumentados, sendo realizados os tratos culturais recomendados para a cultura. Os acessos foram caracterizados de acordo com descritores indicados pelo IBPGR – IRRI (Rice Advisory Committee). Os caracteres avaliados foram: Número de dias para a emergência (NDE); Número de dias para o florescimento (NDF); Ciclo total da planta (CTP); Massa de cem sementes (MCS); Produtividade de grãos (PG); Comprimento do colmo (CC); Número de Perfilhos (NP); Comprimento e Largura da folha bandeira (CFB e LFB); Altura de planta na maturidade (APM); Comprimento da Panícula (CP); Comprimento e largura da espiguetas (CE e LE); Número de ramificações por panícula (NRP). Os dados coletados foram analisados através do software R (R DEVELOPMENT CORE TEAM, 2018). O Banco de Germoplasma de arroz da ESALQ/USP é uma potencial e importante fonte de acessos que podem ser incorporados aos programas de melhoramento, haja vista, o número de acessos e a diversidade genética existente. Não foram identificadas duplicatas entre os acessos avaliados, sendo assim justificável manter-se todos eles na coleção nuclear do Banco. A Análise Discriminante dos Componentes Principais (DAPC) sugeriu a formação de 5 subgrupos genéticos principais na população de estudo de diversidade genética, já a população estudada para análise de associação pode ser subdividida em 3 grupos em função da análise de componentes principais e foram identificados 33 SNP’s associados a oito caracteres quantitativos avaliados nos acessos. A tecnologia de genotipagem por sequenciamento como a DarT-seq mostrou grande eficiência no estudo de diversidade genética e no estudo de associação genômica ampla nos genótipos de arroz testados.

Palavras-chave: *Oryza sativa*; Marcadores SNP; Banco de germoplasma; Mapeamento associativo; Análise de associação

ABSTRACT

Genetic diversity and genome-wide association in rice (*Oryza sativa*) germplasm bank

The Department of Genetics of the "Luiz de Queiroz" College of Agriculture - ESALQ / USP has a rice germplasm bank with approximately 460 accessions, coming from different research institutions in Brazil and the world. Of these, about 190 accessions belonged to Japan, 142 are of Filipino origin and the others are Creole varieties and Brazilian cultivars. Such accesses are undergoing extensive scientific research in order to obtain information regarding their origin, diversity and genetic structuring. The objective of this work was to study genetic diversity and to perform an analysis of broad genomic association (GWAS) of different characteristics of the accessions, in order to know the diversity and genetic structure of the same and to know genomic regions responsible for the control of characteristics that could be used in breeding programs. Genotyping was performed using single nucleotide polymorphisms (SNPs) generated by sequencing genotyping (GBS) DArTseq (Diversity Arrays Sequence) technology. The genetic diversity of the population was analyzed through the "hierfstat" package, implemented in Software R, where 269 accessions (the accesses of Philippine and Brazilian origin) were used. For the analysis of association, the 142 Filipino accessions were used. The establishment of the phenotypic characterization experiments was carried out in rainfed areas, at the experimental farm of the Department of Genetics of ESALQ / USP, in two agricultural years. The experimental design was of increased blocks, being the cultural treatments recommended for the culture. The accesses were characterized according to descriptors indicated by IBPGR - IRRI (Rice Advisory Committee). The evaluated characters were: Number of days for the emergency (NDE); Number of days for flowering (NDF); Total plant cycle (CTP); One hundred seed mass (MCS); Grain yield (PG); Height of high (CC); Number of Profiles (NP); Length and width of the flag leaf (CFB and LFB); Height of plant at maturity (APM); Panicle Length (CP); Length and width of the spikelet (CE and LE); Number of branches per panicle (NRP). The collected data were analyzed through the software R (R DEVELOPMENT CORE TEAM, 2018). The Rice Germplasm Bank of ESALQ / USP is a potential and important source of access that can be incorporated into breeding programs, presence, number of hits and a possible genetic model. No duplicates were used between the accesses evaluated and thus justified. The Principal Components Discriminant Analysis (DAPC) suggested the formation of 5 major genetic subgroups in its genetic study series, already the population studied for analysis of association can be subdivided into 3 groups as a function of the analysis of main components and 33 SNP's associated to eight quantitative traits evaluated in the accessions were identified. Sequencing genotyping technology such as DArT-seq showed great efficiency in the study of genetic diversity and in the study of broad genomic association in the tested rice genotypes.

Keywords: *Oryza sativa*; SNP markers; Germplasm bank; Associative mapping; Association analysis

1. INTRODUÇÃO

O arroz é uma das principais culturas alimentares em todo o mundo. Dentre os cereais cultivados, é um dos que melhor se adapta a diferentes condições geográficas, climáticas e ecológicas (KUMBHAR et al. 2015; MALATHI et al. 2016). Devido a isso, existe um grande número de acessos disponíveis com ampla variabilidade genética e fenotípica. Embora existam importantes fontes de genes conservadas em bancos genéticos globais, percebe-se uma grande similaridade genética entre cultivares comerciais. Isso se deve ao fato de que tem-se utilizado uma pequena quantidade de recursos resultando numa estreita base genética (GARRIS; MCCOUCH; KRESOVICH, 2005; BRAR; KHUSH, 1997). Ganhos genéticos tem sido registrados anualmente graças a forte atuação dos programas de melhoramento, que associados às boas práticas de manejo resultam em elevadas produções e ajudam a garantir segurança alimentar. Por outro lado, percebe-se que estes ganhos tem sido menores com o passar dos anos, tendendo a uma estagnação dos patamares produtivos. Ao longo de décadas, os processos de seleção dirigida para determinadas características podem ter levado ao “esquecimento” ou até mesmo a perda de determinados genes, fato que vem ocorrendo também em outras culturas. Neste sentido, os programas de melhoramento de arroz no Brasil e no mundo enfrentam o desafio de superar patamares de produtividade e além disso precisam levar em conta outros caracteres importantes como a tolerância a estresses bióticos e abióticos e a qualidade de grãos, buscando dessa forma não só trazer bons rendimentos como também atender diferentes nichos de mercado (DEMONT et al. 2013; ASGHAR et al. 2012).

A magnitude da variabilidade genética é muitas vezes determinante para o sucesso de qualquer programa de melhoramento (VANAJA; LUCKINS, 2006). Além do mais, destaca-se a importância de ampliar a base genética dos genótipos atualmente cultivados, introduzindo novos alelos para caracteres de importância agrônômica. Bancos de Germoplasma são tidos como as principais fontes de variabilidade para muitos programas de melhoramento vegetal podendo ser um reservatório de alelos. Sendo assim, esforços em trabalhos de pré-melhoramento são imprescindíveis, visto que por meio deste é possível conhecer a variabilidade genética dos acessos armazenados e conhecer sistematicamente o genoma entendendo as funções dos genes (GOWDA et al. 2003).

Deve-se considerar que a maioria dos caracteres de interesse agrônômico são complexos, chamados de caracteres quantitativos por serem controlados por diversos genes. Regiões do genoma que apresentam genes associados a tais caracteres são chamados de QTLs (derivados do inglês “*Quantitative trait loci*” ou locos de caracteres quantitativos). Desde 1980, tem-se empregado grandes esforços para caracterizar essas regiões no genoma por meio do uso de marcadores moleculares uma vez que não seria possível avaliá-los apenas com base em caracterizações fenotípicas. Essas “marcas” identificadas em nível molecular representam diferenças genéticas entre organismos ou espécies servindo como “etiqueta de genes” e em geral não afetam o fenótipo da característica, pois estão localizadas próximas aos respectivos genes (COLLARD et al. 2005).

Em arroz, o uso de marcadores moleculares tem permitido estudos de diversidade, identificação de características importantes, e a obtenção de mapas saturados (ZHU et al. 1998; CAUSSE et al. 1994; PANTALIÃO et al. 2016). Com o avanço nas tecnologias de sequenciamento de nova geração (NGS), as técnicas de genotipagem por sequenciamento (GBS) surgiram como uma abordagem promissora. Por meio delas, foi possível explorar ainda mais a diversidade genética de plantas e resultaram na descoberta de grande número de marcadores moleculares (POLAND; RIFE, 2012; Elshire et al. 2011). Neste sentido, o GBS vem sendo utilizado para descobrir marcadores de SNPs (*Single-nucleotide polymorphisms* ou polimorfismos em bases únicas), identificar marcadores fortemente ligados a QTLs

(BOUTET et al. 2016; TANG et al. 2016) bem como na seleção genômica em programas de melhoramento de plantas (FURUTA et al. 2017).

Com os avanços nas tecnologias de marcadores baseados em microarranjos, surgiram os marcadores DArT (Diversity Arrays Technology), que devido ao seu baixo custo e eficiências, tem sido muito utilizados para a construção de mapas de alta densidade, mapeamento de QTLs e análises de diversidade genética (GUPTA; RUSTGI; MIR, 2008). Em seguida, combinou-se a redução da complexidade associada ao DArT, com técnicas de sequenciamento de nova geração (NGS), desenvolvendo-se uma plataforma de alto rendimento, o DArTseq resultando em uma nova abordagem de sequenciamento (QIU et al. 2012). A principal diferença desta para outras tecnologias de genotipagem, refere-se a como é feita a redução da complexidade da molécula de DNA. Ao obter-se o alinhamento das sequências, são detectados dois tipos de polimorfismos: i) DArT-Insilico, resultado da variabilidade de distribuição dos sítios de restrição entre as amostras sendo observado na forma de presença (1) ou ausência (2) de determinado fragmento de DNA ou “tag”; ii) SNPs - detecta polimorfismos resultantes da variação de bases únicas ao longo das sequências. A detecção dos SNPs se dá a partir do alinhamento de “reads” de cada “tag” em comum entre as amostras. São então formados grupos de “reads”, e seleciona-se a variante alélica mais abundante em cada grupo sendo esta tida como referência e a variante menos abundante é considerada um SNP (SANSALONI; GRATAPAGLIA, 2012).

O mapeamento por associação tem sido de grande importância pois leva em conta dados fenotípicos e genotípicos e é uma importante ferramenta na identificação de marcadores moleculares ligados à características de interesse podendo ser utilizado na seleção assistida por marcadores. Essa análise tem permitido a utilização de diversos acessos pois fornece uma cobertura alélica ampla não sendo necessário desenvolver populações bi-parentais para os estudos (Tadesse et al. 2015). Marcadores DArTseq vem sendo aplicados para o mapeamento de QTL, estudos de associação genômica ampla (GWAS) e estudos de diversidade genética (Sanchez-Sevilla et al. 2015). Com a disponibilidade pública do sequenciamento do genoma do arroz é possível usar recursos genéticos existentes para a descoberta de genes e alelos com base em estratégias de mapeamento associativo (LU et al. 2005).

O Departamento de Genética da ESALQ/USP, (Escola Superior de Agricultura “Luiz de Queiroz”/Universidade de São Paulo) possui um Banco de Germoplasma de Arroz oriundo de diversas instituições de pesquisa. Nele estão armazenados cerca de 460 acessos mantidos pelo Laboratório de Diversidade Genética e Melhoramento de Plantas. A fim de conhecer a diversidade e estrutura genética desses acessos e conhecer regiões genômicas responsáveis pelo controle de características que poderiam ser utilizadas em programas de melhoramento, este trabalho teve como objetivo estudar a diversidade genética e fazer uma análise de associação genômica ampla (GWAS) de características diversas dos acessos.

Referências

- ASGHAR S, et al. Cooking and eating characteristics or rice (*Oryza sativa* L.). **Pakistan Journal of Food Sciences**, v. 22, n. 3, p. 128–132. 2012.
- BRAR, D; KHUSH, G. Alien gene introgression in rice. **Plant Molecular Biology**, v. 35, n. 1, p. 35–47, 1997.
- CAUSSE, M. A. et al. Saturated molecular map of the rice genome based on an interspecific backcross population. **Genetics**, v. 138, n. 4, p. 1251–1274, 1994.
- COLLARD, B. C. Y. et al. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. **Euphytica**, v. 142, n. 1–2, p. 169–196, 2005.
- DEMONT M. Reversing urban bias in African rice markets: A review of 19 National Rice Development Strategies. **Global Food Security**. v. 2, n. 3 p. 172–181, 2013.

- ELSHIRE R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **Plos One**. v. 6, e19379. 2011.
- FURUTA, T. et al. Adapting Genotyping-by-Sequencing for Rice F2 Populations. G3: Genes | Genomes, **Genetics**. 2017.
- GARRIS, A. S.; MCCOUCH, S.; KRESOVICH, S. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.). **Genetics**, v. 165, n. 2, p. 759–769, 2005.
- GOWDA, M. et al. Advances in rice breeding, genetics and genomics. **Molecular Breeding**, v. 11, n. 5, p. 337–352, 2003.
- GUPTA, P.; RUSTGI, S.; MIR, R. Array-based high-throughput DNA markers for crop improvement. **Heredity**. v. 101, p. 5–18, 2008
- KUMBHAR, S. D. et al. Genetic diversity and population structure in landraces and improved rice varieties from India. **Rice Science**, v. 22, n. 3, p. 99–107, 2015.
- MALATHI S, et al. Genetic characterization and population structure of Indian rice cultivars and wild genotypes using core set markers. **Biotechnology**, v. 6, n. 1, p. 95, 2016.
- PANTALIÃO, G. F. et al. Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. **Genetica**, v. 144, n. 6, p. 651–664, 2016.
- SANSALONI, J. A.; RIFE, T. W. Genotyping-by-sequencing for plant breeding and genetics. **The Plant Genome**. v. 5, p. 92–102, 2012.
- QIU X. et al. Mapping and characterization of the major quantitative trait locus qSS7 associated with increased length and decreased width of rice seeds. **Theoretical and Applied Genetics**, v. 125, p. 1717–1726, 2012.
- SANCHEZ-SEVILLA J. F. et al. Diversity Arrays Technology (DArT) marker platforms for diversity analysis and linkage mapping in a complex crop, the octoploid cultivated strawberry (*Fragaria*×*ananassa*). **Plos One**. v. 10, e0144960, 2015.
- SANSALONI, C. P.; GRATTAPAGLIA, O. D. **Desenvolvimento e aplicações de DArT Diversity Arrays Technology) e genotipagem por sequenciamento (Genotyping-by- Sequencing) para análise genética em Eucalyptus**. [s.l.] Universidade de Brasília, 2012.
- TADESSE, W. Genome-wide association mapping of yield and grain quality traits in winter wheat genotypes. **Plos One**. v. 10: e0141339, 2015.
- VANAJA, T. LUCKINS, T. B. Variability in grain quality attributes of high yielding rice varieties (*Oryza sativa* L.) of diverse origin. **Journal of Tropical Agriculture**, v. 44, n. 1–2, p. 61–63, 2006.
- ZHU, J. et al. AFLP markers for the study of rice biodiversity.pdf. **Theoretical and Applied Genetics**, v. 96, n. 7, p. 602–611, 1998.

2. DIVERSIDADE GENÉTICA E ESTRUTURA POPULACIONAL EM BANCO DE GERMOPLASMA DE ARROZ UTILIZANDO MARCADORES DART-SEQ

Resumo

O Departamento de Genética da Escola Superior de Agricultura “Luiz de Queiroz” – ESALQ/USP possui um banco de germoplasma de arroz com aproximadamente 460 acessos, oriundos de diferentes instituições de pesquisa do Brasil e do mundo. Desses, cerca de 190 acessos pertenciam ao Japão, 142 são de origem Filipina e os demais são variedades crioulas e cultivares brasileiras. Tais acessos estão passando por uma extensa pesquisa científica afim de obter informações a respeito de sua origem, diversidade e estruturação genética. Este trabalho teve como objetivo avaliar a diversidade e estruturação genética em um painel contendo um total de 269 acessos pertencentes ao Banco de Germoplasma de Arroz do Departamento de Genética da ESALQ. A genotipagem foi realizada utilizando marcadores SNPs (Single Nucleotide Polimorfism) gerados por meio da tecnologia DArTseq (Diversity Arrays Sequence) de genotipagem por sequenciamento (GBS). A Diversidade genética da população foi analisada por meio do pacote “hierfstat”, implementado no Software R. A estrutura da população foi estudada por meio de uma Análise Discriminante de Componentes Principais através do pacote “adegenet”. Foi realizada uma análise de agrupamento utilizando marcadores SNPs, Para isso fez-se o cálculo da matriz de distâncias de Rogers (1972) modificada por Wright (1978). Com base nas distâncias calculadas, foi construído um dendrograma pelo método UPGMA. A genotipagem dos 269 acessos utilizando o método de sequenciamento DArT-seq, produziu aproximadamente 35.000 SNPs. Após um processo de filtragem restaram 11.190 SNPs de alta qualidade que foram utilizados nas análises seguintes. Como não havia subpopulações pré-definidas procedeu-se uma análise de estruturação, visto que esta informação é de grande importância em estudos de estrutura genética populacional. Tal análise sugeriu a formação 5 subgrupos genéticos principais na população. O Banco de Germoplasma de arroz da ESALQ/USP é uma potencial e importante fonte de acessos que podem ser incorporados aos programas de melhoramento, haja vista, o número de acessos e a diversidade genética existente. Não foram identificadas duplicatas entre os acessos avaliados, sendo assim justificável manter-se todos eles na coleção nuclear do Banco. A tecnologia de genotipagem por sequenciamento como a DarT-seq mostrou grande eficiência no estudo de diversidade genética entre tais acessos.

Palavras-chave: *Oryza sativa*; Germoplasma; Marcadores SNP; DArT-seq; DAPC

Abstract

The Department of Genetics of the "Luiz de Queiroz" College of Agriculture - ESALQ / USP has a rice germplasm bank with approximately 460 accessions, coming from different research institutions in Brazil and the world. Of these, about 190 accessions belonged to Japan, 142 are of Filipino origin and the others are Creole varieties and Brazilian cultivars. Such accesses are undergoing extensive scientific research in order to obtain information regarding their origin, diversity and genetic structuring. This work aimed to evaluate the genetic diversity and structure in a panel containing a total of 269 accessions belonging to the Rice Germplasm Bank of the Department of Genetics of ESALQ. Genotyping was performed using single nucleotide polymorphisms (SNPs) generated by sequencing genotyping (GBS) DArTseq (Diversity Arrays Sequence) technology. Genetic diversity of the population was analyzed through the "hierfstat" package, implemented in Software R. The population structure was studied through a Discriminant Analysis of Principal Components through the "adegenet" package. A cluster analysis was performed using SNPs markers. For this purpose, the Rogers (1972) matrix of distances was modified by Wright (1978). Based on the calculated distances, a dendrogram was constructed using the UPGMA method. Genotyping of the 269 accesses using the DArT-seq sequencing method yielded approximately 35,000 SNPs. Following a filtering process, 11,190 high-quality SNPs were used which were used in the following analyzes. As there were no pre-defined

subpopulations, a structural analysis was carried out, since this information is of great importance in population genetic studies. Such analysis suggested the formation of 5 major genetic subgroups in the population. The Rice Germplasm Bank of ESALQ / USP is a potential and important source of access that can be incorporated into breeding programs, given the number of accesses and the genetic diversity that exists. No duplicates were identified among the accesses evaluated, so it is justifiable to keep all of them in the Bank's nuclear collection. Sequencing genotyping technology such as DarT-seq showed great efficiency in the study of genetic diversity between such accesses.

Keywords: *Oryza sativa*; Germplasm; SNP markers; DarT-seq; DAPC

2.1 Introdução

Popularmente chamado de grão global, o arroz (*Oryza sativa* L.) é um cereal de grande importância no mundo todo. A cultura desempenha papel estratégico em questões de segurança alimentar, sendo base nutricional para grande parte da população (Kumbhar et al. 2015). É uma das mais diversas espécies cultivadas, sendo capaz de adaptar-se a uma ampla gama de ambientes (CHANG; BUNTING, 1976).

O gênero *Oryza* surgiu há mais de 130 milhões de anos disseminando-se pelo continente Gondwana como uma gramínea silvestre e suas espécies estão distribuídas em todos os continentes à exceção da Antártida. Sua domesticação teve início há cerca de nove mil anos na Ásia, onde era cultivado em solos sem inundação. Posteriormente os processos de alagamento e transplântio de mudas foram aperfeiçoados na China, tornando este cereal plenamente domesticado (KHUSH, 1997). A espécie *O. sativa* apresenta duas subespécies principais *Indica* e *Japonica*. Embora não haja consenso sobre a origem de cada uma, em termos ecogeográficos as variedades *Indica* são conhecidas como arroz cultivado na Ásia Tropical com irrigação por alagamento e as variedades *Japonica* em regiões de clima temperado, elevadas altitudes e áreas de sequeiro (GARRIS; MCCOUCH; KRESOVICH, 2005).

Os processos de cultivo e domesticação, aliados à atuação dos agricultores em selecionar as melhores plantas durante séculos, sob diferentes condições, resultou num grande número de variedades de arroz. Deste longo processo de seleção e ambientação, resultaram mudanças morfológicas e fisiológicas relacionadas ao seu desenvolvimento tanto vegetativo quanto reprodutivo (por exemplo, identificou-se encurtamento no caule devido a uma mutação no gene “semi anão 1 - SD1” - tais mutações os tornaram mais robustos, aumentando a capacidade de produção de grãos da planta) além da redução ou perda de algumas características consideradas indesejáveis ao cultivo, como dormência de sementes, degrana da panícula e a presença de aristas bem como a frequência de polinização cruzada elevando sua taxa de autogamia (CHANG; BUNTING, 1976). Além disso, existe grande diferença no ciclo de desenvolvimento que pode variar de 80 a 280 dias, bem como o tamanho, qualidade e composição dos grãos. A variação também se dá com relação à tolerância a estresses bióticos e abióticos. Estima-se que essa grande variabilidade resulte em mais de 120 mil variedades de arroz armazenadas em bancos de Germoplasma em todo o mundo (KHUSH, 1997).

A forte atuação dos programas de melhoramento vem possibilitando a obtenção de ganhos genéticos que aliados às novas tecnologias de manejo resultam em incremento na produtividade e produção de arroz. Anualmente, são lançadas novas cultivares que oferecem, dentre os principais atributos, uma maior produtividade, melhor qualidade de grãos, tolerância a estresses abióticos e resistência a pragas e doenças. Apesar disso, sabe-se que atualmente a base genética existente nas cultivares comerciais é estreita. Isso se deve ao uso intensivo de variedades modernas, que com o objetivo de preservar as melhores combinações gênicas foi priorizado um grupo restrito de genitores-elite e consequentemente o conjunto gênico (PERVAIZ et al. 2009)

A ampliação da base genética permitiria o surgimento de novas combinações alélicas, proporcionando uma redução da vulnerabilidade a pragas e doenças bem como maior adaptação a ambientes específicos. Além do mais, é uma das principais alternativas para quebrar o atual patamar de produtividade do arroz, resultando em incremento e estabilidade de produção. Isso é possível por meio da introdução de variabilidade genética adicional às populações de melhoramento por meio do uso de parentais mais divergentes oriundos de Bancos de Germoplasma (BRONDANI et al. 2003).

Neste sentido, destaca-se a importância da manutenção e caracterização dos Bancos de Germoplasma, uma vez que um dos maiores desafios da pesquisa é identificar genótipos potenciais, e por meio de cruzamentos amplos sejam resgatados alelos raros para serem incorporados no germoplasma-elite, favorecendo a ocorrência de novas combinações alélicas. A pesquisa genômica, tem gerado novas técnicas resultando em maior eficiência e precisão do melhoramento como a utilização de marcadores moleculares. Sendo assim é possível conhecer a extensão da variabilidade dos recursos genéticos das espécies vegetais de maior interesse econômico (VARSHNEY; GRANER; SORRELLS, 2005).

A utilização de ferramentas biotecnológicas, associadas a métodos convencionais de melhoramento, vem ganhando destaque e tem permitido explorar a variabilidade genética encontrada em acessos dessa cultura. Através do uso da tecnologia de marcadores moleculares, tem-se atingido aprimoramento das avaliações, com medidas mais precisas da variabilidade genética existente entre e dentro dos acessos armazenados, o que permite detectar diferenças entre indivíduos em nível de DNA (BRONDANI et al. 2006). Tanto em pesquisas populacionais relacionadas à conservação quanto no melhoramento genético propriamente dito, uma das técnicas amplamente utilizadas, tem sido a genotipagem por sequenciamento (do inglês Genotyping by Sequence - GBS). Esta técnica de sequenciamento permite a identificação de diversos SNPs (Single Nucleotide Sequences) possibilitando estudo de diversidade genética e compreensão da estrutura populacional, sem conhecimento prévio do genoma estudado e/ou sua diversidade (ELSHIRE et al. 2011).

O Departamento de Genética da Escola Superior de Agricultura “Luiz de Queiroz” – ESALQ/USP possui um banco de germoplasma de arroz com aproximadamente 460 acessos, oriundos de diferentes instituições de pesquisa do Brasil e do mundo. Desses, cerca de 190 acessos pertenciam ao Instituto de Arroz de Sequeiro do Japão, 142 são de origem Filipina e os demais são variedades crioulas e cultivares brasileiras. Tais acessos estão passando por uma extensa pesquisa científica afim de obter informações a respeito de sua origem, diversidade e estruturação genética. Diante da importância econômica que a cultura do arroz assume mundialmente, este trabalho teve como objetivo avaliar a diversidade e estruturação genética em um painel contendo um total de 269 acessos pertencentes ao Banco de Germoplasma de Arroz do Departamento de Genética da ESALQ e foi genotipado com SNP's utilizando a tecnologia de genotipagem por sequenciamento DArT-seq.

2.2 Material e métodos

2.2.1 Material vegetal:

Foi utilizado um painel com 269 acessos de arroz (dentre eles, 142 acessos são de origem filipina, 117 variedades brasileiras e 10 testemunhas comerciais amplamente cultivadas no Brasil), pertencentes ao Banco de Germoplasma de Arroz (Anexo1). Tais recursos genéticos são mantidos pela Escola Superior de Agricultura “Luiz de Queiroz” – USP de Piracicaba, São Paulo.

2.2.2 Extração de DNA:

A extração de DNA genômico, foi realizada a partir de folhas jovens coletadas de uma planta dos respectivos acessos por meio do método de extração CTAB descrito por Doyle; Doyle (1987), modificado. A fim de verificar a qualidade e quantidade, o DNA foi diluído em solução TE (Tris-HCl 10 mM, EDTA 1 mM, com pH 8,0) em volumes de 50µl e as amostras foram aplicadas em gel de agarose a 1%. A integridade do DNA foi verificada pela identificação de apenas uma banda com elevado peso molecular sobre o gel, sendo que a concentração foi estimada com base em marcadores com peso molecular de DNA conhecido de fago λ.

2.2.3 Genotipagem:

O DNA dos acessos foi remetido à empresa Diversity Arrays Technology Pty Ltd. (DArT P/L) na Austrália onde foram sequenciados por meio da tecnologia de genotipagem DArTseq™, descrito por (Courtois et al. 2013). Em síntese, tal metodologia combina “Genotyping by Sequencing”, (GBS) com Diversity Arrays Technology (DArT) que envolve a redução de complexidade do genoma utilizando enzimas de restrição *PstI/TaqI* e em seguida o sequenciamento de sequências curtas através pelo Illumina HiSeq 2500 (Maiores informações podem ser obtidas em <<http://www.diversityarrays.com/dart-application-dartseq-data-types>>).

2.2.4 Filtragem dos dados genotípicos:

De modo a obter estimativas fidedignas, foi realizada a seleção dos SNPs com a melhor qualidade. Neste sentido, fez-se a filtragem dos dados levando em conta alguns parâmetros: *i)* Reprodutibilidade ($Rep\ Avg \geq 0,95$; *ii)* $Call\ Rate \geq 0,9$; *iii)* Conteúdo de informação polimórfica ($PIC \geq 0,1$).

2.2.5 Diversidade genética da população:

A diversidade genética da população foi analisada por meio do pacote “hierfstat”, implementado no Software R (Jerome; Jombart; Goudet, 2015); (R DEVELOPMENT CORE TEAM, 2018). Os parâmetros estimados para cada grupo dentro da população foram:

Heterozigosidade média observada (H_O): Obtida pela soma dos valores de cada loco dividindo-se pelo número total de locos estudados (Equação 1) (WEIR, 1983):

$$H_o = 1 - \sum P_{ii}$$

(Equação 1)

Onde, P_{ii} - a frequência observada de genótipos homozigotos do alelo;

Heterozigosidade média esperada (H_s): Baseada em NEI; Tajima; Tatenó (1978), sendo P_i - a frequência estimada do i ésimio alelo. A heterozigosidade média esperada foi obtida pela média aritmética de todos os locos estudados:

$$H_s = 1 - \sum P_i^2$$

(Equação 2)

Onde - P_i a frequência estimada do i ésimio alelo

Índice de Fixação de Wright (F_{IS}) ou Coeficiente de Endogamia, estimado com base na Heterozigosidade média observada (H_o) e esperada (H_s), (Equação 3):

$$F_{IS} = 1 - \frac{H_o}{H_s}$$

(Equação 3)

Riqueza Alélica (AR): afim de identificar, o número de alelos diferentes segregando em na população.

2.2.6 Estrutura genética da população:

A estrutura da população foi estudada por meio de uma Análise Discriminante de Componentes Principais (DAPC) (JOMBART; DEVILLARD; BALLOUX, 2010) através do pacote *adegenet* (Jombart, 2008) no R (R DEVELOPMENT CORE TEAM, 2018). A DAPC consiste em uma análise exploratória, onde por meio de um subconjunto de marcadores SNPs distribuídos ao longo do genoma, ajuda na determinação do número mais provável de sub-populações dentro da população estudada.

2.2.7 Análise de agrupamento:

Foi realizada uma análise de agrupamento utilizando marcadores SNPs. Para isso fez-se o cálculo da matriz de distâncias de Rogers (1972) modificada por Wright (1978). Com base nas distâncias calculadas, foi construído um dendrograma pelo critério UPGMA (Unweighted Pair Group Method with Arithmetic Mean) (SOKAL; MICHENER, 1958). Este é um método de agrupamento hierárquico que faz uso da média das distâncias de todos os genótipos para formação de cada grupo. A consistência dos dendrogramas foi verificada através do valor do coeficiente de correlação cofenética (CCC) e a consistência dos nós foi verificada com base em 1000 reamostragens “bootstrap” a fim de verificar a estabilidade dos agrupamentos formados. Todas as análises foram realizadas na plataforma R (R DEVELOPMENT CORE TEAM, 2018).

2.3 Resultados e Discussão

2.3.1 Marcadores

A genotipagem dos 269 acessos utilizando o método de sequenciamento DArT-seq, produziu aproximadamente 35.000 SNPs. Após um processo de filtragem restaram 11.190 SNPs de alta qualidade que foram utilizados nas análises seguintes, resultando em 22.380 alelos e o tamanho do genoma de 27,9Mb.

Como não havia subpopulações pré-definidas procedeu-se uma análise de estruturação, visto que esta informação é de grande importância em estudos de estrutura genética populacional.

2.3.2 Estruturação genética

Uma vez que não se tem informação de populações pré-definidas, é importante verificar se existe subestruturação, dada a sua importância em estudos de estrutura genética populacional. Assim sendo, as figuras 1, 2 e 3 apresentam o resultado da Análise Discriminante dos Componentes Principais (DAPC). Tal análise sugeriu a formação de 5 subgrupos genéticos principais na população, onde os mesmos são representados por diferentes cores. Entre os 5 subgrupos formados, o 1 e o 2 (representados pelas cores vermelha e amarela respectivamente) chamaram a atenção pela proximidade e sobreposição parcial, enquanto os grupos 3 (verde), 4 (azul) e 5 (lilás) encontram-se isolados, sendo que o 3 é o menor e mais divergente.

Figura 1. *Scatterplot* da Análise Discriminante dos Componentes Principais (DAPC) com 2 funções discriminantes nos 259 acessos do germoplasma de arroz, mais 10 cultivares comerciais estruturados em 5 grupos. Foram inseridos na parte superior esquerda a PCA (*Principal component analysis*) e na parte inferior a DA (*Discriminant Analysis*) Eigenvalues. Representação em 2 dimensões.

Figura 2. *Scatterplot* da Análise Discriminante dos Componentes Principais (DAPC) com uma função discriminante nos 259 acessos do germoplasma de arroz, mais 10 cultivares comerciais estruturados em 5 grupos. Representação em 2 dimensões.

O número de grupos (K) testados variou de 1 a 10, e o K ótimo foi determinado com base no Bayesian Information Criterion (BIC) (JOMBART; DEVILLARD; BALLOUX, 2010) e proporcionou a formação de 5 subgrupos genéticos principais na população (Figura 1, 2 e 3). A figura 3 apresenta o teste de atribuição de cada genótipo com K=5. A mesma cor em diferentes acessos indica que eles pertencem ao mesmo grupo. Diferentes cores no mesmo indivíduo indicam a probabilidade deste indivíduo pertencer ao grupo k. A identificação dos acessos pode ser melhor visualizada na tabela do Anexo A.

Figura 3. *Barplot* da Análise Discriminante dos Componentes Principais (DAPC) nos 259 genótipos do banco de germoplasma de arroz, mais 10 cultivares comerciais estruturados em 5 subgrupos

A análise de estrutura populacional estima qual a probabilidade de determinado genótipo ter origem de cada subgrupo considerado dentro da população em estudo a partir das frequências alélicas do conjunto de acessos. Neste sentido, foram analisados os dados genotípicos da população para verificar a possível existência de estrutura, o que constitui informação útil na determinação do número de grupos nas análises de agrupamento.

Nas Figuras 1, 2 e 3 visualizam-se os resultados da Análise Discriminante dos Componentes Principais (DAPC). A partir dos dados coletados, é possível verificar que existem evidências para significativa estrutura populacional no painel de acessos de arroz do Banco de Germoplasma estudado. De modo geral, não existe informação detalhada a respeito da origem de cada genótipo. Sabe-se que parte deles procedem de materiais de origem filipina (142 acessos); a outra parte (117 acessos) é resultado de coletas feitas no território brasileiro, provenientes de diversas variedades locais mantidas por agricultores, variedades obsoletas e 10 cultivares comerciais desenvolvidas e cultivadas no Brasil.

A dispersão gráfica, permitiu separar subjetivamente os 269 genótipos avaliados. Foram formados 5 subgrupos principais na população. O maior grupo identificado foi o Grupo 1 (vermelho), com 171 genótipos, o que corresponde a 63% do total. Nele estão contidos as cultivares BRS Pampa; AN Cambará; Esmeralda e Sertaneja, a maioria das variedades brasileiras (111 acessos) e 56 acessos filipinos. Os menores, são o Grupo 3 (verde), com apenas 4 genótipos (2 brasileiros e 2 filipinos) e o Grupo 5 (lilás) que contém apenas 6 genótipo, sendo uma variedade brasileira e outras 5 cultivares testemunhas amplamente utilizadas no Brasil (Epagri 109; Tio Taka; SCS 116 Satoru; Andosan e Markes). O Grupo 4 (azul) compreende 20% dos genótipos (55 acessos + testemunha BRS Primavera). A sobreposição

dos grupos 1 e 2, observável (Figuras 1, 2 e 3), se deve provavelmente ao compartilhamento de características semelhantes.

A DAPC é um método de análise multivariada, utilizado para identificar e descrever agrupamentos de indivíduos relacionados geneticamente em determinada população. Em primeiro lugar, faz-se uma PCA - análise de componentes principais, em seguida uma DA - análise discriminante, levando em consideração escores obtidos na análise dos componentes principais. Em outras palavras, a DAPC minimiza a variação dentro de grupos e maximiza a variação entre grupos resultando num número mais provável de subgrupos evidenciando a estruturação (JOMBART; DEVILLARD; BALLOUX, 2010). Recomenda-se este tipo de análise quando não se tem informações de grupos *a priori*, sendo neste caso mais indicada do que a utilização do programa *Structure* para caracterizar subdivisões populacionais. Destaca-se que neste caso não é necessário assumir ausência de ligação nem pressupostos de Equilíbrio de Hardy-Weinberg (PRITCHARD; STEPHENS; DONNELLY, 2000).

Neste sentido, para identificar o número ótimo de grupos em tais populações, a DAPC realiza os agrupamentos por meio das médias K e “Bayesian Information Criterion” (BIC). Segundo Jombart; Devillard; Balloux (2010), as ferramentas implementadas na DAPC tanto para a identificação de grupos, bem como a representação gráfica de estruturas entre grupos permite desvendar a complexidade estrutural de determinada população. A presença de estrutura genética em arroz tem sido documentada em diversos estudos e isso torna possível explorar os *pools* de genes de arroz de forma mais eficaz com abordagens baseadas em genética de populações utilizando as extensas coleções de recursos genéticos de arroz (GARRIS; MCCOUCH; KRESOVICH, 2005).

2.3.3 Diversidade genética

Na Tabela 1 estão apresentados os resultados da diversidade genética encontrada para cada grupo formado pelo DAPC no painel de acessos em estudo. O valor médio encontrado para heterozigosidade observada (H_O) foi de 0,293 e para a heterozigosidade esperada (H_S) foi de 0,277, o valor relacionado ao coeficiente de endogamia (F_{IS}) foi de 0,014 e de Riqueza Alélica (Ar) que refere-se ao número de alelos diferentes para uma mesma região do genoma, foi de 1,258.

Tabela 1. Índices de diversidade genética. Heterozigosidade observada (H_O); Heterozigosidade esperada (H_S); Coeficiente de endogamia (F_{IS}) e Riqueza alélica (Ar) por grupo formado e média geral para o Banco de Germoplasma da ESALQ-USP

Grupo	Parâmetro			
	H_O	H_S	F_{IS}	Ar
G1	0,149	0,436	0,468	1,337
G2	0,186	0,123	0,052	1,123
G3	0,229	0,151	0,013	1,152

G4	0,376	0,351	-0,014	1,352
G5	0,525	0,323	-0,447	1,326
Média	0,293	0,277	0,014	1,258

A análise de diversidade revelou variabilidade genética no material avaliado. Em estudos de diversidade genética, é importante quantificar as diferenças existentes dentro da população de estudo e entre suas subpopulações. Através do estudo pode-se verificar que existe diversidade entre os genótipos avaliados, destacando-se que a heterozigotidade observada não diferiu estatisticamente para a esperada ($H_o = 0,293$ e $H_s = 0,277$ respectivamente). Como o arroz é uma espécie autógama, portanto, o fluxo de genes é restrito. De maneira geral, a frequência de locos heterozigotos é baixa nessas espécies uma vez que a cada geração de autofecundação eles são reduzidos à metade, resultando num baixo nível de heterozigotidade.

Wright (1969) introduziu o conceito de estatísticas F que utilizam o coeficiente de endogamia para descrever a variabilidade genética existente. Com base nas estimativas das frequências observadas (H_o) e esperadas (H_s) de heterozigotos, é possível calcular o índice de fixação ou coeficiente de endogamia F de Wright ($F_{is} = 0,014$). Este índice é amplamente utilizado em genética de populações, pois mede o grau de endocruzamento de um indivíduo em relação ao resto da sua subpopulação, refletindo a probabilidade de que dois alelos dentro do mesmo indivíduo sejam idênticos por descendência (WRIGHT, 1978). No entanto, é importante destacar que tais conceitos são mais utilizados em estudos de populações naturais e apresentam grande importância para estabelecer programas de conservação. Quando se trata da caracterização e manutenção de Bancos de Germoplasma, tem-se outro cenário, uma vez que não se trata de uma população panmítica. Bancos de Germoplasma são em essência, reservatórios genéticos. É desejável para os programas de melhoramento manter variabilidade, quando se considera espécies autógamas deve-se tomar cuidado com a taxa de fixação de alelos uma vez que naturalmente possuem menor fluxo gênico e assim a perda de alelos ou blocos gênicos pode ser irreversível.

Como componente da diversidade genética também destaca-se a riqueza alélica identificada, ou o número de alelos diferentes segregando em uma população. Para locos bialélicos o número varia de zero a dois, e neste estudo foi observado em média de 1,258 alelos por loco.

2.3.4 Agrupamento

A Figura 4, apresenta o dendrograma obtido no agrupamento dos genótipos. Este refere-se à ordenação em grupos dos indivíduos da população utilizando o critério UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*) baseado na matriz de distâncias de Rogers (1972) modificada por Wright (1978) e sugere a formação de 10 principais grupos. Todos os “*bootstraps*” do dendrograma foram maiores que 100, isso refere-se a consistência dos nós (dada em porcentagem) e indica que os acessos ficaram bem posicionados.

De modo geral, não houve separação relacionada à origem dos acessos. Os Grupo 1, 2, 4, 5, 6, 9 e 10 são formados por apenas um acesso BGA458 de origem filipina; BGA312 variedade brasileira; BGA376 filipino; BGA306 brasileiro; BGA228 brasileiro; BGA291 brasileiro e BGA431 filipino, respectivamente. Os maiores grupos são o 3 e o 7. O grupo 3 contém 98 genótipos sendo importante destacar que neste estão alocadas 4 cultivares utilizadas como testemunhas (BRSPampa, NA Cambará, Esmeralda e Sertaneja) e outros acessos filipinos e brasileiro e o grupo 7 com 32 genótipos, contém as outras testemunhas (Primavera; Epagri 109; Tio Taka; Satoru; Andosan e Markes) bem

como acessos filipinos e brasileiros. O grupo 8 contém 5 genótipos sendo 4 filipinos (BGA410; BGA408; BGA338; BGA429) um brasileiro (BGA276).

Figura 4. Dendrograma gerado pela análise de agrupamento de 269 genótipos obtidos através do critério UPGMA Sokal; Michener, 1958) baseado na matriz de distâncias de Rogers (1972) modificada por Wright (1978)

A análise de agrupamento, busca discriminar os indivíduos geneticamente e permite separá-los por grupos por meio da análise de um conjunto de caracteres, marcadores genéticos ou características inerentes a cada indivíduo. Em outras palavras ela leva em consideração algum critério classificatório para agrupar de modo que haja homogeneidade dentro de determinado grupo e heterogeneidade entre eles (MARDIA; KENT; BIBBY, 1979). O dendrograma resultante do agrupamento dos genótipos refere-se a ordenação dos indivíduos da população pelo critério UPGMA, sugere a formação de 5 grupos genéticos principais (Figura 4). Este método é baseado na matriz de distâncias de Rogers (1972) modificada por Wright (1978).

Destaca-se que as cultivares utilizadas como testemunhas eram pertencentes à subespécie indica, sendo que 6 delas foram desenvolvidas para ambientes irrigados (Ep109; Tio Taka; 116 Satoru; Andosan; Marques e BRS Pampa) e 4 para regiões sequeiro ou terras altas (BRS Primavera, AN Cambará; Esmeralda e Sertaneja). Isso se deve ao fato de que são de maior interesse no mercado brasileiro e por isso representam a maioria das áreas de cultivo de arroz no Brasil. Não estão registrados no Banco de Germoplasma avaliado, a subespécie a que pertencem os acessos brasileiros e filipinos, no entanto sabe-se pelo tipo de grão que os brasileiros possuem mais características atribuídas à subespécie indica, já os filipinos à japônica.

Destaca-se a questão de como as subespécies de arroz indica e japônica poderiam manter-se diferenciadas uma vez que não há um verdadeiro isolamento (MATSUO et al. 1997). O mesmo autor destaca que essas duas subespécies são cultivadas em grandes áreas desde a Índia até a China onde estão presentes vários tipos intermediários. Quando considera-se as Américas, a maioria das cultivares utilizadas nos Estados Unidos pertence à subespécie japônica, em contrapartida, em regiões tropicais, predomina a subespécie indica. Nos estados do sul do Brasil as 10 cultivares mais plantadas pertencem à subespécie indica (onde a cultivar IRGA 424 RI é responsável por 48% da área do RS – maior estado produtor), (IRGA2018). A cultivar IRGA 417, responsável pela maior área cultivada durante anos (IRGA, 2002), é oriunda do cruzamento entre NEW REX e outras duas linhagens. Com base em sua genealogia, percebe-se que mesmo tendo a qualidade de grãos e outras características próprias da subespécie indica, 25 % do seu conjunto gênico é do tipo japônica (oriundo da NEW REX). Brown-Gudiera (2000) fizeram um estudo de agrupamento utilizando RAPD e microssatélites esta cultivar é agrupada junto com as de padrão indica. Desta forma, fica evidenciado que pode haver algum mecanismo associado ao processo seletivo que a levou ao aumento na frequência de alelos da subespécie indica.

2.4 Conclusão

Existe diversidade genética entre os acessos avaliados, sendo que o Banco de Germoplasma de arroz da ESALQ/USP é uma potencial e importante fonte de acessos que podem ser incorporados aos programas de melhoramento. Foram identificados 5 subgrupos na população avaliada. É importante destacar que não foram identificadas duplicatas entre os acessos avaliados, sendo assim justificável manter-se todos eles na coleção nuclear do Banco. A tecnologia de genotipagem por sequenciamento como a *DarT-seq* mostrou grande eficiência no estudo de diversidade genética entre tais acessos.

REFERÊNCIAS

- BROWN-GUEDIRA, G. L. Evaluation of Genetic Diversity of Soybean Introductions and North American Ancestors Using RAPD and SSR Markers. *Crop Science*, v. 40 p. 815, 2000.
- BRONDANI, C. et al. Transferability of microsatellite and sequence tagged site markers in *Oryza* species. *Hereditas*, v.138, p.187-192, 2003.
- BRONDANI, C. et al. Determination of genetic variability of traditional varieties of Brazilian rice using microsatellite markers. *Genetics and Molecular Biology*, v. 29, p. 676-684, 2006.
- CHANG, T. T.; BUNTING, A. H. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, The Rice Cultures*, v. 275, n. 936, p. 143-157 1976.
- DOYLE, J.J.; DOYLE, J.L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, v.19, p.11-15, 1987.
- ELSHIRE R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *Plos One*. v. 6, e19379. 2011.
- GARRIS, A. S.; MCCOUCH, S.; KRESOVICH, S. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). *Genetics*, v. 165, n. 2, p. 759–769, 2005.
- INSTITUTO RIO-GRANDENSE DO ARROZ – IRGA. Equipe de Política Setorial. Dados de Safra. Série Histórica da área Plantada, Produção e Rendimento. Disponível em: <<http://www.irga.rs.gov.br/srquivos/2005020011092011.pdf>>. Acesso em: 10 de abr. de 2018.
- JOMBART, T.; DEVILLARD, S.; BALLOUX, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, v. 11, n. 1, 2010.
- KHUSH, G. S. Origin, dispersal, cultivation and variation of rice. *Oryza: From Molecule to Plant*, p. 25–34. 1997.
- KUMBHAR, S. D. et al. Genetic diversity and population structure in landraces and improved rice varieties from India. *Rice Science*, v. 22, n. 3, p. 99–107. 2015.
- MARDIA, K.V.; KENT, J. T.; BIBBY, J. M. *Multivariate Analysis*. London, Academic, 1979, 529 p.
- MATSUO, T. et al. Science of the rice plant. Food and agriculture policy Research Center, v.3, Genetics, p.88-111, 1997.
- NEI, F.; TAJIMA T; TATENO Y. Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution*, v. 19, n. 2, p. 153–170, 1978.
- PERVAIZ, Z. H. et al. Determination of the genetic variability of asian rice (*Oryza sativa* L.) varieties using microsatellite markers. *Genetics and Molecular Biology*, v. 29, n. 4, p. 2009.
- PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. *Genetics*, v. 155, n. 2, p. 945–959, 2000.
- R CORE TEAM (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL. <https://www.r-project.org/>.
- ROGERS, J. S. Measures of similarity and genetic distance. In *Studies in Genetics VII*. pp. 145–153. 1972.
- SOKAL, R.; MICHENER, C. "A statistical method for evaluating systematic relationships". *University of Kansas Science Bulletin* v. 38, p. 1409–1438, 1958.
- VARSHNEY, R. K.; GRANER, A.; SORRELLS, M. E. Genomics-assisted breeding for crop improvement. *Trends in Plant Science*, v. 10, n. 12, p. 621–630, 2005.

WEIR, B. S.; Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution*, p. 1358–1370, 1983.

WRIGHT, S. *Evolution and the genetics of populations*. v. 2. The theory of gene frequencies. The University of Chicago Press, Chicago, 1969.

WRIGHT, S. *Evolution and the genetics of populations*. v. 4. Variability within and among populations. The University of Chicago Press, Chicago, 1978.

3 ASSOCIAÇÃO GENÔMICA AMPLA PARA CARACTERES DE INTERESSE AGRONÔMICO EM ARROZ UTILIZANDO MARCADORES DART-SEQ

Resumo

Este trabalho teve como objetivo, realizar o mapeamento associativo para catorze caracteres fenotípicos em um painel contendo cento e quarenta e dois acessos de arroz de origem filipina, pertencentes ao Banco de Germoplasma de Arroz do Departamento de Genética da ESALQ/USP. A instalação dos experimentos de caracterização fenotípica foi realizada em áreas de cultivo de sequeiro, na área experimental do Departamento de Genética em dois anos agrícolas. O delineamento experimental foi de blocos aumentados, sendo realizados os tratamentos culturais recomendados para a cultura. Os acessos foram caracterizados de acordo com descritores indicados pelo IBPGR – IRRI (Rice Advisory Committee). Os caracteres avaliados foram: Número de dias para a emergência (NDE); Número de dias para o florescimento (NDF); Ciclo total da planta (CTP); Massa de cem sementes (MCS); Produtividade de grãos (PG); Comprimento do colmo (CC); Número de Perfilhos (NP); Comprimento e Largura da folha bandeira (CFB e LFB); Altura de planta na maturidade (APM); Comprimento da Panícula (CP); Comprimento e largura da espigeta (CE e LE); Número de ramificações por panícula (NRP). Os dados coletados foram analisados através do software R (R DEVELOPMENT CORE TEAM, 2018) utilizando modelos mistos, e as médias corrigidas para os valores de BLUP (Best Linear Unbiased Predictor), utilizadas para o mapeamento. A genotipagem foi realizada utilizando marcadores SNPs (Single Nucleotide Polimorfism) gerados por meio da tecnologia DArTseq (Diversity Arrays Sequence) de genotipagem por sequenciamento (GBS). Foi realizada uma análise de associação genética ampla (GWAS) onde, foram inseridas as matrizes de parentesco (kinship) e estrutura populacional por meio da análise de componentes principais. Após o processo de filtragem e imputação, os dados resultaram num total de 7989 SNPs distribuídos nos 12 cromossomos do arroz. Os marcadores utilizados foram altamente informativos e seletivos para análise de diversidade genética e estudos de associação genômica ampla nos genótipos de arroz testados. A população estudada pode ser subdividida em 3 grupos em função da análise de componentes principais e foram identificados 33 SNP's associados a oito caracteres quantitativos avaliados nos acessos.

Palavras-chave: *Oryza sativa*; Banco de germoplasma; Mapeamento associativo; Análise de associação

Abstract

The objective of this work was to perform associative mapping for fourteen phenotypic characters in a panel containing one hundred and forty two accessions of rice of Filipino origin belonging to the Rice Germplasm Bank of the Department of Genetics of ESALQ/USP. The establishment of the phenotypic characterization experiments was carried out in rainfed areas, at the experimental farm of the Department of Genetics, in two agricultural years. The experimental design was of increased blocks, being the cultural treatments recommended for the culture. The accesses were characterized according to descriptors indicated by IBPGR - IRRI (Rice Advisory Committee). The evaluated characters were: Number of days for the emergency (NDE); Number of days for flowering (NDF); Total plant cycle (CTP); One hundred seed mass (MCS); Grain yield (PG); Height of high (CC); Number of Profiles (NP); Length and width of the flag leaf (CFB and LFB); Height of plant at maturity (APM); Panicle Length (CP); Length and width of the spikelet (CE and LE); Number of branches per panicle (NRP). The collected data were analyzed using R (R DEVELOPMENT CORE TEAM, 2018) using mixed models, and corrected means for BLUP (Best Linear Unbiased Predictor) values, used for the mapping. Genotyping was performed using single nucleotide polymorphisms (SNPs) generated by sequencing genotyping (GBS) DArTseq (Diversity Arrays Sequence) technology. A broad genetic association (GWAS) analysis was performed, where kinship matrices and population structure were

inserted through principal component analysis. After the filtering and imputation process, the data resulted in a total of 7989 SNPs distributed on the 12 rice chromosomes. The markers used were highly informative and selective for analysis of genetic diversity and studies of broad genomic association in the tested rice genotypes. The studied population can be subdivided into 3 groups as a function of principal component analysis and 33 SNPs associated with eight quantitative traits evaluated in the accessions were identified.

Keywords: *Oryza sativa*; Germplasm bank; Associative mapping; Association analysis

3.1 Introdução

O arroz é pertencente ao gênero *Oryza*, da família *Poaceae* (Vaughan, 1989). Dentro deste grupo, duas espécies são cultivadas atualmente (*O. sativa* e *O. glaberrima*). Destaca-se que *O. sativa* representa grande importância para a agricultura mundial, uma vez que mais de 90% do cultivo corresponde a esta espécie (KUSH; BRAR, 2002). Kato et al. (1928), introduziu sua classificação em duas subespécies: *i*) indica e *ii*) japônica, onde observa-se grande diversidade genética. Este fato indica que podem ter ocorrido eventos independentes de domesticação a partir de um ancestral comum (*O. rufipogon*), e as subespécies podem ter se diferenciado há milhares de anos devido a um isolamento geográfico (SWEENEY; MCCOUCH, 2007). Em geral, a subespécie indica é encontrada em regiões inundadas da Ásia tropical, já a subespécie japônica, nas regiões de terras altas no Sul da Ásia (GARRIS; MCCOUCH; KRESOVICH, 2005). Graças à sua diversidade genética, variedades de arroz possuem elevada capacidade de adaptação, podendo tolerar estresses bióticos e abióticos, que combinados com boas práticas de manejo, resultam em rendimentos altos e estáveis (ZONG, 2007; ZHANG, 2009).

O genoma do arroz é compacto, possui 12 cromossomos com cerca de 380 Mpb (Milhões de Pares de Base). Além disso, apresenta alto grau de sintenia com o genoma de outros cereais, podendo ser utilizado como planta modelo para a caracterização de genomas maiores (YU; BUCKLER, 2006). O sequenciamento do seu genoma é de grande importância no sentido social e econômico, e tem sido uma importante ferramenta para o melhoramento genético, pois permitiu a obtenção de informações detalhadas a respeito do funcionamento e estrutura de genes de importância agrônoma, por meio da construção de mapas de ligação, desenvolvimento de marcadores de DNA, bem como a análises de QTL's (Quantitative Trait Loci) (Xu et al. 2004).

Diversas abordagens estatísticas podem ser utilizadas para estudar caracteres complexos. Quando se trata de genômica estrutural, duas abordagens são utilizadas normalmente, o mapeamento de QTLs e o GWAS (Genome Wide Association ou Mapeamento Associativo), ambas baseada no princípio da recombinação. A principal diferença entre os métodos é que no mapeamento de QTLs são necessários cruzamentos biparentais, onde faz-se o estudo em populações derivadas destes. Já no mapeamento associativo, observa-se um grande conjunto de indivíduos com pouco vínculo de parentesco. Neste caso, há maior variabilidade genética quando comparado à populações oriundas do cruzamento de dois parentais, sendo mais eficiente no estudo de caracteres complexos (YU; BUCKLER, 2006).

Estudos de associação (ou mapeamento associativo) estabelecem uma conexão entre o fenótipo e o genótipo, permitindo identificar pequenas variações a nível de DNA que resultam em diferenças fenotípicas ((FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003). Tais estudos podem ser realizados em espécies alógamas como milho e diversas outras florestais, e autógamas como o arroz e o trigo. Em alógamas, exige elevado número de marcadores, pois em geral apresentam baixo desequilíbrio de ligação e o gene-alvo estaria próximo a região genômica

envolvida, uma vez que a cobertura do genoma estaria envolvendo milhões de SNP's (Single Nucleotide Polimorfism). Em autógamias, o desequilíbrio é relativamente maior, consequentemente, um menor número de marcadores assegura a cobertura do genoma, porém neste caso o marcador pode estar longe do gene-alvo (ARANZANA et al. 2005). De acordo com Zhu et al. (2008), isso se deve ao fato de que no mapeamento de QTLs, as oportunidades de ocorrer recombinação entre as famílias resultam num mapeamento de baixa resolução. Em contrapartida, o mapeamento associativo traz como resultado mapas de alta resolução, uma vez que são explorados, a diversidade genética natural da espécie e seu histórico de recombinações.

Diversas vantagens podem ser destacadas quanto ao uso do mapeamento associativo para o melhoramento genético. Permite-se por exemplo, explorar toda a variabilidade genética existente em um banco de Germoplasma, uma vez que não trabalha-se apenas com dois alelos conhecidos (como no caso do estudo de populações oriundas de cruzamentos), mas sim com uma série alélica; Não é necessário o desenvolvimento de populações segregantes, resultando em redução no tempo da pesquisa (YU; BUCKLER, 2006; GUPTA; RUSTGI; KULWAL, 2005); Há um expressivo aumento na resolução do mapa (YU et al. 2006; ZHU et al. 2008), além de que podem ser utilizados dados moleculares e fenotípicos de outros experimentos.

A Escola Superior de Agricultura “Luiz de Queiroz” – ESALQ/USP possui um Germoplasma de arroz provenientes de instituições de pesquisas do Brasil e do mundo. Dentre eles existem diversos acessos das subespécies indica e japônica. Estes acessos vem sendo estudados com a intenção de obter informações a respeito de sua origem, diversidade e estrutura genética, bem como sua caracterização fenotípica de modo que possam ser utilizados em programas de melhoramento. Desta forma, este trabalho teve como objetivo, realizar o mapeamento associativo para 13 características em um painel contendo 142 acessos de arroz de origem filipina, pertencentes ao Banco de Germoplasma de Arroz do Departamento de Genética da ESALQ. Para esse estudo, foram utilizados marcadores SNPs gerados por meio da tecnologia DArTseq (Diversity Arrays Sequence) de genotipagem por sequenciamento (GBS).

3.2 Material e Métodos

3.2.1 Material vegetal:

Este estudo foi realizado por meio de um painel contendo 142 acessos filipinos de arroz pertencentes ao Banco de germoplasma da ESALQ/USP - Departamento de Genética e Melhoramento de Plantas – Campus USP de Piracicaba, São Paulo. Tais acessos tem características predominantes de materiais de sequeiro ou terras altas e pertencem à subespécie japônica.

3.2.2 Fenotipagem:

A fenotipagem dos acessos filipinos foi realizada por Mata, (2010) que estudou tais materiais a fim caracterizar sua diversidade por meio de marcadores agromorfológicos. A instalação dos experimentos de caracterização fenotípica foi realizada em áreas de cultivo de sequeiro, na fazenda experimental do Departamento de Genética da ESALQ/USP em dois anos agrícolas (2007/2008 e 2008/2009). O delineamento experimental foi de blocos aumentados, sendo realizados os tratos culturais recomendados para a cultura.

Os acessos foram caracterizados de acordo com descritores indicados pelo IBPGR – IRRI Rice Advisory Committee (1980). Catorze caracteres avaliados, foram utilizados para o mapeamento associativo: Número de dias para a emergência (NDE) – desde a semeadura até a emergência; Número de dias para o florescimento (NDF) – desde a semeadura até o florescimento de 50% das panículas; Ciclo total da planta (CTP) – desde a semeadura até a maturação completa de 85-90% da parcela; Massa de cem sementes (MCS) – determinada em balança de precisão sob umidade de 13%; Produtividade de grãos (PG) – pesagem de grãos colhidos na parcela útil (2 m²) com umidade corrigida para 13%; Comprimento do colmo (CC) – medido no enchimento de grãos; Número de Perfilhos (NP) – Contados após a floração; Comprimento e Largura da folha bandeira (CFB e LFB); Altura de planta na maturidade (APM); Comprimento da Panícula (CP); Comprimento e largura da espiguetta (CE e LE) - medidos à colheita; Número de ramificações por panícula (NRP) – contados a partir do eixo central da panícula.

A parcela experimental foi composta por 4 linhas de 2 m de comprimento sendo a área útil da parcela composta pelas 2 linhas centrais. Foram realizados tratos culturais recomendados para a cultura. Maiores informações sobre a caracterização fenotípica pode ser observada em Mata, (2010).

Os dados coletados foram analisados através do software R (R DEVELOPMENT CORE TEAM, 2018) utilizando modelos mistos, e as médias corrigidas para os valores de BLUP (Best Linear Unbiased Predictor), utilizadas para o mapeamento.

3.2.3 Extração de DNA:

Foi extraído o DNA genômico a partir de folhas jovens, utilizando o método de extração CTAB descrito por Doyle; Doyle (1987), modificado. O DNA foi diluído em solução de TE (Tris-HCl 10 mM, EDTA 1 mM, com pH 8,0) em concentrações entre 50ng/μl e 100ng/μl. Sua integridade foi verificada em gel de agarose a 1%, sendo atestada devido ao aparecimento de banda única com elevado peso molecular sobre o gel.

3.2.4 Genotipagem DArTseq:

O DNA dos acessos foi remetido à empresa Diversity Arrays Technology Pty Ltd. (DArT P/L) na Austrália onde foram genotipados por meio da plataforma tecnologia de genotipagem DArTseq™, descrito por (COURTOIS et al. 2013). Em síntese, tal metodologia combina “Genotyping by Sequencing”, (GBS) com Diversity Arrays Technology (DArT) que envolve a redução de complexidade do genoma utilizando enzimas de restrição *PstI/TaqI* e em seguida o sequenciamento de sequências curtas através pelo Illumina HiSeq 2500 (Maiores informações podem ser obtidas em <<http://www.diversityarrays.com/dart-application-dartseq-data-types>>) (KILIAN et al. 2012). A empresa fornece além da informação de marcadores, um mapa genético resultado da determinação da posição genética nos respectivos cromossomos em que os marcadores estão inseridos. Tal informação é baseada no alinhamento com o genoma de referência do arroz Rice_v9 (Phytosome assembly v9).

3.2.5 Filtragem dos dados

Os marcadores SNPs obtidos pela genotipagem com a tecnologia DArT-seq passaram por uma filtragem dos dados a fim de garantir a qualidade dos dados para posterior análise. Fez-se a filtragem dos dados utilizando os seguintes parâmetros: *i*) Reprodutibilidade ($Rep\ Avg \geq 0,95$; *ii*) $Call\ Rate \geq 0,9$; *iii*) Conteúdo de informação polimórfica (PIC) $\geq 0,1$. Em seguida, foi realizada a eliminação de SNPs com frequência menor de 5%, (ou seja, foram descartados aqueles com MAF - Minimum Minor Allele Frequency - maior que 0,05. No conjunto de dados foi observado que havia cerca 5% de dados faltantes, então fez-se a imputação dos mesmos por meio do algoritmo Beagle implementado no Software R – pacote “*symbred*”.

3.2.6 Estrutura da população, matriz de parentesco e desequilíbrio de ligação:

A matriz Q (estrutura populacional) foi obtida por meio da análise de componentes principais (PCA) (PRICE et al. 2006) calculada por meio do GAPIT (Genome Association and Prediction Integrated Tool) (LIPKA et al. 2012; TANG et al. 2016) do software R. Em seguida fez-se uma Análise Discriminante de Componentes Principais (DAPC) para verificar a formação de subgrupos. A kinship - K (matriz de parentesco) para os marcadores SNPs foi inferida pelo algoritmo Loiselle (LOISELLE et al. 2008). O desequilíbrio de ligação (DL), foi calculado pelo quadrado da correlação (r^2) (HILL; ROBERTSON, 1968). A média de marcadores adjacentes foi calculada usando uma janela de dez marcadores. As análises foram feitas através do pacote “FarmCPU” (Fixed and random model Circulating Probability Unification, version 1.02) implementado no Software R (LIU, et al. 2016).

3.2.7 Mapeamento associativo:

Para a realização do mapeamento, foram inseridas as matrizes de parentesco (kinship) e estrutura populacional (PCA) de modo a reduzir a possibilidade de ocorrência do erro tipo I. Foi utilizado o conjunto de marcadores SNPs e a abordagem para realizar a análise de associação baseada em Modelos Lineares Mistos (MLM), implementados no FarmCPU (LIU, et al. 2016). Foram consideradas significativas as associações entre marcador e caráter com valor $p < 0,001$.

3.2.8 Anotação de genes:

SNPs que tiveram associação significativa com as características fenotípicas avaliadas foram comparadas ao genoma de referência Nipponbare *Oryza sativa* ssp. *japonica* (Os-Nipponbare-Reference-IRGSP-1.0). As comparações foram feitas utilizando-se uma janela de 200 kb para os respectivos SNPs (VOLANTE et al. 2017), onde foi possível verificar se o mesmo se encontrava posicionado em genes ou em região intergênica. Desta forma foi possível obter a sequência transcrita dos genes baseados no genoma de referência disponível no banco de dados *Gramene* (<http://www.gramene.org/>). Após a identificação dos transcritos, foi possível realizar uma busca referente a anotação funcional de tais genes em três classes de termos GO, depositados no banco de dados do UniProt (<http://www.uniprot.org/>).

3.3 Resultados e Discussão

A genotipagem de 142 acessos de arroz filipino pertencentes ao banco de germoplasma do Departamento de Genética da ESALQ-USP de Piracicaba, foi realizada pela empresa Diversity Arrays Technology Pty Ltd. por meio da tecnologia de genotipagem por sequenciamento DArTseq™, descrita por Courtois et al. (2013). Após o processo de filtragem e imputação, os dados resultaram num total de 7989 SNPs distribuídos nos 12 cromossomos do arroz. Estes foram utilizados nas análises de associação de catorze caracteres fenotípicos. Os resultados, indicaram uma distribuição média de 666 SNPs por cromossomo, sendo que o maior número de marcas estão alocadas no cromossomo 1 (1078). Já o menor número de marcas foram identificadas nos cromossomos 11 e 12, ambos com 496 SNPs (Figura 5).

Cromossomo	SNP
1	1078
2	893
3	917
4	617
5	644
6	669
7	681
8	503
9	497
10	498
11	496
12	496
Número de Marcas (SNPs)	7989

Figura 5. Número de marcas SNPs por cromossomos, identificadas entre os 142 acessos filipinos de arroz.

3.3.1 Matriz de parentesco, estrutura populacional e desequilíbrio de ligação

Matriz de Parentesco

A figura 6, apresenta a matriz de parentesco ou matriz K (Kinship), calculada para os marcadores SNPs por meio do algoritmo de Loiselle (LOISELLE et al. 2008). Similaridades foram representadas entre pares indivíduos, na linha central do gráfico, quanto mais próximo de 0 é o valor, menor a similaridade entre os entre eles.

Figura 6. Matriz de parentesco (Kinship) calculado pelo algoritmo Loiselle utilizando marcadores SNPs.

A matriz de parentesco (ou “Kinship” - k), fornece informações para estudos de caracteres quantitativos, por meio dela, é possível obter informações sobre a similaridade entre dois indivíduos tomados de forma aleatória em uma população (YU et al. 2006; ZHANG et al. 2014). Marcadores moleculares SNPs, foram utilizados para calcular a

matriz k entre pares de indivíduos por meio do algoritmo de Loiselle (LOISELLE et al. 2008) neste trabalho. Tais similaridades estão apresentadas na Figura 5. Observando-se o eixo central da figura, pode ser verificado que os genótipos estudados não apresentam forte relação entre si, visto que quanto mais próximos de zero os valores, menor a similaridade entre pares de indivíduos. Neste sentido, pode-se indicar que os acessos filipinos do Banco de Germoplasma da ESALQ/USP são oriundos de diferentes regiões de cultivo de arroz, uma vez que não apresentam forte relação entre si.

Estrutura Populacional

Na Figura 7, pode-se observar a estruturação genética observada nos acessos filipinos do Banco de Germoplasma de arroz. A PCA revelou uma clara separação, formando 3 grupos distintos, sendo que um indivíduo ficou muito distante destes. Os primeiros dois componentes explicam 56,23% da variância, sendo 49,69% para PC1 e 6,54 para PC2.

Figura 7. Distribuição dos 142 acessos de Arroz Filipino do Banco de Germoplasma da ESALQ/USP baseado em PCA bidimensional.

Sabe-se que tanto a estrutura genética populacional quanto o parentesco entre indivíduos podem influenciar diretamente a ocorrência de falsos positivos das análises de mapeamento associativo e estudos de DL (TOMICIC et al. 20015; CHEN; PATTERSON; REICH 2010; ADETUNJI et al. 2014). Assim, ambos foram estimados para o referido estudo. A estrutura da população do painel de arroz foi inferida por meio de PCA (*Principal Component Analysis*), a qual foi realizada pelo pacote GAPIT (Genome Association and Prediction Integrated Tool) (LIPKA et al. 2012; TANG et al. 2016) do software R. Para a observação da formação de subgrupos dentro da população de acessos filipinos, pode-se verificar a Figura 8, onde tem-se 3 grupos distintos obtidos por meio da Análise Discriminante de

Componentes Principais (DAPC). As médias dos caracteres fenotípicos para cada grupo estão apresentadas na Tabela 2.

Figura 8. Scatterplot da Análise Discriminante dos Componentes Principais (DAPC) com 2 funções discriminantes nos 142 acessos filipinos do germoplasma de arroz, estruturados em 3 grupos. Foram inseridos na parte superior esquerda a PCA (*Principal component analysis*) e na parte inferior a DA (*Discriminant Analysis*) Eigenvalues. Representação em 2 dimensões

Tabela 2. Médias dos grupos formados por meio da Análise Discriminante dos Componentes Principais (DAPC), dos 14 caracteres fenotípicos avaliados nos 142 acessos filipinos de arroz

Número de															
GRUPO	genótipos	NDE	NDF	NP	CC	APM	CFB	LFB	CP	NRP	CE	LE	CTP	MCS	PG
1	30	7.23	123.83	5.04	81.40	102.00	24.49	1.31	20.50	2.37	8.29	3.03	142.47	2.26	0.36
2	55	8.05	121.91	4.19	91.71	111.42	24.78	1.56	19.87	1.71	9.01	3.66	134.38	3.20	0.42
3	57	7.75	115.88	4.06	98.14	102.65	23.43	1.59	19.98	1.61	8.58	3.34	129.42	2.82	0.43
Média	142	7.68	120.54	4.43	90.41	105.35	24.23	1.48	20.12	1.90	8.63	3.35	135.42	2.76	0.41

Desequilíbrio de Ligação

Os resultados do desequilíbrio de ligação (LD), calculados a partir de marcadores SNPs estão representados na Figura 8. Pode-se observar que há uma queda no desequilíbrio à medida que aumenta a distância genética entre eles. Cada ponto plotado no gráfico, representa a distância entre dois marcadores levando em consideração as informações de posição dos marcadores e o r^2 (coeficiente de correlação ao quadrado). Quanto mais próximo de 1 for o valor, maior o DL, quanto mais próximo de 0, o DL entre os marcadores é menor.

Figura 9. Desequilíbrio de ligação – DL, conforme a distância genética entre os pares de marcadores calculado por meio de marcadores SNPs por meio do pacote GAPIT (Software R).

Associações não aleatórias de alelos de diferentes locos são definidas como Associações em Desequilíbrio de Ligação (SLATKIN, 2008). Tais associações, resultam em desvios das frequências alélicas observadas quando comparadas às esperadas se independentes (WEISS, 2002). Estudos de Desequilíbrio de Ligação (DL), podem revelar importantes respostas sobre a história evolutiva de determinada população, uma vez que fornece informações sobre eventos passados como padrões relacionados à subdivisões geográficas e históricos da atuação do melhoramento. Além do mais, análises de DL são importantes ferramentas em estudos genéticos como o GWAS, determinam a densidade de marcadores necessários e a resolução do mapeamento. Segundo Weiss e Clark, 2002, o DL pode ser considerado a diferença entre proporções haplotípicas observadas e esperadas se os alelos segregassem de forma independente. Nota-se que quanto mais próximo de um for o valor entre marcadores, maior o desequilíbrio de ligação, quanto mais próximo de zero menor é o desequilíbrio entre eles, em outras palavras os alelos desses dois marcadores tendem a segregar de forma independente na população.

É importante destacar que a escolha da estatística r^2 em detrimento ao D' para análise do DL deve-se ao fato de que este último considera apenas o histórico de recombinação enquanto o r^2 abrange informações tanto do histórico de recombinação quanto dos eventos de mutação ocorridos na população (FLINT-GARCIA; THORNSBERRY; BUCKLER, 2003). Em espécies de fecundação cruzada o DL é geralmente menor do que em espécies autógamas, como o arroz. Isso se deve ao fato de que a autofecundação tende à homozigose, tornando as recombinações ineficazes na redução da DL (WÜRSCHUM et al. 2013).

3.3.2 Mapeamento associativo

Foram removidos da matriz os dados, marcadores com frequência alélica inferior a 5% ($MAF < 0,05$), para realizar as análises de associação (ou mapeamento associativo – GWAS), em seguida, dados faltantes foram imputados por meio do pacote “*symbreed*” (implementado no software R), levando em conta o genoma de referência do arroz.

Desta forma, as matrizes resultaram em 7989 marcadores SNPs analisados para 14 caracteres fenotípicos de 142 acessos.

O GWAS foi feito com base em modelos lineares mistos (MLM), utilizando a matriz K (kinship) de Loiselle e a estrutura da populacional por meio da análise de componentes principais (PCA), para reduzir os possíveis efeitos de estrutura populacional e parentesco, minimizando as chances de ocorrência de falsos positivos (erro tipo I).

Para verificar a existência de associações significativas entre SNPs resultantes da genotipagem dos 142 acessos de arroz, foram obtidos gráficos *Manhattan Plots*, construídos a partir das coordenadas genômicas de associações apresentadas no eixo X e com o logaritmo negativo do p-valor da associação no eixo Y. Associações significativas, foram consideradas aquelas com um valor $p \leq 0,001$.

Das catorze características analisadas, oito demonstraram marcas associadas, entre elas estão NDE; NDF; CTP; MCS; APM; CE; NP e LFB (Figuras 9 e 10). Foram identificados 33 marcadores associados significativamente aos caracteres, sendo que apenas uma associação foi identificada nos caracteres NDE e APM, e estão localizadas nos cromossomos 5 e 2, respectivamente. O caráter NDF foi o que apresentou o maior número de marcas associadas, sendo um total de 7, localizadas nos cromossomos 2, 3, 4, 5, 6, 10 e 11. Nos caracteres CTP e MCS foram encontradas 6 marcas, NP, 2 e LFB, 4. A localização destas nos respectivos cromossomos pode ser visualizada nas Figuras 9 e 10, bem como na Tabela 3.

Figura 10. *Manhattan plot* e *QQ-Plot* do mapeamento associativo para características fenotípicas NDE, NDF, CTP e MCS utilizando 7989 marcadores SNPs em 142 acessos de arroz filipino.

Figura 11. *Manhattan plot* e *Quantile Quantile-Plot* do mapeamento associativo para características fenotípicas APM, CC, NP e LFB, utilizando 7989 marcadores SNPs em 142 acessos de arroz filipino.

De acordo com Londo et al. (2006), acessos de arroz depositados em bancos de germoplasma podem representar uma fase intermédia entre cultivares elite e variedades oriundas da domesticação. Tais genótipos contêm grande diversidade genética e além de genes exóticos que proporcionam recursos para aplicação em programas de pré-melhoramento e melhoramento. Devido a isso, o mapeamento associativo realizado em acessos de germoplasma é de grande importância para observar variações disponíveis, e permite aos pesquisadores explorarem a diversidade natural existente, localizando genes de interesse através do uso de técnicas genéticas modernas explorar

a diversidade natural e localizar genes valiosos no genoma (ZHANG et al. 2014). Em suma, as análises de associação buscam identificar locos de caracteres quantitativos (QTLs) dissecando as associações entre marcadores e característica atribuídas ao desequilíbrio de ligação entre marcadores e polimorfismos funcionais em um conjunto de germoplasma (ZHU et al. 2008).

Neste estudo, as análises de associação foram feitas com base nos modelos lineares mistos (MLM). Este reúne informações a respeito da Matriz de Parentesco (Kinship) e Estrutura Populacional (PCA), aliado a correção FDR. Tal procedimento é de fundamental importância para reduzir o número de associações espúrias na análise (MAGIN et al. 2015; ADETUNJI et al. 2014), reduzindo a ocorrência do erro tipo I, muito comum devido a existência de populações estruturadas (RICHARDS; FRIESEN; BRUEGGEMAN, 2017). O método de controle de estrutura populacional utilizado foi a análise de componentes principais – PCA - por demandar menor tempo computacional e trabalhos realizados por Zhao et al. (2007) demonstraram este fornece resultados comparáveis com outras abordagens como o algoritmo STRUCTURE.

Um conjunto de marcadores SNPs obtidos da genotipagem por sequenciamento DArTseq, foram utilizados nas análises de associação. O mapeamento associativo foi realizado através das ferramentas disponíveis no pacote pacote “FarmCPU” (Fixed and random model Circulating Probability Unification, version 1.02) implementado no Software R (LIU, et al. 2016). Este método analisa grandes conjuntos de dados com tempo computacional mínimo, sendo por tanto vantajoso em relação aos métodos tradicionais. Foram tidas como significativas as associações entre marcadores e características com um valor p inferior a 0,001 (p -valor < 0,001).

Os SNPs com associação significativa foram demonstrados por meio de gráficos Manhattan Plot. Os mesmos foram elaborados a partir das coordenadas genômicas das associações onde os grupos de ligação estão apresentados no eixo X e no eixo Y estão os logaritmos negativos do p-valor da associação. Na análise de associação realizada entre 14 características, oito demonstraram marcas associadas, (NDE; NDF; CTP; MCS; apm; ce; NP e LFB) Sendo identificados 54 marcadores associados significativamente aos caracteres.

Tais associações foram estudadas nos bancos de dados Gramene e Rice Annotation Project Database (RAP-DB). Estes, são especializados em informações genéticas de espécies gramíneas, e contém uma base de dados específica para *Oryza sativa japonica* (IRGSP-1.0). Nestes bancos estão contidas informações de mapa genético de arroz além de sequências e genes detectados previamente. Tendo-se as informações do grupo de ligação e posição em que cada marcador está inserido, foi possível conhecer detalhes como, a região gênica na qual o SNP está inserido com base no International Rice Genome Sequencing Project (IRGSP) (Tabela 3). Diversos SNPs foram localizados dentro de genes, sendo assim, foi realizada uma busca sobre sua possível anotação funcional baseada nas três classes de termos GO, depositados no banco de dados do Rice Annotation Project Database.

3.3.3 Anotação de genes categorização funcional

A anotação de genes consiste na identificação e caracterização de regiões funcionais. Conhecendo-se a posição dos SNPs foi possível fazer uma busca no banco de dados “Rice Genome Annotation Project” (<http://rice.plantbiology.msu.edu/>), onde pesquisando-se as regiões flangeadoras (± 100 Kb; dados não mostrados (VOLANTE et al. 2017)) dos 33 SNPs significativos, foram observados um total de 491 genes candidatos subjacentes aos SNPs mapeados. A Tabela 3 apresenta informações detalhadas dos marcadores SNPs significativamente associados a alguma característica fenotípica. A fim de sumarizar as informações, foram demonstrados apenas genes contidos na

posição exata ou os mais próximos possíveis aos respectivos SNPs. Portanto, pode-se observar o caractere ao qual o SNP está associado, o cromossomo e a posição genética em que o mesmo está inserido, os p-valores, a MAF (*Minimum Allele Frequency* ou frequência alélica mínima) e o efeito atribuído a cada marca, sendo que este último representa a variação fenotípica explicada pelo SNP. Além do mais tem-se o nome do loco gênico de acordo com informações de bancos de dados a função desses genes e a sua classificação ontológica.

Tabela 3. Marcadores SNPs que apresentaram associação com as características NDE, NDF, CTP, MCS, APM, CC, NP e LFB., identificados entre os 142 acessos de *O. Sativa*. São apresentados: as marcas significativamente associadas (SNPs), o cromossomo, a posição do loco SNP no cromossomo (genoma de referência Rice_v9 - Phytosome assembly v9), o *p* valor, a mínima frequência alélica (MAF) e o efeito do SNP.

Caracter	Cr*	Posição	<i>p</i> -valor	MAF	Efeito	Nome do loco	Descrição	GO_id**	Ontologia
NDE	5	29357064	1.92E-14	0.09859	-0.92057	LOC_Os05g51180.1	hyaluronan/mRNA binding	GO:0003723	Função molecular
							family domain containing	GO:0005634	Componente celular
							protein, expressed	GO:0008150	Processo biológico
								GO:0005737	Componente celular
NDF	2	22474504	1.86E-07	0.12676	-7.42021	Não há gene candidato nesta região			
	3	18641167	8.64E-09	0.47535	-15.1998	LOC_Os03g32590.1	transcription initiation factor, putative, expressed	GO:0006139	Processo biológico
								GO:0009058	Processo biológico
	4	22081448	1.94E-08	0.09507	11.07106	Não há gene candidato nesta região			
	5	27968982	5.19E-07	0.46127	-10.0513	LOC_Os05g48800.2	drought induced 19 protein, putative, expressed	GO:0006950	Processo biológico
								GO:0009628	Processo biológico
	6	4813016	1.11E-08	0.14085	8.225685	Não há gene candidato nesta região			
	10	16018764	6.42E-13	0.38028	-11.3074	LOC_Os10g30750	genomic appr-1-p processing enzyme family protein, putative, expressed	GO:0003674	Função molecular
								GO:0008150	Processo biológico
						LOC_Os03g05660.1	appr-1-p processing enzyme family protein, putative, expressed	GO:0003674	Função molecular
								GO:0008150	Processo biológico
	11	4388206	2.32E-07	0.15141	6.949183	LOC_Os11g08330.1	POLD1 - Putative DNA polymerase delta catalytic subunit, expressed	GO:0003677	Função molecular
								GO:0007049	Processo biológico
								GO:0008152	Processo biológico
								GO:0009058	Processo biológico
								GO:0006259	Processo biológico
								GO:0006950	Processo biológico
								GO:0006139	Processo biológico

Caracter	Cr*	Posição	<i>p</i> -valor	MAF	Efeito	Nome do loco	Descrição	GO_id**	Ontologia
								GO:0003676	Função molecular
								GO:0016740	Função molecular
								GO:0005634	Componente celular
								GO:0000166	Função molecular
								GO:0004518	Função molecular
								GO:0009056	Processo biológico
Caracter	Cr*	Posição	<i>p</i> -valor	MAF	Efeito	Nome do loco	Descrição	GO_id**	Ontologia
CTP	3	33988355	1.42E-10	0.07746	17.59689	LOC_Os03g59680.1	PAPA-1-like conserved region family protein, expressed	GO:0005575	Componente celular
								GO:0008150	Processo biológico
	6	4813016	1.77E-07	0.14085	9.272819	Não há gene candidato nesta região			
	7	29063667	3.87E-08	0.46479	-12.344	Não há gene candidato nesta região			
	9	14109129	1.77E-08	0.11268	-14.4126	LOC_Os07g43300.1	expressed protein	Não encontrada	
	11	17417570	1.77E-08	0.15845	10.62056	Não há gene candidato nesta região			
11	28912694	1.36E-07	0.09859	12.37947	Não há gene candidato nesta região				
MCS	1	28950372	1.68E-06	0.22535	-0.1982	Não há gene candidato nesta região			
	5	26415148	1.19E-15	0.25000	0.512036	LOC_Os05g45580.1	expressed protein		
								GO:0003674	Função molecular
								GO:0005575	Componente celular
								GO:0008150	Processo biológico
	8	544908	1.57E-08	0.15845	-0.31237	Não há gene candidato nesta região			
	10	18169961	3.02E-11	0.11972	-0.5513	LOC_Os10g34040.1	nodulin, putative, expressed	GO:0006950	Processo biológico
								GO:0009987	Processo biológico
								GO:0003674	Função molecular
								GO:0016020	Componente celular
								GO:0005773	Componente celular
								GO:0005575	Componente celular

Caracter	Cr*	Posição	p-valor	MAF	Efeito	Nome do loco	Descrição	GO_id**	Ontologia
	11	4388206	1.29E-06	0.15141	-0.22358	LOC_Os11g08330.1	POLD1 - Putative DNA polymerase delta catalytic subunit, expressed	GO:0003677	Função molecular
								GO:0007049	Processo biológico
								GO:0008152	Processo biológico
								GO:0009058	Processo biológico
								GO:0006259	Processo biológico
								GO:0006950	Processo biológico
								GO:0006139	Processo biológico
								GO:0003676	Função molecular
								GO:0016740	Função molecular
								GO:0005634	Componente celular
								GO:0000166	Função molecular
								GO:0004518	Função molecular
Caracter	Cr*	Posição	p-valor	MAF	Efeito	Nome do loco	Descrição	GO_id**	Ontologia
								GO:0009056	Processo biológico
	11	23422727	0.000207	0.19014	-0.2284	LOC_Os11g39350.1	expressed protein	Não encontrada	
APM	2	35770504	1.02E-08	0.13732	16.1335	Não há gene candidato nesta região			
CC	3	6821863	1.67E-18	0.11268	84.77962	Não há gene candidato nesta região			
	3	32177448	1.81E-09	0.12324	85.06219	LOC_Os03g56460.1	glucose-6-phosphate isomerase, putative, expressed	GO:0003824	Função molecular
								GO:0008150	Processo biológico
								GO:0006950	Processo biológico
								GO:0009607	Processo biológico
								GO:0008152	Processo biológico
								GO:0009058	Processo biológico
								GO:0005975	Processo biológico
								GO:0009987	Processo biológico
								GO:0005829	Componente celular

Caracter	Cr*	Posição	p-valor	MAF	Efeito	Nome do loco	Descrição	GO_id**	Ontologia
	5	4856967	4.20E-10	0.10211	47.71141	LOC_Os05g08810.1	phosphatidylinositol 3-kinase, root isoform, putative, expressed	GO:0006091	Processo biológico
								GO:0009056	Processo biológico
								GO:0008152	Processo biológico
								GO:0009987	Processo biológico
								GO:0009058	Processo biológico
								GO:0006464	Processo biológico
								GO:0016740	Função molecular
								GO:0005488	Função molecular
								GO:0016301	Função molecular
								GO:0007275	Processo biológico
								GO:0006629	Processo biológico
								GO:0005829	Componente celular
								GO:0006950	Processo biológico
								GO:0009628	Processo biológico
								GO:0006810	Processo biológico
								GO:0016043	Processo biológico
	10	16836114	1.19E-09	0.22535	-35.8643	LOC_Os10g32050.1	ankyrin repeat domain containing protein, expressed	GO:0005886	Componente celular
Caracter	Cr*	Posição	p-valor	MAF	Efeito	Nome do loco	Descrição	GO_id**	Ontologia
	12	18040378	5.63E-30	0.14437	137.232	Não há gene candidato nesta região			
	12	18293653	6.14E-24	0.26761	-99.6548	Não há gene candidato nesta região			
NP	2	25515626	1.21E-06	0.49648	8.030243	Não há gene candidato nesta região			
	6	22847044	4.09E-08	0.44718	2.110391	Não há gene candidato nesta região			
LFB	4	21200480	3.79E-07	0.14085	0.122873	Não há gene candidato nesta região			
	5	21271290	1.18E-07	0.07394	0.149747	Não há gene candidato nesta região			
	6	5003718	1.38E-13	0.10915	-0.23358	LOC_Os06g09820.1		GO:0003674	Função molecular

Caracter	Cr*	Posição	<i>p</i> -valor	MAF	Efeito	Nome do loco	Descrição	GO_id**	Ontologia
							cupin superfamily protein,	GO:0005829	Componente celular
							putative, expressed	GO:0008150	Processo biológico
	10	15672084	1.07E-06	0.26408	0.129192	Não há gene candidato nesta região			

*Cr: Cromossomo em que o SNP está inserido

** Go_id: número de acesso GO do banco de dados Gene Ontology

Os genes categorizados funcionalmente segundo sua ontologia podem ser melhor visualizados na figura 7 onde apresenta-se um histograma gerado no WEGO (Web Gene Ontology Annotation Plot) (YE et al. 2006). Nele estão apresentadas comparações no número de genes e porcentagens dos termos GO selecionados. O eixo x exibe os termos GO, o eixo y direito mostra o número de genes dos termos GO selecionados, enquanto o eixo y esquerdo mostra as porcentagens. O projeto “Gene Ontology”, define três categorias: *i)* função molecular, *ii)* processo biológico e *iii)* componente celular. De acordo com os dados obtidos, os marcadores com associação significativa resultaram em 16 genes candidatos com um total de 13 termos GO associados, sendo que a maior parte deles estão relacionados à processos biológicos (7). Com relação à componentes celulares e função molecular, estão envolvidos 4 e 2 genes respectivamente (Figura 11).

Figura 12. Ontologia genética (Termos GO) para os genes candidatos. As barras vermelhas representam a porcentagem e número de genes (eixo Y esquerdo e direito respectivamente) para diferentes classificações de ontologia (Componente Celular – *cellular component* -; Função Molecular – *molecular function* - e Processo Biológico – *Biological Process*) (eixo X); Gerados a partir do Web Gene Ontology Annotation Plot (WEGO).

Os termos GO (Gene Ontology - www.geneontology.org), tem sido amplamente utilizados em projetos de anotação genética, uma vez que reúnem descrições de produtos gênicos em diferentes bancos de dados. O projeto “Gene Ontology” representa um esforço para descrever produtos gênicos bem como suas funções. Neste sentido foram definidas três categorias: *i*) função molecular, *ii*) processo biológico e *iii*) componente celular, onde procura-se entender *i*) em quais processos biológicos o gene está envolvido; *ii*) qual a sua localização no interior da célula e *iii*) como o gene realiza a sua função. Por exemplo, determinado gene pode realizar uma transdução de sinal em uma célula estando assim associado a um processo biológico; estar preso à membrana celular, portanto sua localização seria celular e ter como função fosforilar outra proteína tendo portanto função molecular, assim, este gene seria classificado no GO de acordo com as três características.

Considerando-se os marcadores SNPs foram obtidos diversos genes com anotação no IRGSP:

1-NDE

Foi encontrado uma marca significativa: SNP M18689, localizado junto ao gene LOC_Os05g51180.1 identificado com funções moleculares relacionada a ligação de RNA (GO:0003723), componente celular do núcleo (GO:0005634) e citoplasma (GO:0005737) bem como a processos biológicos (GO:0008150).

2-NDF

O SNP M13383 localizado no gene LOC_Os03g32590.1, foi identificado como parte de processos biológicos (GO:0006139 e GO:0009058), tendo sua função putativa como fator de iniciação da transcrição. M18607 – gene: LOC_Os05g48800.2 - processos biológicos (GO:0006950 e GO:0009628) relacionando-se à respostas a estresses abióticos. M4447 - gene: LOC_Os10g30750 – função molecular (GO:0003674) e processos biológicos (GO:0008150) estando envolvido com o processamento de enzimas da família “*genomic|appr-1-p*”. Esta é uma importante enzima de processamento celular envolvida na via de processamento de tRNA, catalisa a conversão de ADP-ribose-1 'monofosfato (Apr-1 "-p) em ADP-ribose (Kumaran, et al. 2005). M5323 – gene: LOC_Os11g08330.1 – função molecular (GO:0003677; GO:0003676 e GO:0016740) relacionados a ligação de ácidos nucleicos e atividade transferase; Processos biológicos (GO:0007049; GO:0008152; GO:0009058; GO:0006259; GO:0006950; GO:0006139) relacionados a ciclos celulares, processos metabólicos e biossintéticos e resposta a estresses.

3-CTP

M14337 – gene: LOC_Os03g59680.1 – Componente celular (GO:0005575) e processos biológicos (GO:0008150). M26093 – gene: LOC_Os07g43300.1 Não foi encontrada classificação ontológica.

4- MCS

M18495 – gene: LOC_Os05g45580.1- Função molecular (GO:0003674); componente celular (GO:0005575) e processos biológicos (GO:0008150) sendo responsável pela expressão de proteínas. M4602 - gene: LOC_Os10g34040.1 – Processos biológicos (GO:0006950 e GO:0009987) relacionados a resposta à estresses; função molecular (GO:0003674) e componentes celulares (GO:0016020; GO:0005773 e GO:0005575) relacionados à formação de membranas e vacúolo. M5323 – gene: LOC_Os11g08330.1 – Função molecular (GO:0003677; GO:0003676; GO:0016740; GO:0000166 e

GO:0004518), relacionada a ligação de ácidos nucleicos e atividades de nucleases; Processos biológicos relacionados a processos de catabolismo (GO:0007049; GO:0008152; GO:0009058; GO:0006259; GO:0006950; GO:0006139 e GO:0009056). M6831 – gene: LOC_Os11g39350.1 - Não foi encontrada classificação ontológica.

5 – CC

M14183 – gene: LOC_Os03g56460.1 – Função molecular (GO:0003824) atividade catalítica; Processos biológicos (GO:0008150; GO:0006950; GO:0009607; GO:0008152; GO:0009058; GO:0005975; GO:0009987; GO:0005829; GO:0006091; GO:0009056) relacionados à Glicose-6-phosphato-isomerase. M17174 – gene: LOC_Os05g08810.1 – processos biológicos (GO:0008152; GO:0009987; GO:0009058; GO:0006464; GO:0007275; GO:0006629; GO:0005829; GO:0006950; GO:0009628; GO:0006810 e GO:0016043) e funções moleculares (GO:0016740; GO:0005488 e GO:0016301) tendo suas funções relacionadas com o crescimento radicular (fosfatidilinositol 3-quinase) e componentes celulares (GO:0005829). M4510 - gene: LOC_Os10g32050.1 – Componente da membrana celular (GO:0005886).

6-LFB

M19182 – gene: LOC_Os06g09820.1 –Função molecular (GO:0003674); Componente celular (GO:0005829) e processo biológico (GO:0008150).

Não foram encontradas funções associadas nos bancos de dados pesquisados para os caracteres NP e AMP.

Marcadores SNPs ocorrem com grande frequência nos genomas em geral. No arroz pode ser observado em média um SNP em 140 pares de bases (IRGPS, 2005), Além disso, de acordo com ZHU et al. (2008), marcadores co-dominantes (como no caso dos SNPs) são mais informativos do que os dominantes, uma vez que apresentam maior poder em estimar a estrutura populacional (Q) e a matriz de parentesco (K).

Os SNPs se tornaram os marcadores mais utilizados nos estudos de caracteres complexos (ZHU et al. 2008). Estudos de associação utilizando esse tipo de marcador são promissores uma vez que grandes números de SNPs vem sendo identificados devido aos avanços no uso de técnicas de sequenciamento por genotipagem. As principais vantagens dos marcadores SNPs são que eles ocorrem em genomas a uma frequência muito grande, com cerca de um SNP sendo observado por 140 pb no arroz (IRGPS, 2005), detectando uma grande quantidade de polimorfismos entre alelos de determinado gene. Sendo mais reveladores (ou seja, nenhuma ambiguidade alélica) do que os marcadores dominantes e, portanto, mais poderosos na estimativa da estrutura da população (Q) e da matriz de parentesco (K) (ZHU et al. 2008). Shing et al. (2013), comparou a utilização de marcadores SSR e SNPs no estudo de populações, indicando que marcadores SNP devem ser utilizados nos estudos de estrutura de populações. Geralmente os SNPs são derivados de genes portanto, podem ser uma boa ferramenta para estudos de associação entre marcador e característica, além do mais, quando se fala especificamente de arroz, os SNPs derivam de regiões genômicas que apresentam sintonia com o trigo (*Triticum sp.*), sendo assim podem ser úteis para estudos de outros cereais. Em outras palavras, SNPs são de grande importância para abordagens como o mapeamento associativo e seleção genômica. No entanto, um número maior de SNPs poderia gerar informações mais detalhadas dos genótipos estudados. Embora a filtragem e imputação de dados resulte em grande qualidade de marcas,

7989 SNPs é um número relativamente baixo para o mapeamento associativo se comparado com outros trabalhos como o PANTALIÃO et al. (2016), que encontrou mais de 150 mil SNPs utilizando-os no mapeamento associativo para produtividade encontrando 26 associações em arroz e BISCARINI et al. (2016) que utilizou 57000 SNPs, encontrando 42 associações em estudo de características morfológicas da planta, do grão e de raiz em acessos de clima temperado de arroz.

3.4 Conclusão

O presente estudo demonstra o potencial de marcadores SNP derivados de DArTseq altamente informativos e seletivos para análise de diversidade genética e estudos de associação genômica ampla nos genótipos de arroz testados. Um total de 33 SNP's foram identificados associados a oito caracteres quantitativos avaliados em 142 acessos de arroz. A população estudada pode ser subdividida em 3 grupos em função da análise discriminante de componentes principais.

Referências

- ADETUNJI, I. et al. Genetic diversity and linkage disequilibrium analysis in elite sugar beet breeding lines and wild beet accessions. **Theoretical and Applied Genetics**. v. 127, n. 3, p. 559–571, 2014.
- ARANZANA, M. J. et al. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. **PLoS Genetics**. v. 1, n. 3, e60. p. 531-539, 2005.
- BISCARINI, F. et al. Genome-wide association study for traits related to plant and grain morphology, and root architecture in temperate rice accessions. **PLoS ONE**, v. 11, n. 5, p. 1–28, 2016.
- CHEN, H.; PATTERSON, N.; REICH, D. Population differentiation as a test for selective sweeps. **Genome Research**. v. 20, p. 393–402. 2010. doi:10.1101/gr.100545.109
- COURTOIS, B. et al. Genome-wide association mapping of root traits in a japonica rice panel. **PLoS ONE**, v. 8, n. 11, p. 1–18, 2013.
- DOYLE, J.J.; DOYLE, J.L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, v.19, p.11-15, 1987.
- FLINT-GARCIA, S. A.; THORNSBERRY, J. M.; BUCKLER, E. S. Structure of Linkage Disequilibrium in Plants. **Annual Review of Plant Biology**, v. 54, n. 1, p. 357–374, 2003.
- GARRIS, A. S.; MCCOUCH, S.; KRESOVICH, S. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.). **Genetics**, v. 165, n. 2, p. 759–769, 2005.
- GUPTA, P. K.; RUSTGI, S.; KULWAL, P. L. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. **Plant Molecular Biology**, 2005. Disponível em: <<https://pdfs.semanticscholar.org/c07c/539d6b84b92a0aa8b7f4780a2b2566f00c76.pdf>>. Acesso em: 13 jun. 2018.
- HILL, W. G.; ROBERTSON, A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**. v. 38, n. 6, p. 226–231, 1968.
- IRGPS. The map-based sequence of the rice genome. **Nature**, v. 436, n. 7052, p. 793–800, 11 ago. 2005.
- KATO, S.; KOSAKA H.; HARA, S. On the affinity of rice varieties as show by fertility of hybrid plants. Fukuoka: Faculty Agricultural, Kyushu University, 1928. p.132-147. (Bulletin Scientific, 3)

- KILIAN, A. et al. Diversity arrays technology: a generic genome profiling technology on open platforms. **Methods in molecular biology**, v. 888, p. 67–89, 2012.
- KUMARAN, D. et al. Structure and mechanism of ADP-ribose-1"-monophosphatase (Appr-1"-pase), a ubiquitous cellular processing enzyme. **Protein Science : A Publication of the Protein Society**. v. 14, n. 3, p. 719-726, 2005.
- KUSH, G. S.; BRAR, D. S. Biotechnology for rice breeding: progress and potential impact. FAO: The international rice commission. 2002. Available at: <http://www.fao.org>. Acesso em: 13 de junho de 2018.
- LIPKA, A. E. et al. GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, v. 28, n. 18, p. 2397–2399, 2012.
- LIU, X. et al. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. **PLoS Genetics**. v. 12, n. 2, e1005767, 2016.
- LOISELLE, B. A. et al. Spatial Genetic Structure of a Tropical Understory Shrub , *Psychotria officinalis* (Rubiaceae)
Author (s): Bette A . Loiselle , Victoria L . Sork , John Nason , Catherine Graham Published by : Botanical Society of America Stable URL : <http://www.jstor>. **America**, v. 82, n. 11, p. 1420–1425, 2008.
- LONDO, J. P. et al. Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. **Proceedings of the National Academy of Sciences**, v. 103, n. 25, p. 9578–9583, 2006.
- MANGIN, B. et al. Breeding patterns and cultivated beets origins by genetic diversity and linkage disequilibrium analyses. **Theoretical and Applied Genetics**. Springer Berlin Heidelberg; v. 128: p. 2255–2271, 2015.
- MATA, T. L. da., Diversidade genética em Germoplasma de arroz filipino identificada por marcadores moleculares e caracteres agromorfológicos. Dissertação (Mestrado). Universidade de São paulo. Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, 2010. 99p.
- PANTALIAO, G. F. et al. Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. **Genetica**, v. 144, n. 6, p. 651–664, 2016.
- PRICE, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics**, v. 38, n. 8, p. 904–909, 2006.
- R CORE TEAM (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>. 2018.
- RICHARDS, J. K.; FRIESEN, T. L.; BRUEGGEMAN, R. S. Association mapping utilizing diverse barley lines reveals net form net blotch seedling resistance/susceptibility loci. **Theoretical and Applied Genetics**, v. 130, n. 5, p. 915–927, 2017.
- SINGH, N. et al. Comparison of SSR and SNP Markers in Estimation of Genetic Diversity and Population Structure of Indian Rice Varieties. **PLoS ONE**, v. 8, n. 12, p. e84136, 2013.
- SLATKIN, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. **Nature Review Genetics**, v. 9, p. 477–485, 2008.
- SWEENEY, M. T. et al. Caught redhanded: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. **The Plant Cell**, v. 18, p. 283 – 294. 2006.
- TANG, Y. et al. GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. **The Plant Genome**, v. 9, n. 2, p. 0, 2016.

- TOMICIC, M. et al. Tracking footprints of selection associated with soybean adaptation to Central-East Europe environments. **Euphytica**. Springer Netherlands; n. 203, p. 701–713. 2015.
- VAUGHAN, D.A. **The genus *Oryza* L. Current status of taxonomy**. Manila: IRRI, 1989. 21p.
- VOLANTE, A. et al. Genome-Wide Analysis of *japonica* Rice Performance under Limited Water and Permanent Flooding Conditions. **Frontiers in Plant Science**. v. 30, n. 8, doi: 10.3389/fpls.2017.01862. 2017.
- WEISS, K. M.; CLARK, A. G. Linkage disequilibrium and the mapping of complex human traits. **Trends Genetics**, v. 18, p. 19–24. 2002.
- WÜRSCHUM, T. et al. Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. **Theoretical and Applied Genetics**, v. 126, n. 6, p. 1477–1486, 2013.
- XU, Z. J. et al. Correlations between rice grain shapes and main qualitative characteristics. **Acta Agronomica Sinica**, v. 30, n. 9, p. 894–900, 2004.
- YE J. et al. Nucleic Acids Research, v. 34 (Web service issue), p. 293–297, 2006.
- YU, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**, v. 38, n. 2, p. 203–208, 2006.
- ZHANG, D. et al. Genetic structure and differentiation of *Oryza sativa* L. in China revealed by microsatellites. **Theoretical and Applied Genetics**, v. 119, n. 6, p. 1105–1117, 2009.
- ZHANG, P. et al. Association mapping for important agronomic traits in core collection of rice (*Oryza sativa* L.) with SSR markers. **PLoS ONE**, v. 9, n. 10, p. 16, 2014.
- ZHAO, K. et al. An Arabidopsis example of association mapping in structured samples. **PLoS Genetics**, v. 3, n. 1, p. 0071–0082, 2007.
- ZHU, C. et al. Status and Prospects of Association Mapping in Plants. **The Plant Genome Journal**, v. 1, n. 1, p. 5, 2008.
- ZONG Y. et al. Fire and flood management of coastal swamp enabled first rice paddy cultivation in east China. **Nature**, v. 449, 459–462, 2007.

ANEXO

ANEXO A. ANEXO A Acessos de arroz (*Oryza sativa*) pertencentes ao Banco de Germoplasma de arroz cultivado do Departamento de Genética da ESALQ/USP e cultivares comerciais utilizadas como testemunhas no presente projeto (Continua)

Número do Acesso	Subgrupo	Origem	Nome
BGA198	1	Brasileiro	s/ nome
BGA199	1	Brasileiro	s/ nome
BGA200	1	Brasileiro	s/ nome
BGA201	1	Brasileiro	s/ nome
BGA202	1	Brasileiro	s/ nome
BGA203	1	Brasileiro	s/ nome
BGA204	1	Brasileiro	s/ nome
BGA205	1	Brasileiro	s/ nome
BGA206	1	Brasileiro	s/ nome
BGA207	1	Brasileiro	s/ nome
BGA208	1	Brasileiro	s/ nome
BGA209	1	Brasileiro	s/ nome
BGA210	1	Brasileiro	s/ nome
BGA211	1	Brasileiro	s/ nome
BGA212	1	Brasileiro	s/ nome
BGA213	1	Brasileiro	s/ nome
BGA214	1	Brasileiro	s/ nome
BGA215	1	Brasileiro	s/ nome
BGA216	1	Brasileiro	s/ nome
BGA217	1	Brasileiro	s/ nome
BGA218	1	Brasileiro	s/ nome
BGA219	1	Brasileiro	s/ nome
BGA220	1	Brasileiro	s/ nome
BGA222	1	Brasileiro	s/ nome
BGA223	1	Brasileiro	s/ nome
BGA224	1	Brasileiro	s/ nome
BGA225	1	Brasileiro	s/ nome
BGA226	1	Brasileiro	s/ nome
BGA227	1	Brasileiro	s/ nome
BGA229	1	Brasileiro	s/ nome
BGA230	1	Brasileiro	s/ nome
BGA231	1	Brasileiro	s/ nome
BGA232	1	Brasileiro	s/ nome
BGA233	1	Brasileiro	s/ nome
BGA234	1	Brasileiro	s/ nome
BGA235	1	Brasileiro	s/ nome
BGA236	1	Brasileiro	s/ nome
BGA237	1	Brasileiro	s/ nome
BGA238	1	Brasileiro	s/ nome

Número do Acesso	Subgrupo	Origem	Nome
BGA239	1	Brasileiro	s/ nome
BGA240	1	Brasileiro	s/ nome
BGA241	1	Brasileiro	s/ nome
BGA242	1	Brasileiro	s/ nome
BGA243	1	Brasileiro	s/ nome
BGA244	1	Brasileiro	s/ nome
BGA245	1	Brasileiro	s/ nome
BGA246	1	Brasileiro	s/ nome
BGA247	1	Brasileiro	s/ nome
BGA248	1	Brasileiro	s/ nome
BGA249	1	Brasileiro	s/ nome
BGA250	1	Brasileiro	s/ nome
BGA251	1	Brasileiro	s/ nome
BGA252	1	Brasileiro	s/ nome
BGA253	1	Brasileiro	s/ nome
BGA254	1	Brasileiro	s/ nome
BGA255	1	Brasileiro	s/ nome
BGA256	1	Brasileiro	s/ nome
BGA257	1	Brasileiro	s/ nome
BGA258	1	Brasileiro	s/ nome
BGA259	1	Brasileiro	s/ nome
BGA260	1	Brasileiro	s/ nome
BGA261	1	Brasileiro	s/ nome
BGA262	1	Brasileiro	s/ nome
BGA263	1	Brasileiro	s/ nome
BGA264	1	Brasileiro	s/ nome
BGA265	1	Brasileiro	s/ nome
BGA266	1	Brasileiro	s/ nome
BGA267	1	Brasileiro	s/ nome
BGA268	1	Brasileiro	s/ nome
BGA269	1	Brasileiro	s/ nome
BGA270	1	Brasileiro	s/ nome
BGA271	1	Brasileiro	s/ nome
BGA272	1	Brasileiro	s/ nome
BGA273	1	Brasileiro	s/ nome
BGA274	1	Brasileiro	s/ nome
BGA275	1	Brasileiro	s/ nome
BGA277	1	Brasileiro	s/ nome
BGA278	1	Brasileiro	s/ nome
BGA280	1	Brasileiro	s/ nome
BGA281	1	Brasileiro	s/ nome
BGA282	1	Brasileiro	s/ nome
BGA283	1	Brasileiro	s/ nome
BGA284	1	Brasileiro	s/ nome

Número do Acesso	Subgrupo	Origem	Nome
BGA285	1	Brasileiro	s/ nome
BGA286	1	Brasileiro	s/ nome
BGA287	1	Brasileiro	s/ nome
BGA290	1	Brasileiro	s/ nome
BGA292	1	Brasileiro	s/ nome
BGA293	1	Brasileiro	s/ nome
BGA294	1	Brasileiro	s/ nome
BGA295	1	Brasileiro	s/ nome
BGA296	1	Brasileiro	s/ nome
BGA297	1	Brasileiro	s/ nome
BGA298	1	Brasileiro	s/ nome
BGA299	1	Brasileiro	s/ nome
BGA300	1	Brasileiro	s/ nome
BGA301	1	Brasileiro	s/ nome
BGA302	1	Brasileiro	s/ nome
BGA303	1	Brasileiro	s/ nome
BGA304	1	Brasileiro	s/ nome
BGA305	1	Brasileiro	s/ nome
BGA306	1	Brasileiro	s/ nome
BGA307	1	Brasileiro	s/ nome
BGA308	1	Brasileiro	s/ nome
BGA309	1	Brasileiro	s/ nome
BGA310	1	Brasileiro	s/ nome
BGA311	1	Brasileiro	s/ nome
BGA313	1	Brasileiro	s/ nome
BGA314	1	Brasileiro	s/ nome
BGA315	1	Brasileiro	s/ nome
BGA316	1	Brasileiro	s/ nome
BGA318	1	Filipino	Rogur(464799)420
BGA319	1	Filipino	3558
BGA320	1	Filipino	Agbede
BGA322	1	Filipino	Ku 55-1
BGA323	1	Filipino	Ku 91
BGA324	1	Filipino	ku 58
BGA325	1	Filipino	Djoro One
BGA326	1	Filipino	Ketan Tlasih
BGA328	1	Filipino	Tangun
BGA329	1	Filipino	Kyetek
BGA331	1	Filipino	Sawak
BGA332	1	Filipino	Putu
BGA335	1	Filipino	Menurun
BGA336	1	Filipino	Majangan
BGA337	1	Filipino	Lembese
BGA342	1	Filipino	Iac-5100

Número do Acesso	Subgrupo	Origem	Nome
BGA343	1	Filipino	Ligerito
BGA344	1	Filipino	Iac-5544
BGA351	1	Filipino	Khao Kap Xang
BGA352	1	Filipino	Jao Khao
BGA363	1	Filipino	Ngah Chahng
BGA389	1	Filipino	Pah Yan
BGA394	1	Filipino	Khao Chao Met Nhay
BGA397	1	Filipino	khao Namma
BGA401	1	Filipino	Khao Phe Py
BGA402	1	Filipino	Khao Xiou khay
BGA407	1	Filipino	Daw Dan
BGA415	1	Filipino	Cabaysay
BGA416	1	Filipino	Lingkod
BGA417	1	Filipino	Bikyat
BGA419	1	Filipino	Mangglutus
BGA420	1	Filipino	tumendog
BGA421	1	Filipino	Simping
BGA423	1	Filipino	Carcata
BGA424	1	Filipino	Finya
BGA425	1	Filipino	Gette
BGA426	1	Filipino	Cokoba
BGA430	1	Filipino	Liberian Coll. D3-101
BGA432	1	Filipino	Kpanpolohombon
BGA433	1	Filipino	Liberian Coll. D3-167
BGA434	1	Filipino	Gbegbea
BGA435	1	Filipino	Ngelegohun
BGA436	1	Filipino	Gbegbete
BGA437	1	Filipino	Felue
BGA438	1	Filipino	Sole Yoe
BGA439	1	Filipino	Bogutu
BGA440	1	Filipino	Bibilial
BGA443	1	Filipino	Melati
BGA444	1	Filipino	Menurun
BGA446	1	Filipino	Iac-47
BGA447	1	Filipino	Iac-1131
BGA449	1	Filipino	Wittie Sada(Pi-3699815)
BGA453	1	Filipino	Ku 79-1
BGA456	1	Filipino	Jeteh
BGA457	1	Filipino	Selibon Siangan
BGA460	1	Filipino	Gandjah Kutu
BGA462	1	Testemunha	BRSPampa
BGA463	1	Testemunha	ANCambará
BGA464	1	Testemunha	Esmeralda
BGA465	1	Testemunha	Setaneja

Número do Acesso	Subgrupo	Origem	Nome
BGA317	2	Filipino	Br-br-r(C6-1-25-12)
BGA321	2	Filipino	ku 43-2
BGA327	2	Filipino	Ketan Pulosaren
BGA333	2	Filipino	Lokan
BGA334	2	Filipino	Ketji
BGA340	2	Filipino	Metan
BGA348	2	Filipino	Puntas Claras
BGA349	2	Filipino	Khao Pio
BGA350	2	Filipino	Khao Sumneua
BGA353	2	Filipino	Khao Chao Hom
BGA354	2	Filipino	Daw Hen
BGA356	2	Filipino	Leuang Noi
BGA357	2	Filipino	Khao Kai
BGA359	2	Filipino	Khao Mack Fay Khao
BGA360	2	Filipino	Khao Khay
BGA361	2	Filipino	Khao Noi
BGA362	2	Filipino	Khao Bangkhao
BGA364	2	Filipino	KhaoDok Dou
BGA365	2	Filipino	khao None Nhay
BGA367	2	Filipino	khao Nuan
BGA368	2	Filipino	Khao Dam
BGA369	2	Filipino	Mai Kai
BGA370	2	Filipino	Khao Kangkaynoi
BGA371	2	Filipino	Khao Nay Nock
BGA372	2	Filipino	Khao Mone Khan
BGA375	2	Filipino	Khaomu
BGA376	2	Filipino	Gaen Fai
BGA377	2	Filipino	Khao Phe Do
BGA378	2	Filipino	Khao Eo Nhay
BGA379	2	Filipino	Khao Do
BGA380	2	Filipino	Khao Dohome
BGA382	2	Filipino	Khao Pick
BGA383	2	Filipino	Jan
BGA384	2	Filipino	Khao Hido
BGA385	2	Filipino	Khao Dock Keonoi
BGA387	2	Filipino	Pang Leu
BGA388	2	Filipino	Khae
BGA390	2	Filipino	Khao Dok Pout
BGA391	2	Filipino	Ma Hing 267-2-18
BGA392	2	Filipino	Mawai
BGA395	2	Filipino	Khao Seng
BGA396	2	Filipino	Hahng
BGA398	2	Filipino	Khao Toun
BGA399	2	Filipino	Khao Xien Py

Número do Acesso	Subgrupo	Origem	Nome
BGA400	2	Filipino	Khao Xiou Khao
BGA403	2	Filipino	Khao Vanthang
BGA404	2	Filipino	Khao Metto B
BGA405	2	Filipino	Lai Daek Doo
BGA406	2	Filipino	Khao Met Nhay
BGA428	2	Filipino	Fogbandi
BGA445	2	Filipino	ku 101
BGA450	2	Filipino	Ku 10
BGA451	2	Filipino	Ku-94-2
BGA452	2	Filipino	ku 86
BGA454	2	Filipino	ku 101
BGA455	2	Filipino	ku 70-1
BGA291	3	Brasileiro	s/ nome
BGA312	3	Brasileiro	s/ nome
BGA431	3	Filipino	Bulahon
BGA458	3	Filipino	Hill Sel. X Rnbmt54v
BGA197	4	Brasileiro	s/ nome
BGA221	4	Brasileiro	s/ nome
BGA276	4	Brasileiro	s/ nome
BGA330	4	Filipino	Sibakas
BGA338	4	Filipino	Maraja
BGA341	4	Filipino	Majang Djanggut
BGA345	4	Filipino	Tjempo Tsino
BGA346	4	Filipino	Ba Djang Plon
BGA347	4	Filipino	Ba ke Gonh
BGA355	4	Filipino	Khao Kikhouei
BGA358	4	Filipino	Khao Nhouak Nhay
BGA366	4	Filipino	E-naw
BGA373	4	Filipino	Khao Khane
BGA374	4	Filipino	Khao Maet Nhay
BGA381	4	Filipino	Ja Ae
BGA386	4	Filipino	Khao Nhn Py
BGA393	4	Filipino	Khao Chao Hay
BGA408	4	Filipino	Mum1
BGA409	4	Filipino	E-boot
BGA410	4	Filipino	Graboon
BGA411	4	Filipino	khao sim
BGA412	4	Filipino	Sew Glang Dong
BGA413	4	Filipino	Wiang
BGA414	4	Filipino	Unnamed
BGA418	4	Filipino	Kalinayan
BGA422	4	Filipino	Poyeh
BGA429	4	Filipino	Liberian Coll. D3-64
BGA441	4	Filipino	Bata sunduwee

Número do Acesso	Subgrupo	Origem	Nome
BGA442	4	Filipino	Bandan-2
BGA448	4	Filipino	Ku-47
BGA459	4	Filipino	Os.6
BGA461	4	Testemunha	BRS Primavera
BGA228	5	Brasileiro	s/ nome
BGA466	5	Testemunha	Epagri109
BGA467	5	Testemunha	TioTaka
BGA468	5	Testemunha	116Satoru
BGA469	5	Testemunha	Andosan
BGA470	5	Testemunha	Markes

As informações de código e nome do acessos são de registros do banco de germoplasma de arroz do Departamento de Genética da ESALQ/USP.

* Refere-se a coleta de variedades em território nacional onde não se tem o registro dos nomes locais.