UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA

MAPEAMENTO ASSOCIATIVO PARA PRODUTIVIDADE EM ARROZ SOB DÉFICIT HÍDRICO

GABRIEL FERESIN PANTALIÃO

Orientador:
Dr. Claudio Brondani

Goiânia - GO
Brasil

Março - 2013

GABRIEL FERESIN PANTALIÃO

MAPEAMENTO ASSOCIATIVO PARA PRODUTIVIDADE EM ARROZ SOB DÉFICIT HÍDRICO

Dissertação apresentada ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, da Universidade Federal de Goiás, como exigência para obtenção do título de Mestre em Genética e Melhoramento de Plantas.

Orientador:
Dr. Claudio Brondani

Goiânia, GO - Brasil
2013

Dedico este trabalho aos meus pais e à minha família que sempre me apoiaram em todas as minhas decisões, acreditaram no meu potencial e me deram força para a conclusão de mais uma etapa em minha vida. Expresso aqui meu eterno amor e gratidão a vocês.

AGRADECIMENTOS

Agradeço a Deus, primeiramente, por todas coisas boas que trouxe em minha vida, além de obstáculos que tive que ultrapassar e que garantiram meu amadurecimento profissional e pessoal. Aos meus pais que estiveram sempre disponíveis para qualquer conselho, apoiando todos passos que eu dava. À minha avó e minha tia que sempre estiveram presentes nos meus momentos de tristeza e alegria. E a todos meus amigos que de maneira individual torceram para o meu sucesso.

Ao meu orientador, Dr. Claudio Brondani, que me mostrou como é fácil aliar uma boa convivência com resultados produtivos, ensinando não somente conhecimentos profissionais como pessoais, que auxiliaram de forma efetiva meu amadurecimento. À Dra. Tereza Cristina de Oliveira Borba, pelo companheirismo, paciência e disposição, que auxiliou muito na minha chegada até aqui. Ao Dr. Marcelo Narciso pela prestatividade com as análises computacionais. Ao Dr. Cleber Guimarães pela condução do experimento em Porangatu (GO). À Dra. Luíce Gomes Bueno, pela prestatividade e auxílio com as análises dos dados de campo. Ao futuro doutor Ricardo Diógenes, pela amizade e ajuda em diversos momentos. Aos demais colegas do Laboratório de Biotecnologia, pela convivência e pela amizade de todos.

À Embrapa Arroz e Feijão, pelas dependências. À CAPES, pelo financiamento do projeto no qual está inserido meu trabalho e pela concessão da bolsa de mestrado. À Universidade Federal de Goiás, pela oportunidade de aprimoramento da minha formação acadêmica. E a todos os professores do programa pelos conhecimentos adquiridos.

SUMÁRIO

LISTA DE TABELAS 06

LISTA DE FIGURAS 07

RESUMO 08

ABSTRACT 09

1 INTRODUÇÃO 10

2 REVISÃO BIBLIOGRÁFICA 12
2.1 O GÊNERO *ORYZA* 12
2.1.1 A espécie *O. sativa* 13
2.2 COLEÇÃO NUCLEAR DE ARROZ DA EMBPRAPA (CNAE) 14
2.3 O GENOMA DO ARROZ CULTIVADO (*Oryza sativa* L.) 16
2.4 PRINCÍPIOS GERAIS DO MAPEAMENTO ASSOCIATIVO 17
2.5 MAPEAMENTO ASSOCIATIVO E MAPEAMENTO DE QTL 18
2.6 ABORDAGENS DO MAPEAMENTO ASSOCIATIVO 19
2.7 SEQUENCIAMENTO DE NOVA GERAÇÃO (NGS) 21
2.8 ANÁLISE GENÉTICA BASEADA EM NGS 23
2.9 TOLERÂNCIA À SECA EM PLANTAS 25

3 MATERIAL E MÉTODOS 28
3.1 MATERIAL VEGETAL 28
3.2 CARACTERIZAÇÃO DO PAINEL DE MAPEAMENTO ASSOCIATIVO QUANTO AO DÉFICIT HÍDRICO 28
3.3 ANÁLISE DO CONJUNTO DE DADOS FENOTÍPICOS 30
3.3.1 Ajuste dos dados fenotípicos 30
3.3.2 Distribuição de frequências e dispersão 30
3.4 DADOS GENOTÍPICOS 31
3.5 ESTRUTURA GENÉTICA POPULACIONAL 32

3.6 ANÁLISE DE ASSOCIAÇÃO 32
3.7 ANOTAÇÃO DOS GENES 33

4 RESULTADOS 34
4.1 AVALIAÇÃO FENOTÍPICA DOS ACESSOS 34
4.1.1 Conjunto de dados 34
4.1.2 Produtividade e suscetibilidade à seca 38
4.2 DADOS GENOTÍPICOS 40
4.3 ESTRUTURA GENÉTICA POPULACIONAL 41
4.4 ANÁLISE DE ASSOCIAÇÃO 42
4.4.1 Distribuição dos valores de significância 45
4.5 ANOTAÇÃO DOS GENES 46

5 DISCUSSÃO 51
5.1 ANÁLISE DOS CARACTERES FENOTÍPICOS 51
5.2 DADOS GENOTÍPICOS 54
5.3 ANÁLISE DE ASSOCIAÇÃO 55
5.4 ANOTAÇÃO DOS GENES 57

6 CONCLUSÕES 60

7 REFERÊNCIAS BIBLIOGRÁFICAS 61

LISTA DE TABELAS

Tabela 1. Coleção Nuclear de Arroz da Embrapa (CNAE) dividida pelos estratos e número de acessos que fazem parte de cada estrato **16**

Tabela 2. Método dos Quadrados Mínimos para a variável Produtividade (kg ha^{-1}) em ambiente com deficiência hídrica **34**

Tabela 3. Método dos Quadrados Mínimos para a variável Produtividade (kg ha^{-1}) em ambiente sem deficiência hídrica **34**

Tabela 4. Vinte materiais mais produtivos do ambiente com deficiência hídrica **39**

Tabela 5. Vinte materiais mais produtivos do ambiente sem deficiência hídrica **39**

Tabela 6. Vinte materiais mais sensíveis à seca no experimento de 2010 ..... **40**

Tabela 7. Número de marcadores SNPs obtidos pela genotipagem de 283 acessos de arroz de terras altas da CNAE por GBS **41**

Tabela 8. Associações presentes em cada um dos 12 cromossomos **42**

Tabela 9. Associações (SNPs - caractere) e suas significâncias (Cromossomos 1 ao 3) ... **43**

Tabela 10. Associações (SNPs - caractere) e suas significâncias (Cromossomos 4 ao 10) **44**

Tabela 11. Transcritos identificados em cada cromossomo do arroz cultivar Nipponbare **47**

Tabela 12. Genes identificados a partir da presença de marcadores SNPs provenientes da análise de mapeamento associativo e seus respectivos produtos expressos **48**

Tabela 13. Genes identificados a partir da presença de marcadores SNPs provenientes da análise de mapeamento associativo e os prováveis processos biológicos de seus transcritos **49**

Tabela 14. Genes identificados a partir da presença de marcadores SNPs provenientes da análise de mapeamento associativo e as prováveis funções moleculares de seus transcritos **50**

LISTA DE FIGURAS

Figura 1. Distribuição de frequência da produtividade dos acessos do experimento em ambiente com deficiência hídrica **35**

Figura 2. Distribuição de frequência da produtividade dos acessos do experimento em ambiente sem deficiência hídrica **36**

Figura 3. Distribuição de frequência do Índice de Suscetibilidade à Seca para produtividade no experimento com deficiência hídrica **36**

Figura 4. Gráfico boxplot para a produtividade nos ambientes com e sem deficiência hídrica **37**

Figura 5. Gráfico boxplot para o Índice de Suscetibilidade à Seca **38**

Figura 6. Gráfico Manhattan Plot dos valores de significância das associações nos cromossomos 1 ao 6 ..... **45**

Figura 7. Gráfico Manhattan Plot dos valores de significância das associações nos cromossomos 7 ao 12 ..... **46**

RESUMO

PANTALIÃO, G. F. **Mapeamento Associativo para produtividade em arroz sob déficit hídrico.** 2013. 91 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2013.[1]

A seca é um fator ambiental que limita a produção das culturas, como a do arroz de terras altas (*Oryza sativa* L.). O conhecimento de fatores envolvidos na tolerância à deficiência hídrica e das respostas das plantas a esse estresse podem fornecer subsídios aos programas de melhoramento para o desenvolvimento de cultivares tolerantes, e, consequentemente, com uma maior produtividade sob essas condições. O mapeamento associativo, ou análise de associação, tem sido aplicado com sucesso em plantas, sendo utilizado primeiramente na identificação de genes associados a caracteres de importância econômica, e posteriormente, na implementação de seleção assistida por marcadores (SAM). Tecnologias de sequenciamento de nova geração (NGS) têm sido recentemente utilizadas em projetos de sequenciamento e resequenciamento para identificar, validar e avaliar um grande número de SNPs, os quais podem ser utilizados em estudos de mapeamento associativo. Dentre os métodos desenvolvidos para a descoberta de marcadores moleculares e genotipagem de alto desempenho, destaca-se pela rapidez e baixo custo a genotipagem por sequenciamento (GBS). Esse trabalho objetivou detectar, via GBS, o polimorfismo de marcadores SNPs em 283 acessos de arroz de terras altas componentes da CNAE (Coleção Nuclear de Arroz da Embrapa) e associá-los à produtividade sob déficit hídrico. Após a filtragem dos dados brutos de acordo com parâmetros de estringência pré-definidos, foram contabilizados 285.379 SNPs distribuídos ao longo dos 12 cromossomos do arroz. As informações moleculares foram integradas aos dados fenotípicos derivados do experimento de avaliação de produtividade e Índice de Suscetibilidade à Seca (ISS), conduzido no ano de 2010 em Porangatu (GO) em ambiente com e sem deficiência hídrica, para possibilitar a análise de mapeamento associativo, e com isso, detectar marcadores SNPs relacionados à tolerância à seca e oportunizar o desenvolvimento de um conjunto de marcadores úteis para a seleção assistida para esse caráter, assim como genes para estudos de engenharia genética do arroz. Através da análise de associação, foram detectados 48 SNPs relacionados com os caracteres avaliados, dentre os quais 13 foram relacionados ao ISS e 35 à produtividade em condição de déficit hídrico. Dentre os 48 SNPs, foram identificados 35 SNPs ancorados em 31 genes de arroz. Dentre os genes identificados, sete deles continham SNPs associados ao ISS, enquanto que os restantes 24 genes continham SNPs associados à produtividade dos acessos em ambiente com deficiência hídrica. Esses genes podem ser avaliados para serem efetivamente utilizados na seleção assistida por marcadores. Adicionalmente, esses genes podem ser superexpressos para avaliar sua capacidade de aumentar a tolerância à seca, e em caso positivo, gerar cultivares comerciais de arroz geneticamente modificadas mais tolerantes a esse estresse.

Palavras-chave: Mapeamento associativo, *Oryza sativa* L., SNPs, deficiência hídrica, CNAE, Genotipagem por Sequenciamento.

---

[1] Orientador: Prof. Dr. Claudio Brondani. Embrapa Arroz e Feijão.

ABSTRACT

PANTALIÃO, G. F. **Association mapping for rice grain yield under drought stress.** 2013. 91 f. Dissertation (Master's degree in Genetics and Plant Breeding) – Agronomy School, Federal University of Goias, Goiania, 2013.[1]

Drought is an environmental factor which narrows crop production, such as upland rice (*Oryza sativa* L.). The knowledge of aspects related to drought stress, and plant response to it, may furnish plant breeding programs essential data for the development of tolerant cultivars, and hence with higher yields under such conditions. Association mapping has been a successful approach to elucidate the genetic basis of economically important traits in plants, and afterward in the implementation of marker assisted selection (MAS). Next-generation sequencing (NGS) technologies have been applied in a variety of contexts, including SNP identification and development. Among methodologies for marker discovery and high-throughput genotyping, GBS (Genotyping by Sequencing) points out by its low cost and speed at which samples can be analyzed. The aim of this work was to identify, by GBS, the polymorphism from SNP markers within 283 upland accessions from Embrapa Rice Core Collection (ERiCC) and associate them to yield under drought stress. After filtering the raw data of predetermined stringent parameters, 285.379 SNP were identified in the 12 rice chromosomes. For the association mapping, molecular and phenotypic data were combined for the identification of SNP associated to drought, aiming the subsequent development of a marker set for MAS besides the identification of genes for genetic engineering. The analysis identified 48 SNP associated with the evaluated traits, 13 associated to drought susceptibility index (DSI) and 35 to yield under drought stress. Among the 48 SNP, 35 was anchored in 31 rice genes. Seven genes, out of the 31, possessed SNP associated to DSI, and the other 24 genes to yield under drought stress. These genes may be evaluated to be effectively employed for MAS. If the overexpression of such genes provides an enhanced drought tolerance, they may be used in the development of tolerant rice cultivars.

Keywords: Association Mapping, *Oryza sativa* L., SNPs, drought stress, ERiCC, Genotyping by Sequencing.

[1] Adviser: Prof. Dr. Claudio Brondani. Embrapa Rice and Beans.

1 INTRODUÇÃO

O arroz é uma das espécies mais importantes para a alimentação humana, representando a principal fonte de carboidrato para mais da metade da população mundial (Khush, 2005). Estima-se que a população mundial aumentará de forma considerável, o que também irá demandar um aumento na produção de alimentos em relação aos níveis atuais, em um cenário de redução da área cultivável e escassez de recursos hídricos. O uso contínuo da irrigação e a restrição dos recursos hídricos provocando estresse é uma realidade crescente, devido, principalmente, ao aumento da poluição e à competição com o consumo industrial e urbano (Chaves & Oliveira, 2004).

Os programas de melhoramento de arroz, devem, portanto, priorizar a busca por novas estratégias que visem o aumento da tolerância e da produtividade em condições de déficit hídrico. Uma das principais alternativas é o desenvolvimento de genótipos mais tolerantes à seca, a partir da identificação de fontes doadoras de genes tolerantes a esse estresse. Os recursos genéticos do arroz compreendem variedades tradicionais, cultivares e linhagens modernas e espécies silvestres relacionadas. Esta diversidade permite o acesso a diferentes opções para o desenvolvimento de novos cultivares pelos programas de melhoramento, obtendo-se assim cultivares mais produtivas, resistentes a doenças e pragas e mais tolerantes à seca (Rao, 2004).

O arroz é considerado uma planta modelo para estudos genômicos por apresentar o menor genoma entre os cereais, em torno de 382 Mpb (milhões de pares de base). Adicionalmente, a partir do sequenciamento do genoma do arroz, é possível a realização de estudos comparativos com outras espécies de gramíneas, por apresentarem sintenia entre genomas, ou seja, blocos de genes conservados, tanto com relação à organização quanto à similaridade em conteúdo e sequência (Tyagi et al., 2004).

O sequenciamento de nova geração (NGS) está expandindo os conhecimentos sobre a variabilidade genética do arroz. A alta qualidade dos genomas de referência publicamente disponíveis para o arroz e o resequenciamento de diversos acessos de bancos de germoplasma têm formado uma base sólida para os estudos genômica estrutural e funcional. A tecnologia de genotipagem por sequenciamento (GBS) tem sido utilizada na identificação de inúmeros SNPs, a um custo bastante reduzido em comparação a genotipagem utilizando chips de DNA (Elshire et al., 2011).

A tolerância a estresses abióticos, como a seca, geralmente é resultante da ação de caracteres controlados por vários genes e, portanto a identificação de regiões genômicas associadas à tolerância à seca pode ser o primeiro passo para a compreensão da base molecular do controle genético deste tipo de caráter. A utilização da estratégia de mapeamento associativo está baseada no uso de indivíduos não aparentados e todo o histórico de recombinações, permitindo a obtenção de estimativas mais precisas sobre a localização de genes de interesse (Abdallah et al., 2003).

A análise de associação tem o potencial de identificar polimorfismos dentro de genes que sejam responsáveis por variações fenotípicas (Flint-Garcia et al., 2003), estabelecendo uma relação entre o genótipo e o fenótipo. Sendo assim, o mapeamento associativo busca identificar os polimorfismos na sequência de DNA responsáveis por caracteres de interesse, além de facilitar a seleção de genótipos relacionados com os fenótipos (Oraguzie et al., 2007).

Este estudo teve como objetivos: 1) Identificar os materiais mais produtivos em condição de deficiência hídrica e os materiais mais sensíveis à seca, a partir da caracterização fenotípica de 283 acessos de arroz de terras altas da Coleção Nuclear de Arroz da Embrapa (CNAE); 2) Identificar SNPs pela tecnologia de genotipagem por sequenciamento (GBS); 3) Associar os SNPs identificados com os caracteres produtividade de grãos e índice de suscetibilidade à seca através da análise de Mapeamento Associativo; 4) Identificar SNPs associados significativamente aos caracteres avaliados que estejam ancorados a genes já identificados em arroz; e 5) Caracterizar os genes que apresentem SNPs ancorados de acordo com seu produto expresso, função molecular e processo biológico em que atua.

2 REVISÃO BIBLIOGRÁFICA

2.1 O GÊNERO *ORYZA*

O arroz pertence ao gênero *Oryza* e é uma angiosperma monocotiledônea pertencente à família Gramineae (Poaceae), ordem Glumoflorae, sufamília Ehrhartoideae e tribo Oryzae (Vaughan, 1989; Watanabe, 1997). Há duas espécies cultivadas dentro desse grupo, *Oryza sativa* L. e *O. glaberrima* Steud, sendo que a primeira é amplamente cultivada, enquanto que a segunda se restringe à África Central (Second, 1982). Existem também aproximadamente vinte espécies silvestres distribuídas pela África, Ásia, Américas e Austrália (Chang, 1976; Morishima & Martins, 1994; Vaughan, 1994; Khush, 1997).

Segundo Vaughan et al. (2005), o agrupamento das espécies de arroz pode ser realizado em quatro complexos: *O. sativa* (AA), *O. officinalis* (BB, CC, BBCC, CCDD, EE), *O. ridleyi* (HHJJ) e *O. granulata* (GG), sendo que as espécies *O. schlechteri* e *O. brachyantha* não devem ser agrupadas em nenhum destes complexos, pois apresentam características que poderiam classificá-las como pertencentes a outro gênero (*Leersia*). O grupo citogenético HK foi atribuído à espécie *Oryza schlechteri* por Ge et al. (1999), já que até então ela não possuía grupo citogenético conhecido, enquanto que Vaughan et al. (2003) atribuiram genoma FF à *O. brachyanha*.

A classificação do arroz (*O. sativa* L.) em duas subespécies denominadas *indica* (hsien) e *japonica* (keng) foi introduzida por Kato et al. (1928). Entre essas subespécies há uma alta diversidade genética, o que sugere eventos de domesticação independentes de "*pools*" divergentes de *O. rufipogon*, que se diferenciaram há milhares de anos por isolamento geográfico (Garris et al., 2005; Sweeney & McCouch, 2007). A subespécie *indica* é encontrada nas regiões inundadas da Ásia tropical enquanto que a subespécie *japonica* é encontrada nas regiões de terras altas e elevações do Sul da Ásia (Garris et al., 2005). Acredita-se que o arroz cultivado (*O. sativa*) tenha sido domesticado entre os anos 15.000 e 10.000 a.C. na região entre a Índia e Mianmar (antiga Birmânia) (Vieira, 2007).

2.1.1 A espécie *O. sativa*

O. sativa representa um importante papel para a agricultura mundial, sendo que mais de 90% do arroz cultivado pertence a essa espécie (Kush & Brar, 2002). A Ásia é o maior produtor de arroz do mundo, e além desse continente, o arroz é cultivado na África, América Latina, Estados Unidos e Austrália. O Brasil é o maior produtor de arroz fora do continente Asiático (Souza et al., 2010).

A produção mundial de arroz teve um aumento considerável de 1965 (257 milhões de toneladas) para o ano 2011 (723 milhões de toneladas). Ao longo dos anos, porém, o ganho genético para produtividade tem diminuído, fazendo com que haja uma grande preocupação mundial em relação à segurança alimentar diante do crescimento populacional, escassez de novas áreas de plantio, limitação de recursos hídricos e redução da fertilidade das áreas de cultivo (Khush, 2005).

Na safra 2012/2013 está previsto um aumento da produção mundial de arroz, chegando a 724,5 milhões de toneladas, o que representa um aumento de 2% em relação à safra 2011/2012, que apresentava uma área cultivada que chegou a 163 milhões de hectares. Como dito anteriormente, a Ásia tem a maior parte da produção global de arroz com uma previsão de chegar aos 657 milhões de toneladas na safra 2012/2013, ou seja, 0,4% de aumento em comparação com a safra anterior (FAO, 2012).

Segundo Goicoechea et al. (2010), a população consumidora de arroz duplicará nos próximos 23 anos. Sendo assim, um dos principais objetivos dos programas de melhoramento genético do arroz é aumentar a produtividade do grão, para fazer frente ao aumento do consumo. Uma solução para esse problema de segurança alimentar é a exploração da diversidade genética armazenada em bancos de germoplasma de arroz. Essa estratégia permite uma busca por alelos ainda não presentes no pool gênico do arroz cultivado, assim como a obtenção de novas combinações alélicas de genes relacionados a caracteres de importância agronômica, e que podem contribuir significativamente para a obtenção de cultivares mais produtivas (Khush, 2005.; Zhang et al., 2011).

No Brasil, o arroz é cultivado sob dois sistemas de cultivo, o irrigado e o de sequeiro ou de terras altas (Brondani et al., 2006). Segundo Khush (1997), a maioria das variedades de sequeiro é da subespécie *japonica* tropical e as variedades irrigadas são do tipo *indica*. A diferença entre esses dois sistemas de cultivo reside no fato de que

o cultivo irrigado é feito com água cobrindo o solo por inundação contínua em áreas de várzea ou de baixada, enquanto que o cultivo de sequeiro é realizado em terrenos drenados, sendo que o arroz torna-se dependente das precipitações pluviais e procedimentos de irrigação complementares (Khush, 1997).

Arroz é um cereal de grande importância para a segurança alimentar, principalmente, quanto ao seu valor energético, pois apresenta alto teor de carboidratos (Pereira, 1996). De acordo com o último estudo de Gomes & Magalhães (2004), no mundo, estima-se que o consumo médio individual de arroz seja de 60 kg/pessoa/ano, sendo que os países asiáticos apresentam médias mais elevadas, situadas entre 100 e 150 kg/pessoa/ano, enquanto na América Latina consome-se em média 30 kg/pessoa/ano. Entre os países da América Latina, o Brasil é destacado como um grande consumidor, com uma média de 45 kg/pessoa/ano. Em termos regionais, no Brasil, a região Centro-Oeste apresenta maior consumo médio *per capita* (97,18 kg/hab/ano), seguida pelas regiões Sudeste (90,47 Kg/hab/ano), Sul (68,12 kg/hab/ano) e Nordeste (49,64 Kg/hab/ano). Os Estados de Tocantins e Goiás apresentam maior consumo médio *per capita* (101,57 kg/hab/ano), enquanto Pernambuco (33,9 kg/hab/ano) e Bahia (34,22 kg/hab/ano) apresentam os índices mais baixos (Gomes & Magalhães, 2004).

2.2 COLEÇÃO NUCLEAR DE ARROZ DA EMBRAPA (CNAE)

Desde a introdução da agricultura, durante os processos de domesticação e cultivo de plantas, a diversidade genética tem sido utilizada e, em parte, preservada. Em torno de apenas 15% da diversidade potencial de espécies vegetais têm sido utilizadas pela humanidade, o que implica na redução das possibilidades de obtenção de combinações alélicas valiosas relacionadas a caracteres de importância econômica para culturas de interesse econômico. É necessário, portanto, que esta variabilidade genética possa ser descoberta e utilizada para atender aos desafios que ameaçam a segurança alimentar mundial (Nass, 2007).

A conservação e manutenção da diversidade referente às espécies de interesse real e potencial podem envolver o estabelecimento de coleções de germoplasma *ex situ* e *in situ*. Existem catalogadas mais de 1300 coleções de germoplasma para as mais diversas espécies, armazenando mais de seis milhões de acessos (Upadhyaya et al., 2001). As coleções de arroz encontram-se entre as mais

numerosas do mundo, com mais de 120.000 variedades distintas, o que requer uma elevada soma de recursos destinada anualmente para a conservação e multiplicação periódica dos acessos (Ni et al., 2002).

De um modo geral, o nível de utilização do germoplasma conservado é bastante baixo. Percebe-se deste modo que ao longo dos anos ficou evidente a discrepância entre o crescimento das coleções e o uso efetivo destas em programas de melhoramento, gerando lacunas entre a disponibilidade do germoplasma e o uso real desses materiais. De forma a contornar esta situação, estratégias foram propostas visando maximizar o uso adequado e eficiente dos recursos genéticos armazenados nos bancos de germoplasma. Frankel (1984) sugeriu que fosse realizada a seleção de um número reduzido de acessos que representassem a maior proporção possível da diversidade genética da coleção original, denominando este conjunto de acessos representativos do banco de germoplasma de coleção nuclear, a qual deverá ser sempre menor que a coleção original, representando ao redor de 10% do número de acessos da coleção original, evitando-se a ultrapassar o limite de 2000 acessos (Brown, 1989; Nass, 2007).

No Brasil, a partir da década de 1970, o lançamento de cultivares mais produtivas e a melhoria das práticas agrícolas fizeram com que houvesse um aumento considerável na produtividade do arroz. Contudo, a partir da década de 1980 houve uma redução no ganho genético para produção nos programas de melhoramento de arroz no Brasil (Castro et al., 1996). O uso de genitores aparentados na base dos programas de melhoramento do arroz brasileiro foi a principal causa apontada para essa redução.

Melhoristas contrários à utilização de genótipos de base genética ampla argumentam que ocorre uma perda da combinação de genes que garantiram o aumento expressivo da produtividade obtida com as cultivares modernas. Por esse motivo foram estabelecidos programas de pré-melhoramento, com a finalidade de fornecer linhagens com características favoráveis aos programas de melhoramento, porém com base genética mais ampla.

Concluída em 2002, a Coleção Nuclear de Arroz da Embrapa (CNAE) foi concebida para representar a variabilidade genética da cultura do arroz a partir dos 10.000 acessos componentes do Banco Ativo de Germoplasma (BAG) da Embrapa Arroz e Feijão, não sendo incluídas as espécies silvestres do gênero *Oryza* (Abadie et al., 2005).

A CNAE é constituída por 550 genótipos, equivalendo a 5% da coleção do BAG à época de sua elaboração, e foi estratificada da seguinte maneira: variedades tradicionais do Brasil (VT), linhagens e cultivares melhoradas brasileiras (LCB) e linhagens e cultivares melhoradas introduzidas (LCI) (Tabela 1). Dentro de cada estrato também houve uma estratificação secundária, referente ao ambiente de cultivo de cada acesso (irrigado, terras altas ou facultativo, esse exclusivo do estrato VT). Os dois critérios utilizados para a estratificação da CNAE estão entre os mais usados no estabelecimento de coleções nucleares (Abadie et al., 2005).

Tabela 1. Coleção Nuclear de Arroz da Embrapa (CNAE) dividida pelos estratos e número de acessos que fazem parte de cada estrato.

Estratos	Sistema de Cultivo			Total
	Irrigado	Sequeiro	Facultativo*	
Variedades Tradicionais	77	150	81	308
Linhagense Cultivares Brasileiras	37	57	-	94
Linhagens e Cultivares Introduzidas	72	76	-	148
Total	186	283	81	550

*Corresponde aos acessos que podem ser cultivados nas duas condições de cultivo.

2.3 O GENOMA DO ARROZ CULTIVADO (*Oryza sativa* L.)

O estudo do genoma do arroz apresenta grande importância não somente pelo seu aspecto econômico e social, mas também por sua relevância como planta modelo, devido, principalmente, ao seu genoma compacto e ao elevado grau de sintenia entre os genomas de cereais (Yu & Buckler, 2006). Apesar de os cereais terem evoluído independentemente de um mesmo ancestral há cerca de 50 a 70 milhões de anos, os genomas ainda apresentam uma alta conservação (Goff et al., 2002). O arroz possui um genoma em torno de 382 Mpb (Milhões de pares de base) distribuídos em 12 cromossomos, sendo o menor entre os cereais de importância econômica. Devido à alta conservação de genes entre os cereais, sugere-se que o arroz forneça um “roteiro” para a caracterização de genomas maiores como o milho, cevada e trigo (Tyagi et al., 2004). Entre os benefícios do uso do genoma do arroz como modelo, destacam-se a alta densidade de genes (um gene a cada 8 Kb, aproximadamente), fácil manipulação

genética, utilização de mutações para verificação da função de genes e vasto germoplasma de plantas cultivadas e espécies silvestres (Yu & Buckler, 2006).

O sequenciamento do genoma do arroz já foi realizado e permitiu o desenvolvimento de um número quase ilimitado de marcadores baseados em DNA, que podem ser aplicados em caracterização varietal, construção de mapas de ligação, análise de QTL (Quantitative Trait Loci) e mapeamento associativo, permitindo, assim, um acúmulo de informações detalhadas sobre a estrutura e funcionamento de genes relacionados a caracteres de interesse (Xu et al., 2004).

As análises das sequências genômicas do arroz podem facilitar a obtenção de marcadores SSR, que são sequências nucleotídicas repetidas em série, e de marcadores SNP (Single Nucleotide Polymorphism), que se referem a variações de uma única base. Os SNPs estão distribuídos no genoma do arroz com uma frequência de 1 a cada 170 pb, encontrando-se, aproximadamente 2,4 milhões de polimorfismos de uma única base em todo o genoma (Hayashi et al., 2004).

2.4 PRINCÍPIOS GERAIS DO MAPEAMENTO ASSOCIATIVO

A análise de associação tem o potencial de identificar pequenos polimorfismos que sejam responsáveis por variações fenotípicas (Flint-Garcia et al., 2003), fazendo uma conexão entre o genótipo e o fenótipo. Com isto, é possível detectar polimorfismos de sequência de DNA relacionados à caracteres de interesse e/ou seleção de genótipos que estão relacionadas ao fenótipo (Oraguzie et al., 2007). Uma das vantagens é que na busca de uma associação entre uma característica de interesse e um marcador, o efeito ambiental é minimizado, tendo em vista que a população apresenta indivíduos independentes (Abdurakhmonov et al., 2008).

O mapeamento associativo pode ser realizado tanto em espécies autógamas quanto alógamas. Para as espécies alógamas, como o milho e muitas espécies florestais, o desequilíbrio de ligação é menor, implicando no uso de um número maior de marcadores, fazendo com que o gene-alvo esteja muito próximo da região genômica envolvida e a cobertura completa do genoma seria assegurada com milhões de SNPs. Em estudos envolvendo populações de espécies de autofecundação, o desequilíbrio de ligação ainda é baixo, mas maior em comparação as espécies alógamas, sendo assim, relativamente poucos marcadores são necessários para assegurar a

adequada cobertura do genoma, como por exemplo, a cevada e arroz. A desvantagem é que o marcador poderá estar longe do gene-alvo (Aranzana et al., 2005).

Geralmente, as populações de plantas passíveis de estudos de associação incluem-se em um dos seguintes grupos: (i) Amostra com uma estrutura populacional ideal e um pequeno parentesco familiar; (ii) Multi-amostra de uma mesma família; (iii) Amostra com uma população estruturada; (iv) Amostra com estrutura populacional e relações familiares; (v) Amostra com maior estrutura populacional e relações familiares. Devido à adaptação local, seleção e histórico de parentesco em muitas espécies de plantas, muitas populações cairiam na quarta categoria. A diversidade genética, a extensão do desequilíbrio de ligação no genoma completo e o relacionamento genético dentro de uma população determinam a resolução do mapeamento, a densidade de marcadores, os métodos estatísticos e o poder do mapeamento (Zhu et al., 2008).

No mapeamento associativo, a estrutura populacional apresenta grande relevância, pois o entendimento da mesma faz com que associações espúrias entre fenótipos e genótipos sejam evitadas (Pritchard et al., 2000). Para reduzir significativamente o viés entre as estimativas das associações dos marcadores e o efeito real dos alelos do gene alvo, é importante utilizar uma alta densidade de marcadores moleculares em uma população de grande tamanho e em experimentos repetidos sob várias condições ambientais (Wen et al., 2009).

2.5 MAPEAMENTO ASSOCIATIVO E MAPEAMENTO DE QTL

As duas abordagens estatísticas mais comumente usadas para dissecar caracteres complexos, do ponto de vista do genoma estrutural, são o mapeamento de QTLs (Quantitative Trait Loci) e o mapeamento associativo, baseados no mesmo princípio fundamental de recombinação. A principal diferença é que o mapeamento de QTLs explora a herdabilidade de polimorfismos funcionais e marcadores adjacentes, conduzida comumente com populações experimentais derivadas de cruzamentos biparentais. O mapeamento associativo, por sua vez, examina a herança compartilhada por um conjunto de indivíduos com reduzido vínculo genético, sendo um dos principais fatores da sua maior eficiência para estudar caracteres genéticos complexos. A variabilidade genética na população de mapeamento associativo é muito maior quando comparada a uma população derivada de apenas dois parentais (Yu & Buckler, 2006).

No mapeamento de QTLs, há poucas oportunidades para a ocorrência de recombinação entre as famílias e as genealogias com ascendência conhecida, resultando no mapeamento de resolução relativamente baixa. Por sua vez, no mapeamento associativo o histórico de recombinação e a diversidade genética natural são explorados para o mapeamento de alta resolução (Zhu et al., 2008).

Outra limitação do mapeamento de QTLs é a necessidade de gerar populações de mapeamento, que pode ser de difícil manuseio, dependendo da espécie. Como exemplo, pode-se citar a obtenção de populações de mapeamento para *Pinus*, Eucalipto e outras espécies florestais (Gupta et al., 2005). A obtenção de cruzamentos é dificultada principalmente em espécies que possuem duas estações de produção ao ano, podendo levar até cinco anos para gerarem a população necessária para a obtenção de mapas de ligação (Oraguzie et al., 2007).

A utilização da genética de associação foi uma mudança de paradigma, pois os mapeamentos de QTL priorizavam a existência de alto desequilíbrio de ligação. Por outro lado, a genética associativa permite que sejam avaliadas populações em baixo desequilíbrio de ligação, o que aumenta o poder de detecção de genes relacionados a caracteres de interesse agronômico.

No melhoramento de plantas o mapeamento associativo pode ter várias vantagens sobre a análise de ligação clássica. O primeiro fator a se considerar é a ampla variabilidade genética que pode ser incluída na população de origem, permitindo explorar a variabilidade genética encontrada em bancos de germoplasma, trabalhando-se não só com dois alelos conhecidos e sim com toda uma série alélica.

Um segundo fator seria a não necessidade de desenvolver populações segregantes (Yu & Buckler, 2006). Além disto, os dados fenotípicos e moleculares de outros experimentos que estejam armazenados em bancos de dados podem ser incluídos na análise. Finalmente, as vantagens adicionais são aumento da resolução de mapeamento e a redução do tempo de pesquisa (Wen et al., 2009).

2.6 ABORDAGENS DO MAPEAMENTO ASSOCIATIVO

O mapeamento associativo é dividido em duas categorias principais: (i) Mapeamento associativo de genes candidatos, em que sequências selecionadas têm um papel no controle da variação fenotípica específica para os caracteres; e (ii)

Mapeamento associativo por varredura no genoma completo (tradução livre de "*Whole genome scanning"*), em que se procura obter uma ampla cobertura do genoma com marcadores moleculares, permitindo examinar a variação genética em todo o genoma, a fim de encontrar sinais de associação para caracteres complexos. A escolha da abordagem a ser utilizada depende da extensão do desequilíbrio de ligação e o nível de conhecimento do genoma, pois se o genoma já for bem conhecido, a abordagem de genes candidatos pode ser mais promissora (Risch & Merikangas, 1996). No entanto, as análises de associação com a abordagem de "*Whole genome scanning*" têm rendido resultados promissores, corroborando para a identificação de genes candidatos e identificação de novos locos relacionados a caracteres em plantas (Zhu et al., 2008).

A abordagem de genes candidatos é uma metodologia mais direta em comparação ao estudo do genoma completo, uma vez que o mapeamento de associação fica restrito a genes candidatos relevantes, que podem estar envolvidos no controle de caracteres de interesse (Ehrenreich, 2009). A seleção de genes candidatos é baseada em informações obtidas a partir de estudos de genética, bioquímica, fisiologia da espécie-alvo, ou em estudos conduzidos em espécies-modelo, como *Arabidopsis* e arroz (Yu & Buckler, 2006).

A abordagem de genes candidatos é menos exigente em termos de número de marcadores. No entanto, é importante lembrar que esta estratégia é limitada pelos genes identificados e, portanto, sempre existe o risco de se perder a identificação de mutações causais que estão localizadas em genes não identificados (Hall et al., 2010). Esta abordagem obteve sucesso em muitas situações, como na identificação de genes para diversos caracteres em variedades de milho cultivadas (Harjes et al., 2008), *Pinus* (Gonsalez –Martinez et al., 2007) e batata (Malosetti et al., 2007).

Uma terceira abordagem passou a ser utilizada no mapeamento associativo de milho, conhecida por "*Nested Association Mapping* (NAM)" (tradução livre - Mapeamento de associações aninhadas). O NAM é a junção dos conceitos das duas abordagens principais, sendo uma estratégia ainda mais poderosa para dissecar a base genética de caracteres quantitativos de espécies com baixo DL (Yu & Buckler, 2006). A ideia é que possuindo as sequências de genes candidatos, se faça uma varredura com vários marcadores dentro delas, permitindo assim que as associações sejam bem mais confiáveis, se comparadas às obtidas pelas outras duas abordagens. O NAM foi construído para permitir alta significância e alta resolução por meio de

articulação conjunta na análise de associação, capturando as melhores características das abordagens anteriores (Zhang et al., 2005; Yu et al., 2008).

2.7 SEQUENCIAMENTO DE NOVA GERAÇÃO (NGS)

As tecnologias genômicas que são aplicadas às plantas são muitas vezes desenvolvidas e testadas contra dados de seres humanos ou de outros sistemas modelos, como drosófilas ou camundongos (Pool et al., 2010; Metzker., 2010), mas o grande problema reside no fato de que os genomas de plantas são mais dinâmicos do que os modelos testados, apresentando um maior número de genes e de frequência de poliploidia (Lockton et al., 2005).

Os altos níveis de diversidade de nucleotídeos em alguns genomas vegetais representam um desafio para análises comparativas. Embora, muitas culturas não apresentem altos níveis de diversidade, as dificuldades de análise de um genoma com alta diversidade não são exclusivas de plantas alógamas, como o milho, ocorrendo também em plantas de autofecundação, como cevada e arroz (Caldwell et al., 2006).

Outro desafio no estudo de plantas é o tamanho do genoma, que varia por mais de três ordens de magnitude em espécies já caracterizadas (Gaut et al., 2008), principalmente, devido à presença de elementos transponíveis (Tenaillon et al., 2010). A densidade de elementos transponíveis em genomas de plantas significa que uma grande parte dos dados de sequenciamento fica limitada à utilização do genoma de referência para futuras análises. Por isso que a maioria dos genomas de plantas que foram sequenciados até o momento são relativamente pequenos. O maior genoma sequenciado entre os vegetais foi o do milho, o qual é menor que a metade do tamanho médio do genoma de algumas angiospermas (Nordborg et al., 2008).

Apesar de genomas vegetais apresentarem uma série de desafios para a análise genômica, eles oferecem algumas vantagens, principalmente, quanto a utilização de propagação clonal e manutenção de sementes em bancos de germoplasma, auxiliando na replicação de experimentos em diversos ambientes (Nordborg et al., 2008).

O surgimento do sequenciamento de nova geração (NGS – *Next Generation Sequencing*) revolucionou as abordagens genômicas e transcriptômicas para biologia, principalmente na aplicação às plantas, devido à dificuldade encontrada em se trabalhar com genomas extensos e com muitos polimorfismos. Estas novas ferramentas

de sequenciamento também se tornaram valiosas para a descoberta, validação e avaliação de marcadores moleculares em populações para diversos tipos de estudos (Davey et al., 2011).

Marcadores moleculares estão no centro da genética moderna, permitindo o estudo de questões importantes na genética de populações, ecologia e evolução. Com o advento do sequenciamento de segunda geração (NGS), há varias abordagens que são capazes de descobrir, sequenciar e genotipar não centenas, mas milhares de marcadores em quase qualquer genoma de interesse em uma única etapa (Stapley et al., 2010) mesmo em populações nas quais pouca ou nenhuma informação genética está disponível. Esta mudança de patamar na densidade de marcadores permite não só estudos abrangentes de associação de genomas para qualquer organismo, mas também estudos de todo genoma sobre populações naturais, com benefícios substanciais para conservação genética e melhoramento de plantas (Allendorf et al., 2010).

Muitas questões biológicas agora podem ser respondidas com alta precisão, por exemplo, na identificação de pontos de recombinação para o mapeamento associativo e de QTL, localizando regiões genômicas diferenciadas entre as populações para estudos de genética quantitativa e seleção assistida por marcadores. No entanto, o grande fator limitante não é mais a busca por tecnologias que sejam capazes de gerar informação de sequenciamento em genomas extensos e desconhecidos de plantas. No momento, o gargalo se encontra na análise complexa da grande quantidade de dados que são gerados por essas novas abordagens (Helyar et al.; 2011).

Tradicionalmente, o desenvolvimento de marcadores moleculares como SSR (microssatélites) (Jarne & Lagoda, 1996), RFLP (Botstein et al, 1980) e AFLP (Vos et al., 1995) representavam um processo demorado, trabalhoso e de alto custo. O surgimento de ensaios de SNPs de alta performance removeu esse gargalo do processo de genotipagem, mas a produção de um chip de alta qualidade requer um investimento substancial de recursos, além de que esses marcadores são específicos para a população em que foram desenvolvidos (Davey et al., 2011).

Por outro lado, as técnicas baseadas em NGS permitem a descoberta, o sequenciamento e a genotipagem de milhares a centenas de milhares de marcadores, podendo ser executadas diretamente no DNA genômico em uma etapa única de sequenciamento paralelamente a uma construção de biblioteca genômica. Como o custo de genotipagem usando abordagens baseadas em resequenciamento ainda é maior do

que quando se utiliza chips existentes de SNPs, atualmente, ainda pode ser mais econômico para grandes consórcios desenvolver chips de SNPs que serão utilizados em diferentes populações (Davey et al., 2011). No entanto, para estudos em escala reduzida, o custo do sequenciamento (NGS) é muito inferior ao custo do desenvolvimento da um chip de SNPs (Elshire et al., 2011).

2.8 ANÁLISE GENÉTICA BASEADA EM NGS

Há muitos anos, as enzimas de restrição têm sido uma ferramenta essencial para a genotipagem, desde o desenvolvimento e uso de RFLPs para identificar genes relacionados às doenças humanas e para a construção do primeiro mapa de ligação completo do genoma humano (Donis-Kellet et al., 1987). Para facilitar o trabalho com genomas extensos, principalmente em plantas, muitos dos métodos de sequenciamento de nova geração dependem de enzimas de restrição para produzir uma representação reduzida do genoma. A diversidade de enzimas de restrição disponíveis (que variam em comprimento, simetria e sensibilidade à metilação) faz com que essa seja uma ferramenta de análise extremamente versátil (Davey et al., 2011).

Vários métodos foram desenvolvidos para a descoberta de marcadores moleculares e genotipagem de alto desempenho utilizando enzimas de restrição. Todos os métodos envolvem os seguintes passos principais: a digestão de várias amostras de DNA genômico com enzimas de restrição; uma seleção ou redução dos fragmentos de restrição resultantes, e o sequenciamento de nova geração com o grupo de fragmentos final. Entre as abordagens NGS, destacam-se: sequenciamento de representação reduzida, incluindo bibliotecas de representação reduzida (RRLs) e redução da complexidade de sequências polimórficas (CRoPS); a RAD-seq, que é uma técnica de sequenciamento que utiliza códigos de barra e agrupamento dos fragmentos; e genotipagem de baixa cobertura, como é o caso da genotipagem por sequenciamento (GBS) (Elshire et al.; 2011).

Cada abordagem de metodologia de análise genômica baseada em NGS apresenta uma aplicação diferente. Por exemplo, para um estudo envolvendo uma população silvestre em que nenhum genoma de referência está disponível, um grande número de marcadores são necessários para assegurar com precisão os parâmetros populacionais que serão estimados, sendo assim, os métodos RAD-seq ou de

representação reduzida são mais adequados. Genotipagem para aplicações tais como melhoramento assistido por marcadores e mapeamentos de QTL, a genotipagem de baixa cobertura é suficiente para a ligação ser inferida.

O sequenciamento de representação reduzida (RRLs e CRoPS) refere-se ao sequenciamento de um pequeno conjunto de regiões do genoma, tendo em vista que o sequenciamento do genoma inteiro é caro e muitas vezes desnecessário (Altshuler et al., 2000). Marcadores RAD-seq foram inicialmente implementados usando *microarrays* e, mais tarde, adaptados para NGS. Os fragmentos de restrição são aleatoriamente cortados com um comprimento adequado para a plataforma de sequenciamento de escolha, e uma PCR seletiva é usada para amplificar para o sequenciamento somente os fragmentos que contenham um sítio de restrição. RAD-seq tem sido utilizado para estudar a diferenciação populacional, investigar filogeografia, gerar SNPs e construir mapas de ligação (Davey & Blaxter, 2010).

Os métodos acima reduzem a proporção do genoma alvo de sequenciamento de forma que cada marcador pode ser sequenciado em alta cobertura com recursos limitados, possibilitando assim que marcadores sejam genotipados com precisão entre muitos indivíduos. Uma alternativa a esta abordagem é o sequenciamento de muitos marcadores-alvo a uma baixa cobertura por indivíduo, aceitando que um subconjunto diferente de marcadores será genotipado em cada indivíduo. O método de genotipagem por sequenciamento (GBS) envolve a digestão do DNA genômico com enzimas de restrição, que reduzem a complexidade do genoma de maneira rápida e fácil, e promove o sequenciamento das extremidades de todos os fragmentos resultantes da digestão. Sendo assim, o GBS é uma técnica simples, altamente multiplexável, adequada para estudos populacionais, caracterização de germoplasmas, melhoramento genético, dentre outros. Adaptadores contendo códigos de barra são utilizados nessa abordagem, reduzindo o tempo e o custo dessa tecnologia e representando uma alternativa aos protocolos caros e complexos (Elshire et al., 2011).

2.9 TOLERÂNCIA À SECA EM PLANTAS

A seca é um dos principais fatores ambientais que limita a produção das culturas e a distribuição do cultivo das plantas por todo o mundo (Gorantla et al., 2007). Nos últimos anos, estresses abióticos causaram perdas na produção de mais de 50% (Bray, 2004) e apenas a seca, segundo estimativas, tem causado perdas no campo de mais de 15% (Poroyko et al., 2007).

No cultivo do arroz de terras altas, um dos principais fatores limitantes da produção é o déficit hídrico. Os crescentes índices de poluição dos mananciais do planeta e a competição com o consumo urbano e industrial têm reduzido a disponibilidade de água para a irrigação artificial (Liu et al., 2004). Até mesmo na Ásia, onde o arroz é cultivado sob sistema irrigado em várzeas, o aumento da população e a crescente urbanização têm levado a uma significativa redução da disponibilidade de água para irrigação (Lafitte et al., 2006). Sendo assim, o maior desafio para os programas de melhoramento de arroz é promover a estabilidade da produção mesmo sob déficit hídrico (Nguyen et al., 1997).

O déficit hídrico ocorre devido à redução da disponibilidade de água no solo, desencadeando, na planta, uma série de respostas fisiológicas e bioquímicas sob controle de diversos mecanismos genéticos. Para proteger a maquinaria celular e garantir a sobrevivência, respostas como fechamento dos estômatos, repressão do crescimento celular e da fotossíntese e ativação da fotorrespiração são realizadas pela planta (Shinozaki & Yamaguchi-Shinozaki, 2007).

A principal consequência da diminuição da disponibilidade de água para a planta é a redução da fixação de carbono pelas folhas devido ao fechamento dos estômatos e a redução da taxa fotossintética, o que provoca uma redução do seu crescimento (Chaves & Oliveira, 2004). Sob déficit hídrico, a planta reduz seu crescimento celular, permitindo que a energia que seria usada para essa finalidade seja desviada para as vias metabólicas responsáveis pelo combate aos danos provocados pelo estresse (Zhu, 2002).

Estudos moleculares de regulação gênica de plantas submetidas à seca têm identificado alguns genes que respondem à desidratação. Verificou-se a presença de uma família de genes que codificam fatores de transcrição denominados DREB ("*Dehydration Responsive Element Binding protein*") envolvidos na ativação de vários

outros genes que apresentam características de proteção das estruturas celulares durante a desidratação celular (Shinozaki & Yamagushi-Shinozaki, 2000). A introdução do fator de transcrição DREB1A, sob o controle de um promotor estresse-induzido, em *Arabidopsis thaliana*, tabaco e soja resultou em um aumento da tolerância à seca, salinidade e frio nessas espécies (Kasuga et al., 1999; Kasuga et al., 2004; Beneventi, 2006).

Através de uma busca por locos associados ao ajuste osmótico, que é uma das respostas celulares vitais ao déficit hídrico, em uma população segregante de arroz, foram identificadas cinco regiões do genoma associadas ao estresse hídrico, concluindo que o ajuste osmótico contribui para a tolerância à seca e que a delimitação de regiões associadas a essa característica pode ser feita por mapeamento fino, permitindo melhor compreensão de como o ajuste ocorre nas plantas (Zhang et al.; 2001).

O conhecimento dos mecanismos genéticos envolvidos na resposta ao déficit hídrico é importante, pois permite a identificação dos genes expressos nessas condições e a manipulação dessas informações para a obtenção de cultivares mais tolerantes à seca. Quando se trata do desenvolvimento de cultivares tolerantes, é importante também ter em mente que as respostas genéticas que ocorrem nas células em resposta ao déficit hídrico se refletem em mudanças nos aspectos fisiológicos da planta e que esses aspectos também precisam ser entendidos. Além disso, quando se trata de cultivares, somente a sobrevivência da planta sob um período de seca não é o suficiente, pois as mesmas precisam manter níveis desejáveis de produtividade no final do ciclo da cultura, representando a estabilidade da produção (Fukai & Cooper, 1995).

Um sistema radicular eficiente é capaz de buscar água a grandes profundidades no solo. Sob condições de deficiência hídrica essa habilidade pode representar um maior potencial produtivo e maior estabilidade de produção, pois a planta será capaz de obter água das camadas mais profundas do solo, onde geralmente a retenção de umidade é maior (Nguyen et al., 1997). De maneira geral o arroz possui sistema radicular superficial e é mais eficiente em profundidades até 60 cm, onde possui sua capacidade máxima de busca de água no solo (Fukai & Cooper, 1995). Isso pode explicar a sua maior susceptibilidade à seca quando comparado com outras gramíneas, especialmente durante o período de floração (Fukai & Cooper, 1995; Lafitte et al., 2004). O arroz de várzea possui sistema radicular com maior número de raízes

adventícias e mais superficiais, enquanto que o arroz de terras altas possui raízes mais espessas e profundas (Nguyen et al., 1997).

Segundo Babu et al. (2003), existe forte correlação entre os caracteres desenvolvimento de raiz e produção de grãos. Esses autores identificaram regiões do genoma que estavam associadas a caracteres relacionados à raiz, localizadas na mesma região de locos identificados para produção de grãos sob déficit hídrico. Isso indica que plantas com sistema radicular melhor adaptado produzem mais grãos na seca e que a avaliação de plantas tolerantes pode, dependendo da situação, ser estimada somente pela produção. Isso é vantajoso, pois características de raiz são difíceis de serem avaliadas e a tomada dos dados é bastante trabalhosa (Nguyen et al., 1997; Babu et al., 2003). Para Fukai & Cooper (1995), a produção total de grãos e a proporção de grãos cheios são caracteres muito influenciados pelo déficit hídrico, pois estão diretamente relacionados com a condução de fotoassimilados para as espiguetas no momento da emergência da panícula. A redução da disponibilidade de água durante a floração aumenta a proporção de espiguetas estéreis, diminuindo a produção, uma vez que esses caracteres são altamente correlacionados (Fukai & Cooper, 1995).

O conhecimento de fatores envolvidos na tolerância à seca e das respostas das plantas de arroz ao estresse fornece as informações que servem de base para a obtenção de cultivares tolerantes. As novas tecnologias de genotipagem e sequenciamento oferecem a oportunidade de detecção de regiões do genoma associadas a caracteres quantitativos, abrindo caminho para o melhor entendimento do controle genético desses caracteres e para a possibilidade de utilização das informações do genoma na seleção assistida por marcadores (MAS – *Marker Assisted Selection*) (Tuberosa & Salvi, 2006).

3 MATERIAL E MÉTODOS

3.1 MATERIAL VEGETAL

O mapeamento associativo foi conduzido a partir de um painel composto por 283 acessos da Coleção Nuclear de Arroz da Embrapa (CNAE) pertencentes ao sistema de cultivo de sequeiro. Dentre esses acessos, têm-se 150 variedades tradicionais, 57 linhagens e cultivares brasileiras e 76 linhagens e cultivares introduzidas, sendo que todos acessos da coleção nuclear foram purificados, ou seja, sofreram uma série de autofecundações para se alcançar linhas puras nesses materiais.

3.2 CARACTERIZAÇÃO DO PAINEL DE MAPEAMENTO ASSOCIATIVO QUANTO AO DÉFICIT HÍDRICO

Experimentos em ambiente com e sem deficiência hídrica foram conduzidos no Sítio de Fenotipagem para Seca da Embrapa Arroz e Feijão, na Estação Experimental da Agência Rural (SEAGRO), localizada no município de Porangatu – Goiás (49º 06’ 47” W, 13º 18’ 31” S e altitude média de 396 m). Os 283 acessos componentes da CNAE e pertencentes ao sistema de cultivo sequeiro foram avaliados através de um delineamento experimental do tipo Blocos Aumentados de Federer com três testemunhas: Sertaneja, Pepita e Curinga, durante o ano de 2010.

O ambiente sem deficiência hídrica foi caracterizado por condições adequadas de água no solo (- 0,025 MPa a 15 cm de profundidade monitorada através de tensiômetros), durante todo o desenvolvimento das plantas. O ambiente submetido à deficiência hídrica foi mantido sob as mesmas condições de irrigação até trinta dias após a emergência, quando foi conduzido à deficiência hídrica. Ao ambiente submetido ao período de deficiência hídrica foi disponibilizada aproximadamente 50% da lâmina de água aplicada no ambiente sem deficiência hídrica. De acordo com experiências anteriores, dependendo das condições atmosféricas, este nível de estresse pode ocasionar uma redução de 50% a 70% na produtividade média da cultura.

Foram avaliados os potenciais produtivos dos 283 acessos através das medidas de produtividade de grãos, ou seja, após a completa maturação fisiológica dos

grãos de cada parcela, as plantas de arroz da área útil foram colhidas manualmente. As panículas colhidas foram acondicionadas em sacos de fibra de algodão, identificadas por meio de etiquetas e conduzidas para a trilha e determinação do peso de grãos por parcela, transformado para $kg.ha^{-1}$.

Também foi calculado o Índice de Susceptibilidade à Seca (ISS) conforme a metodologia de Fischer & Maurer (1978) realizada por meio da fórmula:

ISS = Yw - Yd / Yw x IE,

Em que:

Yd - valor da produtividade do genótipo sob deficiência hídrica;

Yw - valor da produtividade do genótipo sob boas condições hídricas;

IE - intensidade do estresse, severidade da deficiência hídrica.

Intensidade do Estresse (IE): avaliada pela fórmula: **IE=1-(Xd/Xw)**,

Em que:
Xd - valor médio da produtividade dos genótipos sob deficiência hídrica;

Xw - valor médio da produtividade dos genótipos sob boas condições hídricas.

3.3 ANÁLISE DO CONJUNTO DE DADOS FENOTÍPICOS

3.3.1 Ajuste dos dados fenotípicos

Na avaliação fenotípica dos acessos do sistema de cultivo sequeiro da Coleção Nuclear de Arroz da Embrapa (CNAE) utilizados para o mapeamento associativo, foi realizado um ajuste das médias com base nas testemunhas dispostas pelo delineamento de Blocos Aumentados de Federer. A análise estatística dos dados fenotípicos foi realizada pelo *software* SAS v.9.2 (Statistical Analysis System) por meio do procedimento de modelos mistos através do comando *PROC MLM* (SAS Institute, 1995).

No procedimento de modelos mistos, os testes foram aplicados interblocos e intergenotípicos, considerando os blocos e acessos como sendo de efeito aleatório e as testemunhas de efeito fixo. A correção dos dados foi feita para o caractere produtividade (Kg/ha), obtendo-se os melhores ajustes estatisticamente significativos para o conjunto de dados, minimizando a soma dos quadrados das diferenças entre o valor estimado e os dados observados nos experimentos com e sem deficiência hídrica.

3.3.2 Distribuição de frequências e dispersão

Com base nas médias ajustadas de produtividade e índice de suscetibilidade à seca, foram feitos gráficos de distribuição e gráficos do tipo Boxplot para cada ambiente com e sem deficiência no ano de 2010 avaliado, com a finalidade de captar importantes aspectos do conjunto de dados trabalhado. Os Boxplots foram construídos através de um script específico no *software* R.

Comparações entre os acessos também foram realizadas, buscando-se obter os 20 materiais mais produtivos e os mais sensíveis à seca nas duas condições estudadas. Para comprovar que esses materiais foram os mais produtivos, foi feito um teste com base no desvio padrão das médias dos acessos em cada ambiente.

3.4 DADOS GENOTÍPICOS

O DNA genômico dos 283 acessos componentes do painel de associação foi obtido a partir de folhas jovens utilizando-se kit comercial DNeasy 96 Plant Kit (QIAGEN).

Para a genotipagem em larga escala de SNP utilizou-se a tecnologia de sequenciamento de nova geração denominada de genotipagem por sequenciamento (GBS, *Genotyping by sequencing*). O DNA genômico foi enviado e analisado pelo Instituto de Diversidade Genômica da Universidade de Cornell (*Buckler Laboratory*). Neste instituto foram construídas as bibliotecas genômicas e conduzido o GBS.

A coleta de dados foi realizada em uma plataforma Genome Analyzer II (Illumina) e o sequenciamento foi do tipo *single-end* com plexagem de 96 amostras. As bibliotecas foram preparadas e analisadas de acordo com Elshire et al. (2011), utilizando-se a enzima de restrição "APK1" para digestão e desenvolvimento da biblioteca com *barcodes* únicos.

A bioinformática básica, que é responsável pela identificação dos SNPs e pela construção de um arquivo em formato HAPMAP, e a filtragem dos dados brutos, para se retirar alelos raros que podem gerar falsas associações, basearam-se em estimativas prévias de desequilíbrio de ligação e índices de endogamia obtidos para a cultura do arroz.

O procedimento (*pipeline*) para a identificação de SNP (SNP *call*) foi composto por diversas etapas. A primeira etapa foi composta pelo processamento dos *reads* para a identificação de quais alelos estavam presentes entre os acessos analisados, em seguida, os *reads* foram ancorados ao genoma de referência.

Posteriormente, polimorfismos na ancoragem dos *reads* foram reconhecidos, confirmando a detecção de SNP presente em cada um dos acessos. Também foi definido o valor de Frequência Mínima dos Alelos (FMA) como 0,01. Este valor corresponde à frequência mínima que o alelo deve apresentar para ser considerado na análise, retirando alelos raros que comprometam a análise. Além disto, foram considerados como requisitos o Coeficiente de Endogamia igual a 0,9 e Cobertura Mínima dos Locos igual a 0,1, a qual por sua vez corresponde ao número de acessos com pelo menos uma marca (*tag*) no locus.

3.5 ESTRUTURA GENÉTICA POPULACIONAL

Foram utilizados em torno de 1000 SNPs derivados do conjunto obtido originalmente, os marcadores selecionados foram filtrados a partir de sua frequência alélica. As hipóteses de divisão em subconjuntos foi testada através do software Structure (Pritchard et al., 2000), que faz uso de inferência Bayesiana. Para se obter uma boa convergência dos dados, utilizou-se um burn-in no valor de 10000 e um MCMC no valor de 50000, testando 10 tipos de estruturações ($1 < K < 10$) com 5 repetições para cada K.

Durante o processo de estimação admitiu-se a possibilidade de ocorrência de genótipos resultantes de cruzamentos entre subpopulações (*admixture model*). O método descrito por Evanno et al. (2005) foi utilizado para se determinar o número efetivo de subgrupos (K) na população, através do software Structure Harvester v. 0.6 (Earl, 2011), disponível on-line.

3.6 ANÁLISE DE ASSOCIAÇÃO

A análise de associação entre os marcadores SNPs e os caracteres de interesse, potencial produtivo em ambiente com e sem deficiência hídrica e Índice de Suscetibilidade à Seca (ISS) no ano de 2010 avaliado foi realizada pelo programa computacional TASSEL versão 3.0 standalone (Bradbury et al., 2007), utilizando-se o módulo *Mixed Linear Model* (MLM).

A matriz de coancestralidade (matriz K ou kinship) foi obtida a partir do mesmo software. Os marcadores SNPs identificados e os dados de estruturação foram considerados como fatores de efeito fixo, enquanto que a matriz de parentesco foi considerada como fator de efeito aleatório.

Para confirmar a significância das associações entre locos e caracteres fenotípicos, foi utilizado o método FDR (*false discovery rate*), obtido pelo *software* Qvalue versão 1.0 (Storey, 2002), que tem por finalidade reduzir o número de falsos positivos, que podem decorrer de associações por ancestralidade comum.

Os valores mais relevantes para uma associação são os P-valores, que indicam a significância das associações entre os marcadores e os caracteres envolvidos. Neste trabalho, esses valores foram ajustados, pelo método FDR, fornecendo, assim, os

Q-valores, sendo que, após a obtenção dos Q-valores, foi utilizado um nível de 0,05 de significância para indicar os SNPs associados de forma significativa.

Com base nos valores de significância ajustados (Q-valores), foram construídos gráficos do tipo Manhattan Plot, para se observar a dispersão das associações entre marcadores SNPs e os caracteres de interesse, delimitando as associações estatisticamente significativas (< 0,05).

3.7 ANOTAÇÃO DOS GENES

Depois que os SNPs foram associados aos caracteres de interesse através do Modelo Linear Misto, foram utilizadas as posições fornecidas de cada marcador SNP com a finalidade de se encontrar transcritos conhecidos do genoma do arroz publicamente disponíveis (cultivar Nipponbare; MSU Rice Genome versão 7.0).

A partir da informação da sequência dos genes, através da análise de BLASTx realizada pela ferramenta blast+ do pacote NCBI C++ Toolkit (NCBI), foi possível realizar a identificação dos genes e suas proteínas já anotadas utilizando os códigos (IDs) identificados na busca de termos descritos no Gene Ontology (GO *terms*). Estes termos do GO agrupam os genes pela função ou provável função das proteínas por categorias funcionais principais, utilizou-se a base de dados agriGO (http://bioinfo.cau.edu.cn/agriGO). Essa análise utiliza um conjunto de genes predefinidos ou genes da referência e ranqueia os genes para identificar os prováveis processos biológicos e funções moleculares dos transcritos.

4 RESULTADOS

4.1 AVALIAÇÃO FENOTÍPICA DOS ACESSOS

A avaliação fenotípica dos acessos do sistema de cultivo de sequeiro da Coleção Nuclear de Arroz da Embrapa (CNAE) foi utilizada para a análise de mapeamento associativo. Os valores para produtividade (Kg ha^{-1}) foram corrigidos pelo ajuste de médias (Tabelas 2 e 3) das testemunhas, uma vez que o delineamento utilizado foi o de Blocos Aumentados de Federer. Os valores dos acessos ajustados para ambos os experimentos foram estatisticamente significativos (<.0001).

Tabela 2. Método dos Quadrados Mínimos para a variável Produtividade (kg ha^{-1}) em ambiente com deficiência hídrica.

	QM							
F.V	**Médias**	**Erro Padrão**	**GL**	**t Valor**	**Pr > |t|**	**Alpha**	**Inferior**	**Superior**
Acessos	1081.40	94.1192	14	11.49	< .0001	0.05	879.55	1283.25
BRSMG_Curinga	589.13	219.06	28.1	2.69	0.0119	0.05	140.49	1037.77
BRS_Pepita	2122.13	201.38	27.9	10.54	< .0001	0.05	1709.58	2534.68
BRS_Sertaneja	1933.17	230.18	28.2	8.40	< .0001	0.05	1461.82	2404.52

Tabela 3. Método dos Quadrados Mínimos para a variável Produtividade (kg ha^{-1}) em ambiente sem deficiência hídrica.

	QM							
F.V	**Médias**	**Erro Padrão**	**GL**	**t Valor**	**Pr > |t|**	**Alpha**	**Inferior**	**Superior**
Acessos	2385.47	78.7573	27.5	30.29	< .0001	0.05	2224.00	2546.93
BRSMG_Curinga	2908.93	191.04	31.4	15.23	< .0001	0.05	2519.50	3298.35
BRS_Pepita	3144.18	175.73	31.4	17.89	< .0001	0.05	2785.96	3502.41
BRS_Sertaneja	3961.99	182.91	31.4	21.66	< .0001	0.05	3589.14	4334.84

4.1.1 Conjunto de dados

A produtividade média dos acessos de arroz de sequeiro da CNAE avaliados em ambiente com deficiência hídrica variou de 342,62 kg.ha^{-1} (Linhagem Introduzida TB47H-MR-11-51-3) a 3312,88 kg.ha^{-1} (Linhagem Introduzida IRAT 141), com uma média geral para o experimento de 1081,4 kg.ha^{-1}. A produtividade média dos acessos de arroz de sequeiro da CNAE avaliados em ambiente sem deficiência hídrica variou de 614,82 kg.ha^{-1} (Variedade Tradicional JAPONÊS) a 5477,08 kg.ha^{-1} (Linhagem Introduzida CT11632-3-3-M), com uma média geral de 2385,47 kg.ha^{-1}.

O índice de suscetibilidade à seca (ISS) leva em consideração as reações de rendimento produtivo em resposta às condições com e sem déficit hídrico. Neste estudo ele variou de -1,8560 a 1,5739 com média de 0,8490 para produtividade. O acesso mais sensível à seca foi a Linhagem Introduzida CT11632-3-3-M, enquanto que o acesso menos sensível à seca foi a Linhagem Introduzida IRAT 13.

Com base nas médias ajustadas de produtividade e ISS, foram construídos gráficos de distribuição de frequências e gráficos do tipo Boxplot, com a finalidade de captar importantes aspectos do conjunto de dados trabalhado.

No experimento com deficiência hídrica, foram obtidas as frequências relativas das classes dos dados de produtividade, com a classe (942,62 kg.ha^{-1} a 1242,62 kg.ha^{-1}) apresentando a maior frequência (46,64%). No experimento sem deficiência hídrica, também foram obtidas as frequências relativas das classes dos dados de produtividade, com a classe (2114,82 kg.ha^{-1} a 2614,82 kg.ha^{-1}) apresentando a maior frequência (23,67%) (Figuras 1 e 2). A distribuição de frequências do ISS também foi obtida, com a classe (0,94 a 1,34) apresentando a maior frequência (45,58%) (Figura 3).

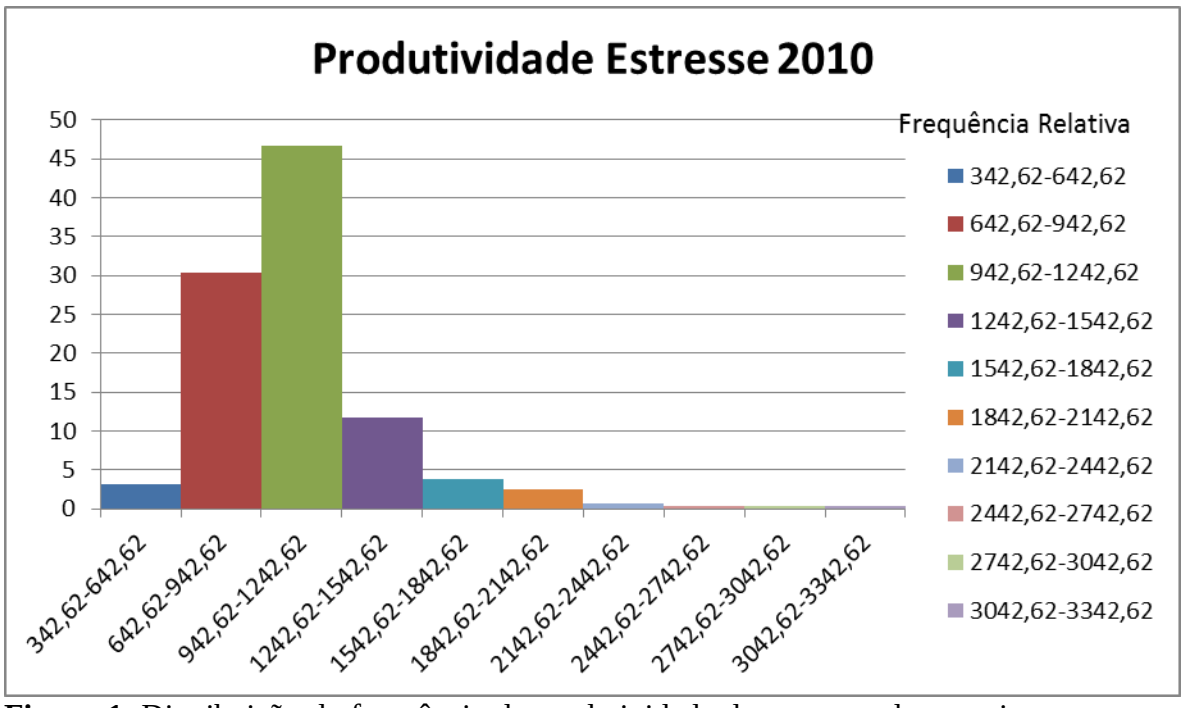

Figura 1. Distribuição de frequência da produtividade dos acessos do experimento em ambiente com deficiência hídrica.

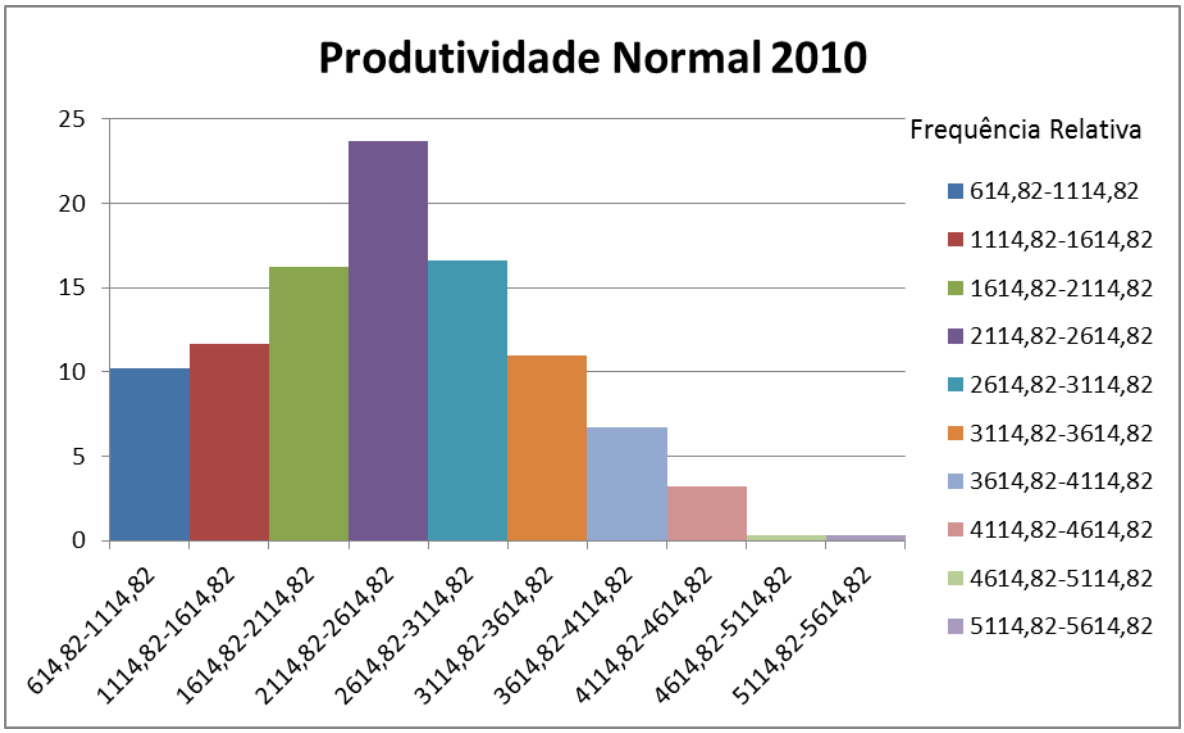

Figura 2. Distribuição de frequência da produtividade dos acessos do experimento em ambiente sem deficiência hídrica.

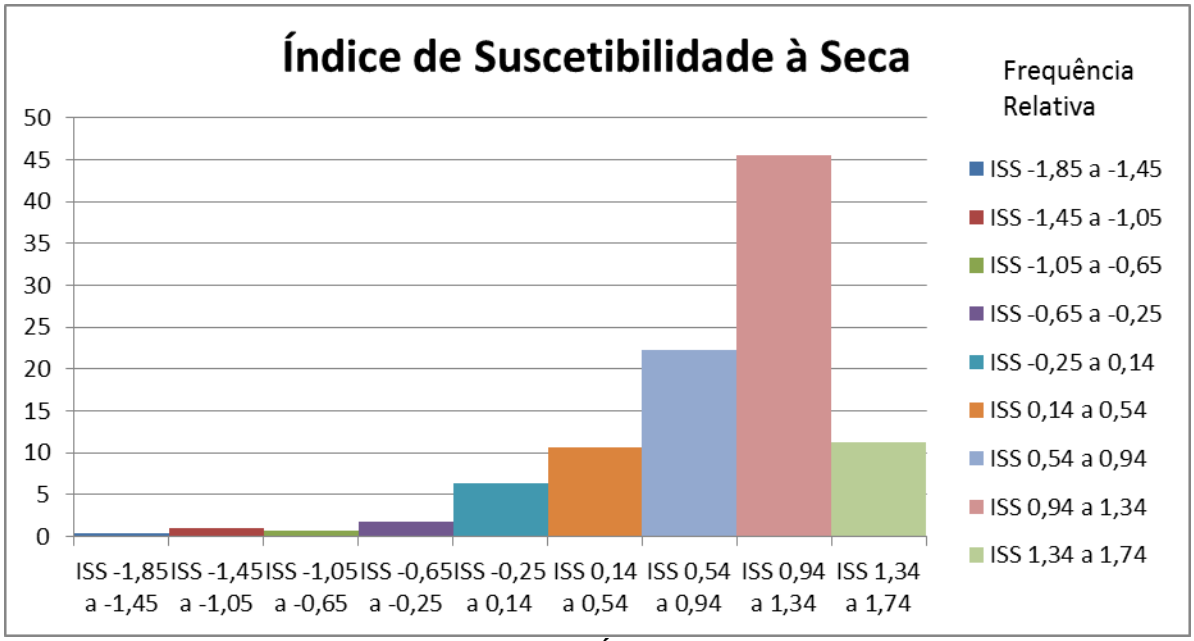

Figura 3. Distribuição de frequência do Índice de Suscetibilidade à Seca para produtividade no experimento com deficiência hídrica.

O gráfico boxplot de produtividade mostrou uma grande dispersão dos dados em ambiente sem deficiência hídrica, com o valor mínimo não discrepante de 614,82 kg.ha^{-1}, o valor máximo não discrepante de 4731,13 kg.ha^{-1} e com um Intervalo Interquartil (IIQ) de 1286,3 kg.ha^{-1}. Considerando os dados em ambiente com deficiência hídrica, houve uma menor dispersão, com o valor mínimo não discrepante de 519,26 kg.ha^{-1}, o valor máximo não discrepante de 1624,08 kg.ha^{-1} e com um Intervalo Interquartil (IIQ) de 310,8553 kg.ha^{-1} (Figura 4).

No gráfico boxplot para o Índice de Suscetibilidade à Seca, a dispersão dos dados comportou-se com o valor mínimo não discrepante de -0,1534, o valor máximo não discrepante de 1,5739 e com um Intervalo Interquartil (IIQ) de 0,5434 (Figura 5).

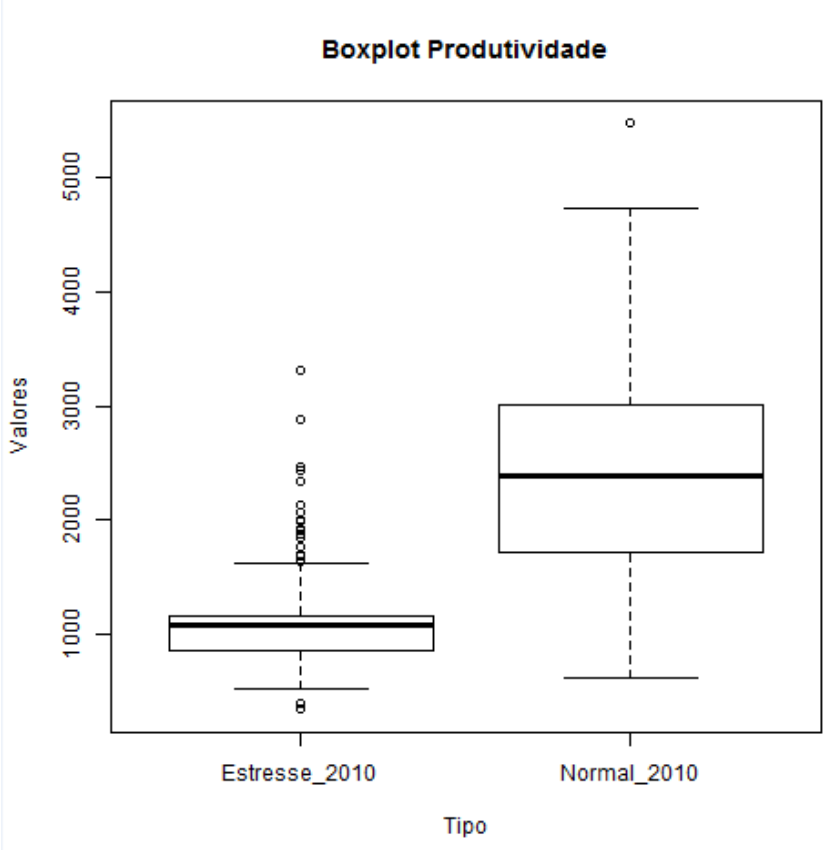

Figura 4. Gráfico boxplot para a produtividade nos ambientes com e sem deficiência hídrica.

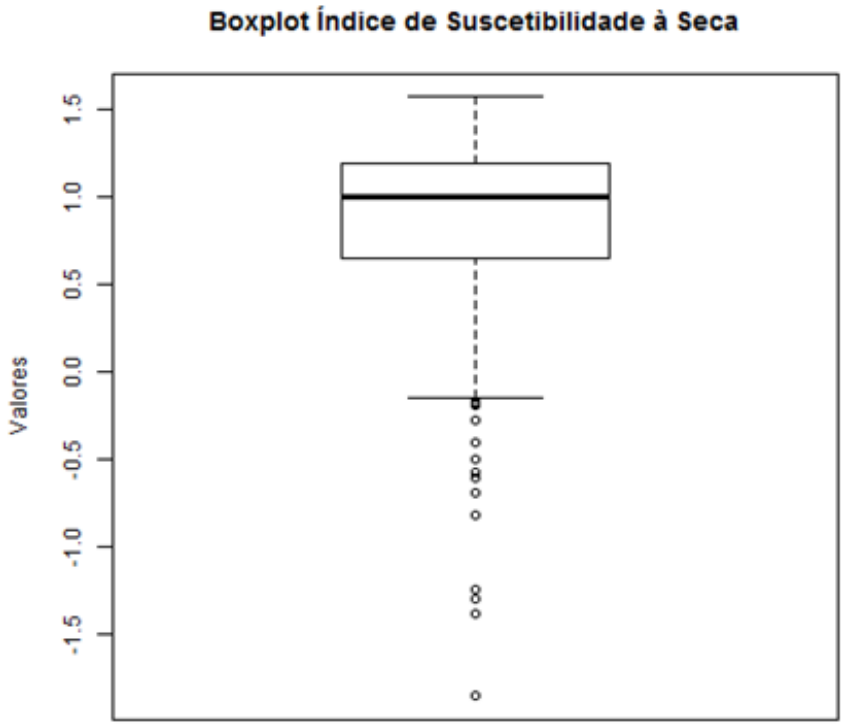

Figura 5. Gráfico boxplot para o Índice de Suscetibilidade à Seca.

4.1.2 Produtividade e suscetibilidade à seca

Com base nos dados das médias ajustadas, foram realizadas comparações entre os acessos com a finalidade de se obter os materiais mais produtivos e os mais sensíveis à seca. No ambiente com deficiência hídrica, foram obtidos os vinte materiais mais produtivos, sendo que, desses materiais, nove eram Variedades Tradicionais, três eram Linhagens Brasileiras (material melhorado) e oito eram Linhagens Introduzidas (material melhorado no exterior) (Tabela 4). No ambiente sem deficiência hídrica, dentre os vinte materiais mais produtivos, quatro eram Variedades Tradicionais, sete eram Linhagens Brasileiras e nove eram Linhagens Introduzidas (Tabela 5).

Dentre os vinte materiais mais sensíveis à seca, seis eram Variedades Tradicionais, seis Linhagens Brasileiras e oito Linhagens Introduzidas (Tabela 6).

Tabela 4. Vinte materiais mais produtivos do ambiente com deficiência hídrica.

ACESSOS	Nome	Origem	PROD. ($kg.ha^{-1}$)
BGA_3395	IRAT 141	LCI	3312,88
BGA_13138	JAGUARY	VT	2884,50
BGA_12356	AGULHA	VT	2470,40
BGA_5975	IDSA 6	LCB	2433,28
BGA_13020	AGULHÃO	VT	2339,74
BGA_9102	CT10006-7-2-M-5-1P-3	LCI	2128,68
BGA_5970	FAROX 299	LCI	2071,03
BGA_12954	SAIA VELHA	VT	1999,20
BGA_13877	AGULHA DA TERRA	VT	1992,71
BGA_8305	CARISMA	LCB	1923,06
BGA_13890	CACHO GRANDE	VT	1907,22
BGA_13731	AGULHINHA BRANCO	VT	1876,17
BGA_11224	MATO GROSSO	VT	1838,53
BGA_5972	FAROX 301	LCI	1766,05
BGA_4193	IREM 656	LCI	1702,58
BGA_10527	BR4742-B-19-23	LCI	1679,10
BGA_13466	MONTANHINHA 90 DIAS	VT	1642,53
BGA_8411	BLUE BELLE	LCI	1624,08
BGA_4168	RIO DOCE	LCB	1607,24
BGA_6574	IRAT 112	LCI	1595,05

Tabela 5. Vinte materiais mais produtivos do ambiente sem deficiência hídrica.

ACESSOS	Nome	Origem	PROD. ($kg.ha^{-1}$)
BGA_9115	CT11632-3-3-M	LCI	5477,08
BGA_5660	IPEACO 77-P	LCB	4731,13
BGA_3281	CABAÇU	LCB	4582,49
BGA_3395	IRAT 141	LCI	4552,08
BGA_13050	CHATÃO AMARELO	VT	4542,82
BGA_13880	LEVANTA TESTA	VT	4487,13
BGA_13928	FARROUPILHA	VT	4461,55
BGA_4640	TOX 1785-19-18	LCI	4359,26
BGA_3397	IRAT 144	LCI	4248,42
BGA_3362	IRAT 142	LCI	4237,30
BGA_8305	CARISMA	LCB	4143,74
BGA_9280	CT13573-11-2	LCI	3980,82
BGA_10533	1R65907-188-1-B	LCI	3960,47
BGA_9223	CT13569-5-7	LCI	3906,69
BGA_9154	CT13370-2-M	LCI	3873,83
BGA_7024	CNAX 1503-12-9-4-B	LCB	3842,92
BGA_6187	CAIAPÓ	LCB	3836,49
BGA_11224	MATO GROSSO	VT	3819,09
BGA_5180	TANGARÁ	LCB	3798,59
BGA_8540	TALENTO	LCB	3770,00

Tabela 6. Vinte materiais mais sensíveis à seca no experimento de 2010.

ACESSOS	Nome	Origem	ISS
BGA_9115	CT11632-3-3-M	LCI	1,573
BGA_4120	RIO PARAGUAY	LCB	1,538
BGA_5660	IPEACO 77-P	LCB	1,515
BGA_10469	TB47H-MR-11-51-3	LCI	1,510
BGA_9280	CT13573-11-2	LCI	1,499
BGA_6422	IAPAR L 99-98	LCB	1,476
BGA_3362	IRAT 142	LCI	1,463
BGA_6940	LEMONT	LCI	1,463
BGA_10682	IAC 202	LCB	1,454
BGA_13307	AGULHA	VT	1,452
BGA_6187	CAIAPÓ	LCB	1,451
BGA_3397	IRAT 144	LCI	1,449
BGA_13329	PINGO DE OURO	VT	1,436
BGA_13753	BICO GANGA	VT	1,428
BGA_12141	PRATA BRANCO	VT	1,428
BGA_13405	4 MESES BRANCO	VT	1,423
BGA_7024	CNAX 1503-12-9-4-B	LCB	1,420
BGA_12770	101	VT	1,405
BGA_8093	L 285	LCI	1,397
BGA_9154	CT13370-2-M	LCI	1,397

As médias obtidas para os vinte acessos mais produtivos em cada condição diferiram de forma estatisticamente significativa do resto dos dados, devido à presença de cinco desvios padrões de diferença entre a média da produtividade em ambiente sob déficit hídrico e o 20° material mais produtivo nessa condição, e também devido à presença de 17 desvios padrões de diferença entre a média da produtividade em ambiente com irrigação normal e o 20° material mais produtivo nessa condição.

4.2 DADOS GENOTÍPICOS

A caracterização molecular dos 283 acessos componentes do painel de associação foi obtida pela tecnologia de genotipagem por sequenciamento (GBS, *Genotyping by sequencing*). A bioinformática básica realizada pelo Instituto de Diversidade Genômica da Universidade de Cornell (*Buckler Laboratory*) forneceu um total de 516.240 SNPs distribuídos nos 12 cromossomos de arroz, utilizando uma frequência alélica mínima (FMA) de 0,01.

Tendo em vista a redução de alelos raros e uma maior confiabilidade dos dados para o mapeamento associativo, foi utilizada uma FMA de 0,05, obtendo- se assim um total de 285.379 SNPs nos 12 cromossomos de arroz (Tabela 7).

Tabela 7. Número de marcadores SNPs obtidos pela genotipagem de 283 acessos de arroz de terras altas da CNAE por GBS.

CROMOSSOMO	FMA = 0,01	FMA = 0,05
1	68866	37233
2	53217	29301
3	55374	28999
4	49842	27552
5	39588	20469
6	44376	25017
7	40913	22078
8	39421	22217
9	31866	17468
10	33344	19559
11	41780	25110
12	17653	10376
Total	516.240	285.379

4.3 ESTRUTURA GENÉTICA POPULACIONAL

Aproximadamente 1.000 SNPs, obtidos através de filtragem da frequência mínima alélica do conjunto de dados original, foram utilizados para avaliar a estruturação dos dados utilizados no mapeamento associativo, através do *software* Structure (Pritchard et al., 2000). O método descrito por Evanno et al. (2005) foi utilizado para se determinar o número efetivo de subgrupos (K) na população, através do software Structure Harvester v. 0.6 (Earl, 2011). Como resultado, o valor de K obtido foi igual a 2, identificando, portanto, 2 subgrupos a partir do conjunto de dados total de SNPs.

4.4 ANÁLISE DE ASSOCIAÇÃO

A análise de mapeamento associativo foi realizada a partir dos dados de fenotipagem para o potencial produtivo em dois ensaios, com e sem aplicação de déficit hídrico, juntamente com os dados de genotipagem por GBS dos 283 acessos de sequeiro da CNAE.

A análise de associação foi baseada na abordagem de MLM (Modelo Linear Misto), em que foram utilizadas as informações de parentesco (K) e estrutura da população (Q), reduzindo com isto o erro tipo I. Sendo assim, os SNPs identificados foram associados aos caracteres avaliados nos experimentos com e sem deficiência hídrica (Produtividade e Índice de Suscetibilidade à Seca - ISS).

Dos 285.379 SNPs identificados nos 12 cromossomos a partir da FMA de 0,05, 48 marcadores SNPs (aproximadamente 0,016%) foram associados significativamente a um dos caracteres avaliados em ambos os ambientes (com e sem deficiência hídrica) (Tabela 8). Foram verificadas 13 associações entre SNPs e o ISS (Índice de Suscetibilidade à Seca) e 35 associações entre SNPs e a produtividade em ambiente com deficiência hídrica (Tabelas 9 e 10).

Tabela 8. Associações presentes em cada um dos 12 cromossomos.

Cromossomo	Número de associações
1	10
2	3
3	9
4	6
5	1
6	3
7	3
8	2
9	9
10	2
11	0
12	0
TOTAL	48

Tabela 9. Associações (SNPs – caractere) e suas significâncias (Cromossomos 1 ao 3).

Caractere	SNP	Cromossomo	Posição	P-valor	Q-valor	R^2	Var. genética
Produtividade_ESTRESSE_2010	S1_31622748	1	31622748	2,04E-09	3,4854E-04	0,224122091	338823,6124
Produtividade_ESTRESSE_2010	S1_42199857	1	42199857	3,48E-09	3,4854E-04	0,172492635	338823,6124
Produtividade_ESTRESSE_2010	S1_23885433	1	23885433	6,61E-08	4,4135E-03	0,148900814	338823,6124
Produtividade_ESTRESSE_2010	S1_23885324	1	23885324	1,66E-07	8,3129E-03	0,138725818	338823,6124
ISS_2010	S1_40258448	1	40258448	8,61E-07	3,0314E-02	0,120539607	0,753560933
ISS_2010	S1_16062057	1	16062057	9,08E-07	3,0314E-02	0,144425421	0,753560933
Produtividade_ESTRESSE_2010	S1_38138811	1	38138811	1,19E-06	3,4053E-02	0,122370861	338823,6124
ISS_2010	S1_30824493	1	30824493	1,80E-06	3,6966E-02	0,111678631	0,753560933
ISS_2010	S1_12033317	1	12033317	1,95E-06	3,6966E-02	0,118498599	0,753560933
ISS_2010	S1_27076684	1	27076684	2,03E-06	3,6966E-02	0,118432311	0,753560933
ISS_2010	S2_20530291	2	20530291	2,41E-08	3,7388E-03	0,187705558	0,753560933
Produtividade_ESTRESSE_2010	S2_1344963	2	1344963	8,22E-08	6,3761E-03	0,158687067	338823,6124
Produtividade_ESTRESSE_2010	S2_35190297	2	35190297	4,15E-07	2,1460E-02	0,407307094	338823,6124
Produtividade_ESTRESSE_2010	S3_5310087	3	5310087	5,60E-09	9,0616E-04	0,273977303	338823,6124
Produtividade_ESTRESSE_2010	S3_3311741	3	3311741	8,58E-08	6,9418E-03	0,141906645	338823,6124
Produtividade_ESTRESSE_2010	S3_5318886	3	5318886	4,39E-07	2,2039E-02	0,178480921	338823,6124
Produtividade_ESTRESSE_2010	S3_16573069	3	16573069	5,49E-07	2,2039E-02	0,140720975	338823,6124
Produtividade_ESTRESSE_2010	S3_14119502	3	14119502	6,81E-07	2,2039E-02	0,164012381	338823,6124
Produtividade_ESTRESSE_2010	S3_20945176	3	20945176	8,49E-07	2,2897E-02	0,220192256	338823,6124
Produtividade_ESTRESSE_2010	S3_15583251	3	15583251	1,72E-06	3,1877E-02	0,432426051	338823,6124
ISS_2010	S3_20799716	3	20799716	1,87E-06	3,1877E-02	0,111292419	0,753560933
ISS_2010	S3_9946984	3	9946984	1,97E-06	3,1877E-02	0,111328208	0,753560933

Tabela 10. Associações (SNPs – Caractere) e suas significâncias (Cromossomos 4 ao 10).

Caractere	SNP	Cromossomo	Posição	P-valor	Q-valor	R^2	Var. genética
Produtividade_ESTRESSE_2010	S4_35281220	4	35281220	7,44E-08	1,0685E-02	0,141592434	338823,6124
ISS_2010	S4_5915544	4	5915544	3,25E-07	1,5558E-02	0,159262409	0,753560933
Produtividade_ESTRESSE_2010	S4_10892731	4	10892731	5,65E-07	1,7635E-02	0,464899798	338823,6124
Produtividade_ESTRESSE_2010	S4_10892742	4	10892742	6,14E-07	1,7635E-02	0,461464752	338823,6124
Produtividade_ESTRESSE_2010	S4_27195948	4	27195948	1,49E-06	3,5663E-02	0,111961414	338823,6124
ISS_2010	S4_34524598	4	34524598	2,33E-06	4,7802E-02	0,117772254	0,753560933
Produtividade_ESTRESSE_2010	S5_21266402	5	21266402	7,29E-08	7,8872E-03	0,14283999	338823,6124
Produtividade_ESTRESSE_2010	S6_24319610	6	24319610	2,01E-08	1,7613E-03	0,300840846	338823,6124
Produtividade_ESTRESSE_2010	S6_24319611	6	24319611	2,65E-08	1,7613E-03	0,293206137	338823,6124
Produtividade_ESTRESSE_2010	S6_11453278	6	11453278	5,15E-08	2,2819E-03	0,265797425	338823,6124
Produtividade_ESTRESSE_2010	S7_7475601	7	7475601	1,77E-08	2,1559E-03	0,161713714	338823,6124
Produtividade_ESTRESSE_2010	S7_27697339	7	27697339	6,02E-08	3,6662E-03	0,144919005	338823,6124
Produtividade_ESTRESSE_2010	S7_27803592	7	27803592	1,01E-07	4,1006E-03	0,175701527	338823,6124
Produtividade_ESTRESSE_2010	S8_27223483	8	27223483	5,77E-09	6,7665E-04	0,497497474	338823,6124
Produtividade_ESTRESSE_2010	S8_17506042	8	17506042	5,65E-07	3,3129E-02	0,170117842	338823,6124
Produtividade_ESTRESSE_2010	S9_18281732	9	18281732	1,84E-08	1,3788E-03	0,153123371	338823,6124
Produtividade_ESTRESSE_2010	S9_10768968	9	10768968	3,04E-08	1,3788E-03	0,221085546	338823,6124
Produtividade_ESTRESSE_2010	S9_10160417	9	10160417	6,50E-08	1,9653E-03	0,235782352	338823,6124
Produtividade_ESTRESSE_2010	S9_15027131	9	15027131	2,92E-07	6,6217E-03	0,280849162	338823,6124
ISS_2010	S9_5922759	9	5922759	9,58E-07	1,4483E-02	0,218479052	0,753560933
ISS_2010	S9_5922760	9	5922760	9,58E-07	1,4483E-02	0,218479052	0,753560933
ISS_2010	S9_7996621	9	7996621	1,90E-06	2,4621E-02	0,11298558	0,753560933
Produtividade_ESTRESSE_2010	S9_15560708	9	15560708	4,09E-06	4,1222E-02	0,206889421	338823,6124
Produtividade_ESTRESSE_2010	S9_15560717	9	15560717	4,09E-06	4,1222E-02	0,206889421	338823,6124
Produtividade_ESTRESSE_2010	S10_14409085	10	14409085	6,21E-08	6,5559E-03	0,198362058	338823,6124
Produtividade_ESTRESSE_2010	S10_18591273	10	18591273	6,60E-07	3,4838E-02	0,138819034	338823,6124

4.4.1 Distribuição dos valores de significância

Com a finalidade de observar a dispersão do grande número de dados oriundos da genotipagem dos 283 acessos da CNAE, foram obtidos gráficos conhecidos como Manhattan Plot. Eles foram elaborados a partir das coordenadas genômicas das associações exibidas ao longo do eixo X e com o logaritmo negativo do P-valor da associação exibido no eixo Y.

No gráfico foi delimitado um P-valor de 0,05 para considerar as associações significativas. Como as associações mais fortes têm os menores P-valores, consequentemente possuem os logaritmos de menor valor (eixo Y). Desse modo, as associações acima do valor marcado foram consideradas estatisticamente significativas na análise de mapeamento associativo, resultando em 48 associações (Figuras 6 e 7).

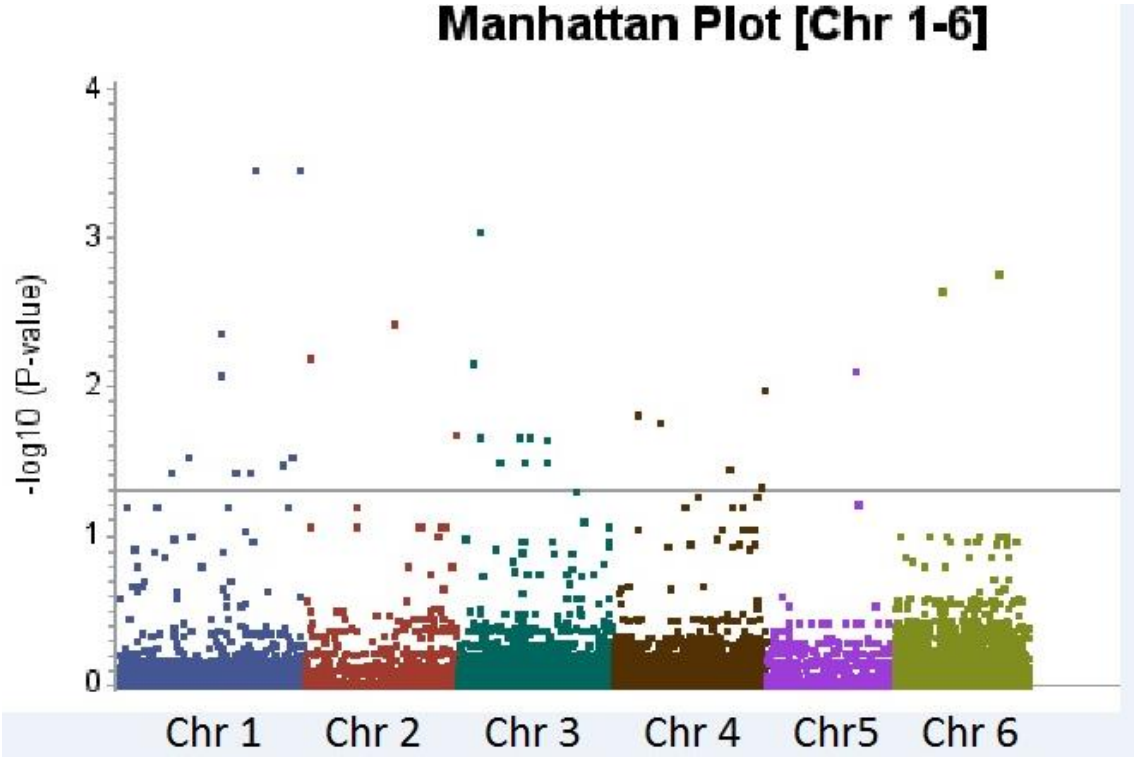

Figura 6. Gráfico Manhattan Plot dos valores de significância das associações nos cromossomos 1 ao 6.

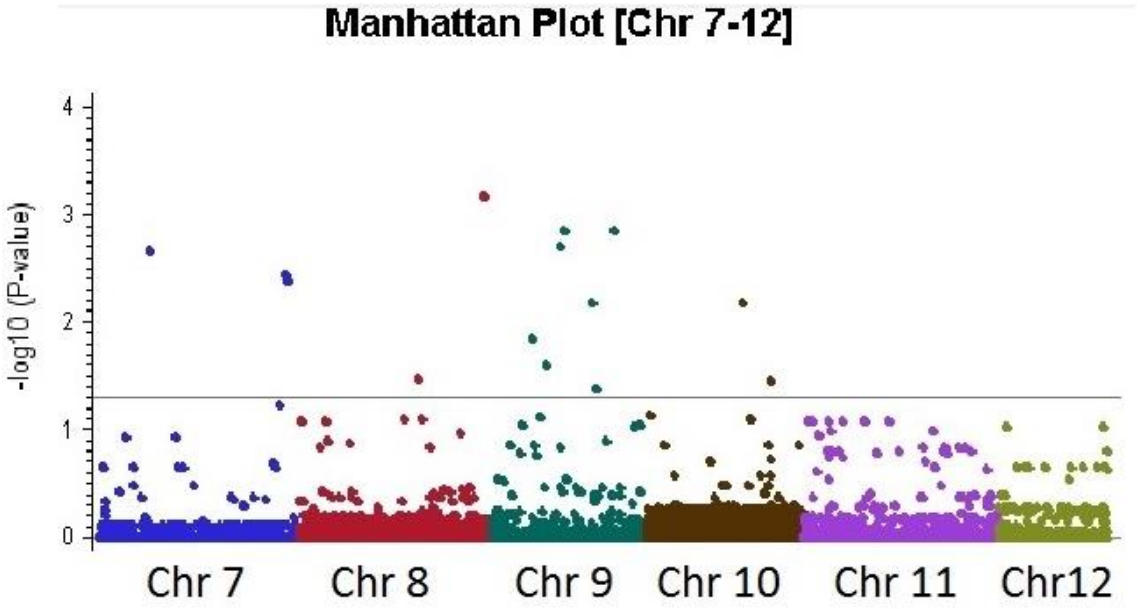

Figura 7. Gráfico Manhattan Plot dos valores de significância das associações nos cromossomos 7 ao 12.

4.5 ANOTAÇÃO DOS GENES

A posição de cada um dos 48 SNPs que apresentaram associações significativas na análise de mapeamento associativo foi utilizada como âncora para a identificação de genes. O genoma de referência utilizado foi o da cultivar Nipponbare (MSU Rice Genome versão 7.0). Foram identificados 35 SNPs em 31 genes de arroz. Dentre os genes identificados, sete deles continham SNPs associados ao Índice de Suscetibilidade à Seca, enquanto que os outros 24 genes continham SNPs associados à produtividade dos acessos em ambiente com deficiência hídrica.

O cromossomo 3 apresentou o maior número de genes que continham SNPs identificados pela análise de associação, enquanto que nos cromossomos 11 e 12 não foram identificados SNPs associados a genes (Tabela 11).

A partir da informação da sequência dos genes, através da análise de BLASTx foi possível identificar as proteínas relacionadas aos transcritos identificados (Tabela 12). A partir dos códigos dos genes (Ids) também foi possível a busca de termos descritos no Gene Ontology (GO *terms*) (Tabela 13). Adicionalmente, os transcritos dos genes identificados foram relacionados aos prováveis processos biológicos e as funções moleculares através da base de dados agriGO (Tabela 14).

Tabela 11. Transcritos identificados em cada cromossomo do arroz cultivar Nipponbare.

Cromossomo	Transcritos
1	5
2	2
3	7
4	3
5	1
6	1
7	3
8	1
9	6
10	2
11	0
12	0
Total	31

Tabela 12. Genes identificados a partir da presença de marcadores SNPs provenientes da análise de mapeamento associativo e seus respectivos produtos expressos.

Gene	Produto Expresso
LOC_Os01g28690.2	nucleoporin autopeptidase domain containing protein, expressed
LOC_Os01g42130.1	expressed protein
LOC_Os01g54990.1	auxin response factor, expressed
LOC_Os01g69270.1	OsFBO2 - F-box and other domain containing protein, expressed
LOC_Os01g72720.1	expressed protein
LOC_Os02g03330.1	expressed protein
LOC_Os02g34310.1	expressed protein
LOC_Os03g06570.1	IQ calmodulin-binding motif family protein, expressed
LOC_Os03g10440.1	glycosyl hydrolase family 10 protein, expressed
LOC_Os03g17850.1	glycosyltransferase family 43 protein, expressed
LOC_Os03g24770.1	expressed protein
LOC_Os03g29170.1	sterol-4-alpha-carboxylate 3-dehydrogenase, decarboxylating, expressed
LOC_Os03g37490.1	MATE efflux family protein, expressed
LOC_Os03g37780.1	expressed protein
LOC_Os04g19570.1	retrotransposon protein, Ty3-gypsy subclass, expressed
LOC_Os04g45920.1	protein kinase domain containing protein, expressed
LOC_Os04g59340.1	RNA-binding region RNP-1, expressed
LOC_Os05g35840.1	expressed protein
LOC_Os06g19980.1	MYB family transcription factor, expressed
LOC_Os07g13030.1	expressed protein
LOC_Os07g46410.1	bifunctional thioredoxin reductase/thioredoxin, expressed
LOC_Os07g46555.1	F-box domain containing protein, expressed
LOC_Os08g43070.1	helix-loop-helix DNA-binding domain containing protein, expressed
LOC_Os09g10840.2	transcription factor, expressed
LOC_Os09g13680.1	ZOS9-04 - C2H2 zinc finger protein, expressed
LOC_Os09g16550.1	ankyrin repeat family protein, expressed
LOC_Os09g17610.1	RING finger protein, expressed
LOC_Os09g25910.1	xylanase inhibitor, expressed
LOC_Os09g30070.1	expressed protein
LOC_Os10g27310.1	expressed protein
LOC_Os10g34820.2	CDT1B - DNA replication initiation protein, expressed

Tabela 13. Genes identificados a partir da presença de marcadores SNPs provenientes da análise de mapeamento associativo e os prováveis processos biológicos de seus transcritos.

Gene	Processo Biológico
LOC_Os01g28690.2	transport (GO:0006810)
LOC_Os01g42130.1	NÃO CLASSIFICADO
LOC_Os01g54990.1	transcription (GO:0006350)
LOC_Os01g69270.1	biological_process (GO:0008150)
LOC_Os01g72720.1	NÃO CLASSIFICADO
LOC_Os02g03330.1	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0006139)
LOC_Os02g34310.1	NÃO CLASSIFICADO
LOC_Os03g06570.1	NÃO CLASSIFICADO
LOC_Os03g10440.1	carbohydrate metabolic process (GO:0005975)
LOC_Os03g17850.1	biosynthetic process (GO:0009058)
LOC_Os03g24770.1	NÃO CLASSIFICADO
LOC_Os03g29170.1	lipid metabolic process (GO:0006629)
LOC_Os03g37490.1	transport (GO:0006810)
LOC_Os03g37780.1	NÃO CLASSIFICADO
LOC_Os04g19570.1	DNA integration (GO:0015074)
LOC_Os04g45920.1	protein modification process (GO:0006464)
LOC_Os04g59340.1	NÃO CLASSIFICADO
LOC_Os05g35840.1	NÃO CLASSIFICADO
LOC_Os06g19980.1	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0006139)
LOC_Os07g13030.1	NÃO CLASSIFICADO
LOC_Os07g46410.1	response to stress (GO:0006950)
LOC_Os07g46555.1	NÃO CLASSIFICADO
LOC_Os08g43070.1	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0006139)
LOC_Os09g10840.2	signal transduction (GO:0007165)
LOC_Os09g13680.1	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0006139)
LOC_Os09g16550.1	NÃO CLASSIFICADO
LOC_Os09g17610.1	NÃO CLASSIFICADO
LOC_Os09g25910.1	response to abiotic stimulus (GO:0009628)
LOC_Os09g30070.1	NÃO CLASSIFICADO
LOC_Os10g27310.1	NÃO CLASSIFICADO
LOC_Os10g34820.2	protein modification process (GO:0006464)

Tabela 14. Genes identificados a partir da presença de marcadores SNPs provenientes da análise de mapeamento associativo e as prováveis funções moleculares de seus transcritos.

Gene	Função Molecular
LOC_Os01g28690.2	transporter activity (GO:0005215)
LOC_Os01g42130.1	NÃO CLASSIFICADO
LOC_Os01g54990.1	sequence-specific DNA binding transcription factor activity (GO:0003700)
LOC_Os01g69270.1	binding (GO:0005488)
LOC_Os01g72720.1	NÃO CLASSIFICADO
LOC_Os02g03330.1	sequence-specific DNA binding transcription factor activity (GO:0003700)
LOC_Os02g34310.1	NÃO CLASSIFICADO
LOC_Os03g06570.1	protein binding (GO:0005515)
LOC_Os03g10440.1	hydrolase activity (GO:0016787)
LOC_Os03g17850.1	transferase activity (GO:0016740)
LOC_Os03g24770.1	NÃO CLASSIFICADO
LOC_Os03g29170.1	catalytic activity (GO:0003824)
LOC_Os03g37490.1	transporter activity (GO:0005215)
LOC_Os03g37780.1	NÃO CLASSIFICADO
LOC_Os04g19570.1	zinc ion binding (GO:0008270)
LOC_Os04g45920.1	kinase activity (GO:0016301)
LOC_Os04g59340.1	NÃO CLASSIFICADO
LOC_Os05g35840.1	NÃO CLASSIFICADO
LOC_Os06g19980.1	sequence-specific DNA binding transcription factor activity (GO:0003700)
LOC_Os07g13030.1	NÃO CLASSIFICADO
LOC_Os07g46410.1	enzyme regulator activity (GO:0030234)
LOC_Os07g46555.1	NÃO CLASSIFICADO
LOC_Os08g43070.1	sequence-specific DNA binding transcription factor activity (GO:0003700)
LOC_Os09g10840.2	DNA binding (GO:0003677)
LOC_Os09g13680.1	sequence-specific DNA binding transcription factor activity (GO:0003700)
LOC_Os09g16550.1	NÃO CLASSIFICADO
LOC_Os09g17610.1	binding (GO:0005488)
LOC_Os09g25910.1	hydrolase activity (GO:0016787)
LOC_Os09g30070.1	NÃO CLASSIFICADO
LOC_Os10g27310.1	NÃO CLASSIFICADO
LOC_Os10g34820.2	kinase activity (GO:0016301)

5 DISCUSSÃO

5.1 ANÁLISE DOS CARACTERES FENOTÍPICOS

A seca é um dos estresses ambientais mais importantes para a agricultura no mundo. O desenvolvimento de genótipos mais tolerantes a esse estresse, portanto, é um dos maiores desafios dos programas de melhoramento genético de plantas. A análise dos dados fenotípicos do experimento em ambiente com e sem deficiência hídrica foram ajustados pelo método de modelos mistos. Como base para esse ajuste foram utilizadas as testemunhas BRSMG Curinga, BRS Pepita e BRS Sertaneja.

Esse experimento revelou que os efeitos entre os genótipos e os ambientes foram significativos, indicando especificidades adaptativas de certos acessos às condições de seca. A produtividade média dos acessos em condição com disponibilidade de água no solo mostrou-se superior à condição sob deficiência hídrica (2385,47 $kg.ha^{-1}$ e 1081,4 $kg.ha^{-1}$, respectivamente), equivalendo a uma redução de aproximadamente 55% na produtividade, que representa um padrão severo de deficiência hídrica para materiais de ciclo curto, de acordo com Heinemann (2010). Essa redução no rendimento da produtividade é esperada, principalmente porque a restrição de água foi aplicada no início da fase reprodutiva, o que reduz o número de panículas por m^2, o número de espiguetas por panícula, uma proporção significativa do número de grãos cheios e um aumento considerável no número de espiguetas estéreis (Jongdee et al., 2002).

A distribuição de frequências da produtividade no ambiente sob déficit hídrico apresentou a classe mais frequente (942,62 $kg.ha^{-1}$ a 1242,62 $kg.ha^{-1}$) contendo o valor da média para essa condição, enquanto que no ambiente sem deficiência hídrica também se observou a média de produtividade contida na classe mais frequente (2114,82 $kg.ha^{-1}$ a 2614,82 $kg.ha^{-1}$).

Analisando os gráficos Boxplot para o caráter produtividade em ambos os ambientes, verificou-se uma grande diferença na dispersão dos dados. Em ambiente sem deficiência hídrica, observou-se uma maior dispersão dos dados variando de 614,82 $kg.ha^{-1}$ a 4731,13 $kg.ha^{-1}$, enquanto que em ambiente com estresse hídrico, observou-se a ação da deficiência hídrica, que além de reduzir a produtividade média, diminuiu a dispersão dos dados, causando um "achatamento" verificado pela variação de 519,26 $kg.ha^{-1}$ a 1624,08

kg.ha^{-1} na produtividade dos acessos. Nesse caso, a resposta ao déficit hídrico reduziu a dispersão dos dados de produtividade, ou seja, os acessos tiveram um comportamento bastante semelhante frente ao estresse. Em contrapartida, no ambiente com fornecimento adequado de água, os acessos tendem a expressar de modo diferenciado seu potencial produtivo.

Outra característica avaliada nesse trabalho foi o Índice de Suscetibilidade à Seca (ISS), que é um modelo matemático elaborado para estimar o rendimento dos genótipos sob déficit hídrico, assegurando que os acessos selecionados terão caracteres para a tolerância à seca (Fischer & Maurer, 1978). Essa característica apresenta bastante relevância para a seleção de plantas com alelos superiores sob condições de estresse.

A distribuição de frequências do ISS, calculado a partir dos valores de produtividade em ambos os ambientes, apresentou a classe mais frequente de 0,94 a 1,34. Esse valor de ISS em torno de 1,0 representa uma redução de aproximadamente 50% da produtividade de cada acesso quando condicionado ao déficit hídrico. A dispersão dos dados de ISS foi observada pelo gráfico Boxplot, que variou de -0,1534 a 1,5739, ou seja, houve materiais que reduziram em mais de 50% sua produtividade, mostrando-se extremamente sensíveis a essas condições (por exemplo, o acesso CT11632-3-3-M). Por outro lado, houve materiais que não alteraram sua produtividade com a redução na disponibilidade de água, como o IRAT 13.

Foram identificados os 20 materiais mais produtivos em ambiente com e sem deficiência hídrica, dentre os quais três acessos apresentaram maiores produtividades em ambas as condições: IRAT 141 (linhagen melhorada no exterior), CARISMA (cultivar brasileira lançada em 1999) e MATO GROSSO (variedade tradicional coletada em Santa Catarina em 2001), indicando, portanto, materiais que podem ser bem explorados em programas de melhoramento. De acordo com Bueno (2012), as médias ajustadas em experimentos conduzidos em oito locais no Brasil para esses acessos foram, respectivamente, 4.363, 3.547 e 3.821 Kg.ha^{-1}. A média geral dos 550 acessos avaliados nesses locais foi de 3.450 Kg.ha^{-1}, ou seja, são genótipos com desempenho superior com relação à produtividade, e que podem ser indicados para avaliações adicionais pelo programa de melhoramento genético de arroz da Embrapa para uso como genitores no bloco de cruzamentos para o desenvolvimento de futuras linhagens e cultivares.

No ambiente sob deficiência hídrica, dentre os 20 materiais mais produtivos, nove eram Variedades Tradicionais, três eram Linhagens Brasileiras e oito eram Linhagens

Introduzidas, ou seja, apenas 15% dos materiais mais tolerantes à seca foram desenvolvidos por programas de melhoramento genético do arroz brasileiro. Obviamente um programa de melhoramento utiliza uma série de caracteres que tornam uma cultivar apta para o cultivo e aceitação no mercado. Contudo, especificamente para a tolerância à seca, existe um grande potencial de melhoria de desempenho para essas cultivares, e a utilização de acessos mais tolerantes à seca identificados por esse trabalho pode ser o ponto de partida para atingir esse resultado.

No ambiente com irrigação normal, dentre os 20 materiais mais produtivos, quatro eram Variedades Tradicionais, sete eram Linhagens Brasileiras (material melhorado) e nove eram Linhagens Introduzidas. Nesse caso, 35% dos acessos mais produtivos foram desenvolvidos por programas nacionais de melhoramento genético, evidenciando a importância da incorporação de genótipos mais tolerantes à seca, que pode resultar em impactos econômicos positivos para a agricultura brasileira.

Foram identificados também os 20 materiais mais sensíveis à seca no experimento, e dentre esses, nove acessos mais sensíveis também estiveram entre os 20 mais produtivos em ambiente sem deficiência hídrica. Todos os nove acessos foram oriundos de programas de melhoramento do Brasil e exterior, ou seja, esses materiais foram desenvolvidos para uma condição ótima de disponibilidade hídrica. Constata-se com isso que os programas de melhoramento genético de arroz de terras altas devem aumentar seu foco no desenvolvimento de linhagens e cultivares mais produtivas tanto sob condições hídricas ótimas, quanto sob condições de déficit hídrico.

Em geral, os acessos menos sensíveis à seca (menor ISS) podem representar o ponto de partida para selecionar materiais superiores de arroz sob deficiência hídrica, auxiliando na identificação de genótipos que expressem características morfofisiológicas específicas para a adaptabilidade à seca (Pantuwan et al., 2002). Segundo Bernier et al. (2007), a seleção dos genótipos tolerantes à seca deve conduzir à melhoria igual do rendimento sob ambas condições, ou seja, o material selecionado deve ser produtivo baseando-se diretamente na condição com deficiência hídrica e indiretamente na condição sem esse estresse.

5.2 DADOS GENOTÍPICOS

O surgimento das tecnologias de sequenciamento de nova geração (NGS) tem impulsionado os estudos genômicos funcionais em arroz. O arroz é considerado um organismo modelo para estudos genéticos e o acúmulo rápido de informações genômicas tem sido facilitado, dentre outros motivos, pela alta qualidade dos genomas de referência das duas subespécies (Indica e Japônica), o que permite um alinhamento rápido e preciso com pequenos *reads* provenientes das tecnologias NGS. Tendo em vista o número relativamente modesto de sequências repetitivas, uma grande proporção desses pequenos *reads* pode ser precisamente mapeada no genoma de referência do arroz. O sistema reprodutivo autógamo em conjunto com o genoma relativamente pequeno do arroz, permite o sequenciamento de indivíduos a uma baixa cobertura, porém com uma alta eficiência.

Nesse trabalho, a tecnologia de genotipagem por sequenciamento (GBS) foi utilizada para identificar SNPs nos 283 acessos componentes do painel de associação. Os comandos (*Pipeline*) utilizados para a identificação e posterior validação desses marcadores resultaram em um total de 516.240 SNPs distribuídos nos 12 cromossomos do arroz.

Atualmente, o grande gargalo diz respeito à análise de um grande volume de dados, pois requer grande capacidade computacional, além de profissionais com experiência em informática que possam manipular tamanho volume de informação. A utilização de servidores (*clusters*) tornou-se praticamente obrigatória para o trabalho com grande volume de dados de sequências de genoma. Outra questão que deve ser tratada com cautela é que a tecnologia de GBS, apesar de gerar um grande número de dados de SNPs, resulta em um grande número de dados faltantes, sendo necessário realizar a imputação desses dados, ou seja, estimar qual o provável nucleotídeo estaria faltando em determinado ponto do genoma. Uma alternativa para solucionar esse problema e aumentar a precisão da análise GBS é a utilização do *software* FastPHASE, que estima genótipos faltantes e reconstrói haplótipos a partir de dados de SNPs de indivíduos não aparentados, conforme descrito por Scheet & Stephens (2006).

Para aumentar a confiabilidade da análise baseada nos marcadores SNPs, foi realizada uma "filtragem" dos dados com uma frequência alélica menor que 0,05, resultando em 285.379 SNPs presentes nos 12 cromossomos de arroz (redução de 55% do

número inicial de SNPs). Essa redução de SNPs ainda manteve uma grande quantidade de informação, ao mesmo tempo em que preveniu possíveis associações com alelos raros, aumentando a confiabilidade da análise de associação entre SNPs e caracteres fenotípicos.

O cromossomo 1 foi o que apresentou mais SNPs identificados após a filtragem (37.233 SNPs **),** enquanto que os cromossomos 9, 10 e 12 apresentaram menos SNPs **(**17.468, 19.559 e 10.376, respectivamente). Segundo os dados dos tamanhos dos cromossomos disponíveis no *Rice Genome Annotation Project* (http://rice.plantbiology.msu.edu/index.shtml), uma explicação para a quantidade de SNPs identificados seria em relação ao tamanho em pares de bases de cada cromossomo. O cromossomo 1, por exemplo, que resultou no maior número de SNPs detectados, também é o maior cromossomo do arroz (43.270.923 pb). Em contrapartida, cromossomos 9**,** 10 e 12, que apresentaram a menor quantidade de SNPs**,** também são os menores cromossomos do arroz **(**23.012.720 pb, 23.207.287 pb e 27.531.856 pb, respectivamente). Pode-se concluir com isso que o número de SNPs detectados por cromossomo, considerando um conjunto de genótipos geneticamente divergentes, está diretamente relacionado com o tamanho desses cromossomos.

Com base nos dados de SNPs dos 283 acessos utilizados no mapeamento associativo, foi realizada uma análise de estruturação, em que se verificou um "K" (número de subconjuntos) igual a dois. Esses dois subgrupos podem ser explicados pela utilização de materiais provenientes de linhagens melhoradas e variedades tradicionais (Garris et al., 2005).

5.3 ANÁLISE DE ASSOCIAÇÃO

A análise de associação emergiu como uma poderosa abordagem para identificar genes relacionados a características complexas em humanos (Altshulher et al.; 2008). A aplicação desse método é plenamente viável para o arroz, principalmente devido ao seu pequeno genoma e sua grande diversidade genética, permitindo a identificação de regiões haplotípicas que possibilitam a obtenção precisa de associações entre marcadores moleculares e caracteres fenotípicos.

Estudos de mapeamento associativo já listaram associações entre marcadores e 26 características agronômicas importantes para a cultura do arroz (Jena et al., 2008). Neste trabalho, os dados de SNPs provenientes da tecnologia GBS foram

associados aos caracteres avaliados nos experimentos com e sem deficiência hídrica (Produtividade e Índice de Suscetibilidade à Seca - ISS) no painel contendo os 283 acessos do sistema de cultivo sequeiro da Coleção Nuclear de Arroz da Embrapa (CNAE).

A partir da análise com 285.379 SNPs distribuídos nos 12 cromossomos de arroz, através da abordagem de modelos mistos, 48 desses marcadores foram associados de forma significativa aos caracteres avaliados. A grande vantagem da análise de associação é a determinação específica da posição do loco de interesse e a proporção da variação no caráter de interesse que o polimorfismo no loco consegue explicar. Este tipo de análise também permite a identificação do alelo favorável e a estimação do efeito associado à sua presença sobre a expressão do caráter.

Dentre os 48 SNPs estatisticamente significativos, 13 foram associados ao ISS (Índice de Suscetibilidade à Seca) e 35 associados à produtividade em ambiente com deficiência hídrica. De acordo, com os gráficos do tipo Manhattan Plot, que evidenciam a distribuição dos p-valores dos SNPs associados, as cinco associações mais significativas foram entre os SNPs S1_31622748, S1_42199857, S8_27223483, S3_5310087 e S9_18281732 e o caráter produtividade em ambiente sob déficit hídrico, evidenciando uma importante relação entre polimorfismos no genoma relacionados com a maior produtividade dos acessos em condições de seca. Dos SNPs associados no mapeamento associativo, somente os cromossomos 11 e 12 não apresentaram associações.

Algumas associações primárias entre regiões genômicas e caracteres relacionados à produtividade e resistência à seca já foram descritas em arroz (Kamoshita et al., 2008). Em estudos com populações segregantes, foram encontrados quatro marcadores SNPs relacionados à tolerância à seca, nos cromossomos 1, 4, 5 e 7 (Ye et al.; 2012), e três regiões significativas para a produtividade nos cromossomos 1, 8 e 10 do arroz (Vikram et al.; 2012). Painéis de associação utilizando genótipos com reduzido vínculo genético, como o descrito nesse trabalho, ainda não foram utilizados em arroz para a identificação de locos relacionados à produtividade sob déficit hídrico. A utilização de populações sem vínculo genético favorece a identificação de uma série alélica dos genótipos que podem estar relacionados a caracteres de interesse, não ficando restrita aos alelos provenientes dos genitores de uma população segregante, podendo resultar em um maior número de associações, como foi identificado nesse trabalho (48 associações).

A identificação de SNPs associados a caracteres de interesse é informação suficiente para o desenvolvimento de marcadores baseados em chips de DNA, os quais

podem auxiliar na seleção assistida pela genotipagem de linhagens de programas de melhoramento genético do arroz. Contudo, é apenas o primeiro passo caso o objetivo seja a identificação de genes em rotas metabólicas relacionadas com o caráter em estudo.

No presente estudo só foram detectados SNPs associados à produtividade sob déficit hídrico. Contudo, houve SNPs associados ao ISS, que leva em conta a produtividade nas duas condições de disponibilidade hídrica. Um mesmo marcador SNP não foi relacionado a mais de um caractere simultaneamente, no entanto, diferentes SNPs foram associados ao mesmo caractere, podendo indicar a ocorrência de diferentes rotas metabólicas associadas ao fenótipo observado.

5.4 ANOTAÇÃO DOS GENES

A ocorrência de estresse de seca ocasiona uma grande quantidade de respostas fisiológicas e bioquímicas na planta, influenciando simultaneamente a expressão de genes de diferentes rotas metabólicas. Com o advento das tecnologias de sequenciamento de nova geração (NGS), a identificação de genes associados aos mecanismos de resposta à seca de modo rápido e preciso tornou-se possível, gerando um grande avanço nos estudos de genômica funcional. Neste trabalho, o mapeamento associativo forneceu 48 SNPs associados aos caracteres produtividade sob deficiência hídrica e ISS. Desses 48 SNPs, aproximadamente 73% deles (35 SNPs) estavam contidos em regiões codantes do genoma de arroz, disponível no *Rice Genome Annotation Project* (http://rice.plantbiology.msu.edu/index.shtml).

Dessas regiões genômicas codantes, alguns SNPs estavam presentes no mesmo gene, resultando em 31 genes identificados. Desse total, sete genes apresentavam SNPs relacionados com o Índice de Suscetibilidade à Seca (ISS) e 24 genes apresentavam SNPs relacionados com a produtividade dos acessos sob deficiência hídrica.

Dentre os 31 genes descritos, 10 deles apresentaram, como produto, proteínas expressas cuja função e processo biológicos ainda não foram determinados. Dos sete genes relacionados com o ISS, pode-se destacar o gene LOC_Os09g13680.1, que codifica a proteína do tipo Dedo de Zinco ZOS9-04 - C2H2, que atua como fator de transcrição no processo metabólico dos ácidos nucléicos. Proteínas do tipo Dedo de Zinco (Zinc finger protein, ou ZFPs) são um grupo que se expressa em diferentes condições de estresse, sendo que a proteína expressa C2H2 (Cys2/His2 finger protein) é induzida por

vários tipos de estresses abióticos, representando um importante gene candidato relacionado à tolerância à seca.

Outro gene de interesse dentre os sete relacionados ao ISS é o LOC_Os03g37490.1, que codifica a proteína da família MATE efflux (*multidrug and toxic compound extrusion*) , o qual é responsável por proteger as células de compostos tóxicos (Brown et al., 1999). Durante a ocorrência de mecanismos de resposta decorrentes da deficiência hídrica, uma série de compostos dispensáveis acabam sendo gerados, sendo assim, esse produto expresso poderia atuar na retirada desses compostos, auxiliando no reparo dos danos ocasionados pela seca.

O gene LOC_Os01g69270.1 também apresenta relevância por codificar a proteína OsFBO2 que contém o domínio F-Box, que atua como mediadora na ubiquitinação de proteínas para degradação pelo proteossoma, além de estar associada a funções celulares como transdução de sinal e regulação do ciclo celular, podendo ter extrema importância nas rotas metabólicas relacionadas à produtividade sob déficit hídrico.

Dos 24 genes relacionados com à produtividade em ambiente com déficit hídrico, pode-se destacar quatro genes: LOC_Os09g16550.1, LOC_Os06g19980.1, LOC_Os04g45920.1 e LOC_Os01g54990.1. O gene LOC_Os09g16550.1 codifica uma proteína com domínio de regiões repetidas de Anquirina (proteínas do citoesqueleto), que são comuns na interação proteína-proteína (Stogios et al., 2005), tendo diversas funções, como regulação da transcrição, transporte de íons e transdução de sinal (Mosavi el al., 2004; Breeden et al., 1987). Em estudos com *Arabidopsis*, uma proteína ACD6 contendo domínio de Anquirina foi relacionada com a regulação dos mecanismos de resposta contra estresses abióticos, como o alto teor de sal (Lu et al., 2003).

O gene LOC_Os06g19980.1 codifica uma proteína que funciona como fator de transcrição MYB, que constitui a mais extensa família de fatores de transcrição em plantas, auxiliando na regulação do ciclo celular, e sendo ativado somente após o acúmulo significativo de ácido abscísico (ABA) endógeno. O ABA é um hormônio vegetal que tem como função a regulação de vários aspectos ligados à fisiologia das plantas, tais como respostas ao déficit hídrico, e por esse motivo, esse gene tem grande probabilidade de ser um gene candidato relacionado à maior produtividade sob déficit hídrico.

O gene LOC_Os04g45920.1 codifica proteínas quinases, que são enzimas responsáveis pelo controle de vários processos celulares, incluindo respostas a estímulos ambientais. Segundo Dardick et al. (2007), há 1.429 proteínas quinases no arroz, e estão

relacionadas com regulação de mecanismos de estresse. Como exemplo, a proteína SAPK4 está envolvida na resposta ao estresse salino (Diédhiou et al., 2008), assim como a proteína TaSnRK2.4, que, de acordo com de Mao et al. (2010), foi responsável pelo aumento da tolerância ao déficit hídrico, alto teor salino e de congelamento.

Por fim, o gene LOC_Os09g16550.1 codifica uma proteína de fator resposta à auxina, que regula a ativação e repressão da transcrição, e portanto capazes de regular o desenvolvimento das plantas, dependendo do ambiente a que estão submetidas (Okushima et al., 2005).

6 CONCLUSÕES

Neste trabalho, foram avaliados 283 acessos provenientes do sistema de cultivo de sequeiro da Coleção Nuclear de Arroz da Embrapa (CNAE) em um ambiente com e e um ambiente sem deficiência hídrica. As médias gerais de produtividade revelaram uma diferença favorável aos acessos em cultivados em ambiente com irrigação normal em relação ao ensaio sob déficit hídrico, demonstrando que a aplicação desse tratamento foi efetiva.

Os materiais mais produtivos sob déficit hídrico foram IRAT 141, JAGUARY, AGULHA, IDSA 6 e AGULHÃO, enquanto que os acessos mais produtivos em ambiente sem deficiência hídrica foram CT11632-3-3-M, IPEACO 77-P, CABAÇU, IRAT 141 e CHATÃO AMARELO. Três materiais estavam presentes entre os 20 mais produtivos em ambos os ambientes: IRAT 141, CARISMA e MATO GROSSO.

A tecnologia de sequenciamento de nova geração (NGS) denominada Genotipagem por Sequenciamento (GBS) identificou 516.240 SNPs distribuídos nos 12 cromossomos do arroz através da análise de DNA dos 283 acessos do experimento. Um filtro de frequência mínima alélica de 0,05 foi aplicado, reduzindo o número de SNPs para 285.379, os quais foram utilizados no estudo de associação juntamente com os caracteres de interesse avaliados. O mapeamento associativo identificou 48 marcadores SNPs estatisticamente significativos associados aos dois caracteres fenotípicos avaliados, a produtividade e Índice de Suscetibilidade à Seca (ISS). Desses SNPs, 13 estavam associados ao ISS e 35 associados à produtividade em ambiente com deficiência hídrica.

As posições desses SNPs no genoma de referência (Cultivar MSU 7.0 Niponbare) foram utilizadas para identificar se estariam presentes em genes, sendo então posicionados em 31 genes. Dentre os genes que continham os SNPs associados ao ISS, os mais relevantes foram LOC_Os09g13680.1, LOC_Os06g19980.1 e LOC_Os01g69270.1, enquanto que dentre os genes que continham os SNPs associados à produtividade, os mais relevantes foram LOC_Os09g16550.1, LOC_Os06g19980.1, LOC_Os04g45920.1 e LOC_Os01g54990.1. Esses genes podem ser avaliados para serem efetivamente utilizados na seleção assistida por marcadores. Adicionalmente, esses genes podem ser superexpressos para avaliar sua capacidade de aumentar a tolerância à seca, e em caso positivo, gerar cultivares comerciais de arroz geneticamente modificadas mais tolerantes a esse estresse.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ABADIE, T.; CORDEIRO, C. M. T.; FONSECA, J. R.; ALVES, R. B. N.; BURLE, M. L.; BRONDANI, C.; RANGEL, P. H. N; CASTRO, E. M.; SILVA, H. T.; FREIRE, M. S.; ZIMMERMANN, F. J. P.; MAGALHÃES, J. R. S. O. Constructing a rice core collection for Brazil. **Pesquisa Agropecuária Brasileira**, Brasília, v. 40, n. 2, p. 129-136, 2005.

ABDALLAH, J. M.; GOFFINET, B.; CIERCO-AYROLLES, C.; PÉREZ-ENCISO, M. Linkage disequilibrium fine mapping of quantitative trait loci: A simulation study. **Genetics Selection Evolution**, Les Ulis, v. 35, n.5, p. 513-532, 2003.

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

ALLENDORF, F. W.; HOHENLOHE, P. A. & LUIKART, G. Genomics and the future of conservation genetics. **Nature Reviews Genetics**, v. 11, p. 697–709, 2010.

ALTSHULER, D.; POLLARA, V. J.; COWLES, C. R.; VAN ETTEN, W. J.; BALDWIN, J.; LINTON, L.; LANDER, E. S. An SNP map of the human genome generated by reduced representation shotgun sequencing. **Nature**, v. 407, p. 513–516, 2000.

ALTSHULER, D.; DALY, M. J.; LANDER, E. S. Genetic mapping in human disease. **Science**, v. 322, p. 881–888, 2008.

ARANZANA, M. J.; KIM, S.; ZHAO, K.; BAKKER, E.; HORTON, M.; JAKOB, K.; LISTER, C.; MOLITOR, J.; SHINDO, C.; TANG, C.; TOOMAJIAN, C.; TRAW, B.; ZHENG, H.; BERGELSON, J.; DEAN, C.; MARJORAM, P.; NORDBORG, M. Genome-Wide association mapping in *Arabidopsis* indentifies previously known flowering time and pathogen resistance genes. **Plos Genetics**, v. 5, p. 531-539, 2005.

BABU, R. C.; NGUYEN, B. D.; CHAMARERK, V.; SHANMUGASUNDARAM, P.; CHEZHIAN, P.; JEYAPRAKASH, P.; GANESH, S. K.; PALCHAMY, A.; SADASIVAM, S.; SARKARUNG, S.; WADE, L. J.; NGUYEN, H. T. Genetic analysis of drought resistance in rice by molecular markers: association between secondary traits and field performance. **Crop Science**, Madison, v. 43, n.4, p.1457-1469, jul.-ago. 2003.

BENEVENTI, M. A. **Transformação genética em soja por inserção da construção gênica contendo a região promotora do gene rd29A e a região codante do gene DREB1A de *Arabidopsis thaliana,* visando tolerância à seca**. 2006. 126 f. Dissertação (Mestrado em Genética e Biologia Molecular). Universidade Estadual de Londrina, UEL, Londrina, 2006.

BERNIER, J., KUMAR, A., RAMAIAH, V., SPANER, D., ATLIN, G. N. A large-effect QTL for grain yield under reproductive-stage drought stress in upland Rice. **Crop Science**, v. 47, n. 2, p. 507-518, mar. 2007.

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

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

BRAY, E. A. Genes commonly regulated by water-deficit stress in Arabidopsis thaliana. **Journal of Experimental Botany**, Chicago, v. 55, p. 2331-2341, 2004.

BREEDEN, L.; NASMYTH, K. Similarity between cell-cycle genes of budding yeast and fission yeast and the Notch gene of Drosophila. **Nature**, v. 329, p. 651-654, 1987.

BRONDANI, C.; RANGEL, P. H. N.; BRONDANI, R. P. V.; BORBA, T. C. O**. Coleção Nuclear de Arroz da Embrapa – Parte I:** Caracterização agronômica. Série Documentos 189. Embrapa, Santo Antônio de Goiás, 2006. 22 p.

BROWN, A. H. D. Core collections: a practical approach to genetic resources management. **Genome**, North York, v. 31, n. 2, p. 818-824, 1989.

BROWN, M. H.; PAULSEN, I. T. AND SKURRAY, R. A. The multidrug efflux protein NorM is a prototype of a new family of transporters. **Molecular Microbiology**, v. 31, p. 394–395, 1999.

BUENO, L. G. Adaptabilidade e estabilidade de acessos de uma coleção nuclear de arroz. **Pesquisa Agropecuária Brasileira**, Brasília, v.47, n.2, p.216-226, fev. 2012.

CALDWELL, K. S.; RUSSELL, J.; LANGRIDGE, P. & POWELL, W. Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. **Genetics**, v. 172, p. 557–567, 2006.

CASTRO, A. M. G. de; WRIGHT, J.; GOEDERT, W. Metodologia para viabilização do modelo de demanda na pesquisa agropecuária. In: **Anais do XIX Simpósio de Gestão da Inovação Tecnológica**. São Paulo: USP/PGT/FIA/PACTO, 1996.

CHANG, T. T. Manual on genetic conservation of rice germplasm for evaluation and utilization. 2. ed. Los Baños, Philippines: **International Rice Research Institute**, 81p., 1976.

CHAVES, M. M.; OLIVEIRA, M. M. Mechanisms underlying plant resilience to water deficits: prospects for water-saving agriculture. **Journal of Experimental Botany**, Oxford, v. 55, n. 407, p. 2365-2384, nov. 2004.

DARDICK, C.; CHEN, J.; RICHTER, T.; OUYANG, S.; RONALD, P. The rice kinase database. A phylogenomic database for the rice kinome. **Plant Physiology,** vol. 143, p. 579–586, 2007.

DAVEY, J. W. & BLAXTER, M. L. RADSeq: next-generation population genetics. **Brief. Funct. Genomics**, v. 9, p. 416–423, 2010.

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

DIEDHIOU, C. J.; POPOVA, O. V.; DIETZ, K. J.; GOLLDACK-BROCKHAUSEN, D. The SNF1-type serine-threonine protein kinase SAPK4 regulates stress-responsive gene expression in rice. **BMC PLANT BIOLOGY**, v. 8, 2008.

DONIS-KELLER, H.; GREEN, P.; HELMS, C.; CARTINHOUR, S.; WEIFFENBACH, B.; STEPHENS, K.; KEITH, T. P.; BOWDEN, D. W.; SMITH, D. R.; LANDER, E. S. A genetic linkage map of the human genome. **Cell**, v. 51, p. 319–337, 1987.

EARL, D. A.; VONHOLDT, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conservation Genetics Resources**, 2011.

EHRENREICH, I. M.; HANZAWA, Y.; CHOU, L.; ROE, J. L.; KOVER, P. X.; PURUGGANAN, M. D. Candidate gene association mapping of *Arabidopsis* flowering time. **Genetics**, Austin, v. 183, p. 325-335, 2009.

ELSHIRE, R. J.; GLAUBITZ, J. C.; SUN, Q.; POLAND, J. A.; KAWAMOTO, K.; BUCKLER, E. S.; MITCHELL, S. E. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. **PLoS ONE**, 2011.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular Ecology**, v. 14, p. 2611 – 2620, 2005.

FAO – **Food and Agriculture Foundation**. Disponível em <http:/www.fao.org>. Acesso em: 08 de janeiro de 2013.

FISCHER, R. A., MAURER, R. Drought resistance in spring wheat cultivars. I grain yield responses. **Australian Journal of Agricultural Research**, v. 29, n. 5, p. 897-912, 1978.

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

FRANKEL, O. H. Genetic perspectives of germplasm conservation. In: ARBER, W.; LLIMENSEE, K.; PEACOCK W. J.; STARLINGER, P. (Ed). **Genetic Manipulation: Impact on Man and Society.** Cambridge: Cambridge University Press, p. 161-170, 1984.

FUKAI, S.; COOPER, M. Development of drought-resistant cultivars using physio-morphological traits in rice. **Field Crops Research**, Elsevier, v. 40, n. 2, p. 67-86, fev. 1995.

GARRIS, A. J.; TAI, T. H.; COBURN, J.; KRESOVICH, S.; MCCOUCH, S. Genetic Structure and Diversity in *Oryza sativa* L. **Genetics**, Pittsburgh, v. 169, n. 3, p. 1631-1638, mar. 2005.

GAUT, B. S.; & ROSS-IBARRA, J. Selection on major components of angiosperm genomes. **Science**, v. 320, p. 484–486, 2008.

GE, S.; SANG, T.; LU, B. R.; HONG, D. Y. Phylogeny of rice genomes with emphasis on origins ofallotetraploid species. **Proceedings of the National Academy of Sciences (PNAS)**, Estados Unidos da América, v. 96, p. 14400-14405, 1999.

GOFF, S. A.; RICKE, D.; LAN, T. H.; PRESTING, G.; WANG, R.; DUNN, M.; GLAZEBROOK, J.; SESSIONS, A.; OELLER, P.; VARMA, H. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). **Science**, Washington, v. 296, n. 5565, p. 92-100, 2002.

GOICOECHEA, J. L., AMMIRAJU, J. S. S., MARRI, P. R., CHEN, M., JACKSON, S., YU, Y., ROUNSLEY, S., WING R. A. The Future of Rice Genomics: Sequencing the Collective *Oryza* Genome. **Rice**, v. 3, n. 2-3, p. 89-97, sep. 2010.

GOMES, A. da S.; MAGALHÃES JUNIOR, A. M. **Arroz irrigado no Sul do Brasil**. Brasília, DF: Embrapa Informação Tecnológica, 899p. 2004.

GONSÁLEZ-MARTÍNEZ, S. C.; WHEELER, N. C.; ERSOZ, E.; NELSON, C. D. NEALE, D. B. Association genetics in *Pinus taeda* L. I. wood property traits. **Genetics**, California, v. 175, p. 399–409, 2007.

GORANTLA, M.; BABU, P. R.; LACHAGARI, V. B. R.; REDDY, A. M. M.; WUSIRIKA, R.; BENNETZEN, J. L.; REDDY, A. R. Identification of stress-responsive genes in an *indica* rice (*Oryza sativa* L.) using ESTs generated from drought-stressed seedlings. **Journal of Experimental Botany**, Hyderabad, v. 58, p. 253-265, 2007.

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

HALL, D.; TEGSTRO, C.; INGVARSSON, P. K. Using association mapping to dissect the genetic basis of complex traits in plants. **Briefings in Functional Fenomics and Proteomics**, Oxford, v. 1, p. 1–9, 2010.

HARJES, C. E.; ROCHEFORD, T. R.; BAI, L.; BRUTNELL, T. P.; KANDIANIS, C. B.; SOWINSKI, S. G.; STAPLETON A. E.; VALLABHANENI, R.; WILLIAMS, M.; WURTZEL, E. T.; YAN, J.; BUCKLER, E. S. Natural genetic variation in *Lycopene epsilon* cyclase tapped for maize biofortification. **Science**, Leesburg, v. 319, p. 319–333, 2008.

HAYASHI, K.; HASHIMOTO, N.; DAIGEN, M.; ASHIKAWA, I. Development of PCR-based SNP markers for rice blast resistance genes at the Piz locus. **Theoretical and Applied Genetics**, New York, v. 108, n. 7, p. 1212–1220, 2004.

HEINEMANN, A. B. Caracterização dos padrões de estresse hídrico para a cultura do arroz (ciclo curto e médio) no estado de Goiás e suas consequências para o melhoramento genético. **Ciência e Agrotecnologia**, Lavras, v. 34, n. 1, p. 29-36, feb. 2010.

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

JARNE, P. & LAGODA, P. J. Microsatellites, from molecules to populations and back. **Trends in Ecology & Evolution**. v. 11, p. 424–429, 1996.

JENA, K. K.; MACKILL, D. J. Molecular markers and their use in marker-assisted selection in rice. **Crop Science**. Vol. 48, p. 1266-1276, 2008.

JONGDEE, B., FUKAI, S., COOPER, M. Leaf water potential and osmotic adjustment as physiological traits to improve drought tolerance in rice. **Field Crops Research**, v. 76, n. 2-3, p. 153-163, jul. 2002.

KAMOSHITA, A., BABU, R. C., BOOPATHI, N. M., FUKAI S. Phenotypic and genotypic analysis of drought-resistance traits for development of rice cultivars adapted to rainfed environments. **Field Crops Research**, v. 109, n. 1-3. p. 1-23, oct./dec. 2008.

KASUGA, M.; LIU, Q.; MIURA, S.; YAMAGUSCHI-SHINOZAKI, K.; SINOZAKI, K. Improving plant drought, salt, and freezing tolerance by gene transfer of a single stress-inducible transcription factor. **Nature America**, Yokohama, p. 287-291, 1999.

KASUGA, M.; MIURA, S.; SHINOZAKI, K.; YAMAGUCHI-SHINOZAKI, K. A combination of the *Arabidopsis* DREB1A gene and stress-inducible *rd29A* promoter improved drought and low temperature stress tolerance in tobacco by gene transfer. **Plant Cell Physiology**, Yokohama, v. 45, n. 3, p. 346-350, 2004.

KATO, S.; HARA, S. **On the affinity of rice varieties as shown by the fertility of rice plants.** Kyushu: Central Agricultural Institute of Kyushu Imperial University, 1928, 276p.

KHUSH, G. S. Origin, dispersal, cultivation and variation of rice. **Plant Molecular Biology**, Belgium: Zurich, v. 35, n. 1/2, p. 25-34, 1997.

KHUSH, G. S. What it will take to feed 5.0 billion rice consumers in 2030. **Plant Molecular Biology**, Belgium: Zurich, v. 59, n.. 1, p. 1-6, fev. 2005.

KUSH, G. S.; BRAR, D. S. **Biotechnology for rice breeding: progress and potential impact**. Thailand: FAO - The international rice commission, 2002. Disponível em: <http://www.fao.org/DOCREP/MEETING/004/AC347E/AC347E00.HTM>.

LAFITTE, H. R.; PRICE, A. H.; COURTOIS, B. Yield response to water deficit in an upland rice mapping population: associations among traits and genetic markers. **Theoretical and Applied Genetics**, Springer-Verlag, v. 109, n. 6, p. 1237-1246, jul. 2004.

LAFITTE, H R.; LI, Z. K.; VIJAYAKUMAR, C. H. M.; GAO, Y. M.; SHI, Y.; XU, J. L.; FU, B. Y.; YU, S. B.; ALI, A. J.; DOMINGO, J.; MAGHIRANG, R.; TORRES, R.; MACKILL, D. Improvement of rice drought tolerance through backcross breeding: evaluation of donors and selection in drought nurseries. **Field Crops Research**, Elsevier, v. 97, n. 1, p. 77-86, maio. 2006.

LIU, L.; LAFITTE, R.; GUAN, D. Wild *Oryza* species as potential sources of drought-adaptive traits. **Euphytica**, Netherlands, v. 138, n. 2, p. 149-161, jun. 2004.

LOCKTON, S. & GAUT, B. S. Plant conserved non-coding sequences and paralogue evolution. **Trends in Genetics.** v. l. 21, p. 60–65, 2005.

LU, H.; RATE, D. N.; SONG, J. T.; GREENBERG, J. T. ACD6, a novel ankyrin protein, is a regulator and an effector of salicylic acid signaling in the *Arabidopsis* defense response. **Plant Cell**, v.15, p. 2408–2420, 2003.

MALOSETTI, M.; VAN DER LINDEN, C. G.; VOSMAN, B.; VAN EEUWIJK, F. A. A Mixed-Model Approach to Association Mapping Using Pedigree Information With an Illustration of Resistance to Phytophthora infestans in Potato. **Genetics**, Wageningen, v. 175, p. 879–889, 2007.

MAO, X.; ZHANG, H.; TIAN, S.; CHANG, X.; JING, R. TaSnRK2.4, an SNF1-type serine/threonine protein kinase of wheat (*Triticum aestivum* L.), confers enhanced multistress tolerance in Arabidopsis. **Journal of Experimental Botany.** v. 61, n. 3, p. 683-696, 2010.

METZKER, M. L. Sequencing technologies — the next generation. **Nature Reviews Genetics.** v. 11, p. 31–46, 2010.

MORISHIMA, H.; MARTINS, P.S. **Investigations of plant genetic resources in Amazon Basin with emphasis on the genus *Oryza***. Mishima, Japan: The Monbusho International Scientific Research Program, 100p. 1994.

MOSAVI, L. K.; CAMMETT, T. J.; DESROSIERS, D. C.; PENG, Z. Y. The ankyrin repeat as molecular architecture for protein recognition. **Protein Science.** v. 13, n. 6, p. 1435-1448, 2004.

NASS, L. L. **Recursos Genéticos Vegetais**. Brasília – DF: Embrapa Recursos Genéticos e Biotecnologia, 858 p. 2007.

NGUYEN, H. T.; BABU, R. C.; BLUM, A. Breeding for drought resistance in rice: physiology and molecular genetics considerations. **Crop Science**, Madison, v. 37, p. 1426-1437, set. 1997.

NI, J.; COLOWIT, P. M.; MacKILL, D. Evaluation of genetic Diversity in rice using microsatellite markers. **Crop Science**, Madison, v. 42, n. 2, p. 601-607, 2002.

NORDBORG, M. & WEIGEL, D. Next-generation genetics in plants. **Nature**, v. 456, p. 720–723, 2008.

OKUSHIMA, Y.; OVERVOORDE, P. J.; ARIMA, K.; ALONSO, J. M.; CHAN, A., CHANG, C.; ECKER, J. R.; HUGHES, B.; LUI, A., NGUYEN, D. ET AL. Functional genomic analysis of the AUXIN RESPONSE FACTOR gene family members in Arabidopsis thaliana: Unique and overlapping functions of ARF7 and ARF19. **Plant Cell**, v. 17, p. 444 -463, 2005.

ORAGUZIE, N. C.; RIKKERINK, E. H. A.; GARDINER, S. E.; SILVA, H. N. **Associaion Mapping in Plants**. Hardcover, 277p. 2007.

PANTUWAN, G., FUKAI, S., COOPER, M., RAJATESEREEKUL, S., O'TOOLE, J. C. Yield response of rice (Oryza sativa L.) genotypes to drought under rainfed lowlands. Part 2. Selection of drought resistant genotypes. **Field Crops Research**, v. 73, p. 169-180, 2002.

PEREIRA, J. **Alterações na qualidade tecnológica de grãos de arroz (*Oryza sativa* L.) durante o armazenamento**. 1996. 107 f. Dissertação (Mestrado em Ciência e Tecnologia de Alimentos) – Universidade Federal de Viçosa, Viçosa, 1996.

POOL, J. E.; HELLMANN, I.; JENSEN, J. D. & NIELSEN, R. Population genetic inference from genomic sequence variation. **Genome Research**. v. 20, p. 291–300, 2010.

POROYKO, V.; SPOLLEN, W. G.; HEJLEK, L. G.; HERNANDEZ, A. G.; LENOBLE, M. E.; DAVIS, G.; NGUYEN, H. T.; SPRINGER, G. K.; SHARP, R. E.; BOHNERT, H. J. Comparing regional transcript profiles from maize primary roots under well-watered and low water potential conditions. **Journal of Experimental Botany**, Urbana, v. 58, p. 279-289, 2007.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, Oxford, v. 155, n. 2, p. 945-959, 2000.

RAO, N. K. Plant genetic resources: advancing conservation and use through biotechnology. **African Journal of Biotechnology**, Nairobi, v. 3, n. 2, p. 136-145, 2004.

RISCH, N.; MERIKANGAS, K. The future of genetic studies of complex human diseases. **Science**, Stanford, v. 273, p. 1516–1517, 1996.

SAS INSTITUTE. **SAS language and procedures: usage**. Version 6. Cary, 373 p. 2005.

SCHEET, P. AND STEPHENS, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. **American Journal of Human Genetics.** 2006.

SECOND, G. Origin of the genic diversity of cultivated rice *(Oryza* spp*):* study of the polymorphism scored at 40 isozyme loci. **Japanese journal of genetics**, Shizuoka, v. 57, n. 1, p. 25-57, 1982.

SHINOZAKI, K; YAMAGUCHI-SHINOZAKI, K. Molecular Responses to Dehydration and low temperature: differences and cross-talk between two stress signaling pathways. **Current Opinion in Plant Biology**, Tsukuba, v. 3, p. 217-223, 2000.

SHINOZAKI, K.; YAMAGUCHI-SHINOZAKI, K. Gene networks involved in drought stress response and tolerance. **Journal of Experimental Botany**, Tsurumi-ku, v. 58, n. 2, p. 221-227, 2007.

SOUZA, G. S.; WANDER, A. L.; GAZZOLA, R.; SOUZA, R. S. Evolução da produção e do comércio internacional do arroz e projeção de preços. **Pesquisa Operacional para o Desenvolvimento**, Rio de Janeiro - RJ, v. 2, n. 1, p. 1-86, 2010.

STAPLEY, J.; REGER, J.; FEULNER, P. G. D.; SMADJA, C.; GALINDO, J.; EKBLOM, R.; BENNISON, C.; BALL, A. D.; BECKERMAN, A. P.; SLATE, J. Adaptation genomics: the next generation. **Trends in Ecology & Evolution.** v. 25, p. 705–712, 2010.

STOGIOS, P. J.; DOWNS, G. S.; JAUHAL, J. J.; NANDRA, S. K.; PRIVÉ**,** G. G. Sequence and structural analysis of BTB domain proteins. **Genome Biology**. v. 6, 2005.

STOREY, J. D. A direct approach to false discovery rates. **Journal of the Royal Statistical Society**, Stanford, v. 64, p. 479-498. 2002.

SWEENEY, M.; MCCOUCH, S. The complex history of the domestication of Rice. **Annals of Botany**, New York, v. 100, p. 951-957, 2007.

TENAILLON, M. I.; HOLLISTER, J. D. & GAUT, B. S. A triptych of the evolution of plant transposable elements. **Trends in Plant Science.** v. 15, p. 471–478, 2010.

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

TYAGI, A.;.KHURANA, J. P.; KHURANA, P.; RAGHUVANSHI, S.; GAUR, A.; KAPUR, A.; GUPTA, V.; KUMAR, D.; RAVI, V.; VIJ, S.; KHURANA, P.; SHARMA, S. Structural and functional analysis of rice genome. **Journal of Genetics**, Bangalore, v. 83, n. 1, p. 79-99. 2004.

UPADHYAYA, H. D.; ORTIZ, R. A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. **Theoretical and Applied Genetics**, Berlin, v. 102, n. 8, p. 1292-1298, 2001.

VAUGHAN, D. A. **The genus *Oryza* L.:** current status of taxonomy. Manila: International Rice Research Institute. 21p. 1989.

VAUGHAN, D. A. **The Wild Relatives of Rice**: A Genetic Resources Handbook. Manila, Philippines: International Rice Research Institute. 137p. 1994.

VAUGHAN, D. A.; MORISHIMA, H.; KADOWAKI, K. Diversity in the *Oryza* genus. **Current Opinion in Plant Biology**, Elsevier, v. 6, n. 2, p. 139-146, abr. 2003.

VAUGHAN, D. A.; KADOWAKI, K.; KAGA, A.; TOMOOKA, N. On the phylogeny and biogeography of the genus *Oryza*. **Breeding Science**, Tsukuba, v. 55, n. 2, p. 113-122, fev. 2005.

VIEIRA, J. **Caracterização morfológica e molecular do banco de germoplasma de arroz irrigado (*Oryza sativa* L.) da EPAGRI**. 2007. 115 f. Dissertação (Mestrado em Recursos Genéticos Vegetais) – Centro de Ciências Agrárias, Universidade Federal de Santa Catarina, 2007.

VIKRAM, P., SWAMY B. P. M., DIXIT S., AHMED H., STA CRUZ M. T., SINGH A. K., YE G., KUMAR A. Bulk segregant analysis: An effective approach for mapping consistent-effect drought grain yield QTL in rice. **Field Crops Research**, v. 134, p. 185-192, aug. 2012.

VOS, P.; HOGERS, R.; BLEEKER, M.; REIJANS, M.; VAN DE LEE, T.; HORNES, M.; FRIJTERS, A.; POT, J.; PELEMAN, J.; KUIPER, M.; ZABEAU, M. AFLP: a new technique for DNA fingerprinting. **Nucleic Acids Research**, London, v. 23, n. 21, p. 4407–4414, 1995.

WATANABE, Y. Phylogeny ang geographical distribution of genus *Oryza*. In: MATSUO, T.; FUTSUHARARA, Y.; KIKUCHI, F.; YAMAGUCHI, H. **Science of the rice plant genetics**. Tokyo: Food and Agriculture Plicy Research Center. p.29-39. 1997.

WEN, W.; MEI, H.; FENG, F.; YU, S.; HUANG, Z.; WU, J.; CHEN, L.; XU, X.; LUO, L. Population structure and association mapping on chromosome 7 using a diverse panel of chines germplasm of rice. **Theoretical and Applied Genetics**, Wuhan, v. 119, p. 459–470, 2009.

XU, Y. B.; BEACHELL, H.; McCOUCH, S. R. A marker-based approach to broadening the genetic base of rice in the USA. **Crop Science**, Madison, v. 44, n. 6, p. 1947-1959, 2004.

YE, C., ARGAYOSO, M. A., REDOÑA, E. D., SIERRA, S. N., LAZA, M. A., DILLA, C. J., MO, Y., THOMSON, M. J., CHIN, J., DELAVIÑA, C. B., DIAZ, G. Q., HERNANDEZ, J. E. Mapping QTL for heat tolerance at flowering stage in rice using SNP markers. **Plant Breeding**, v. 131, n. 1, p. 33-41, feb. 2012.

YU, J.; BUCKLER, E. S. Genetic Association mapping and genome organization of maize. **Current Opinion in Biotechnology**, New York, v. 17, p. 155–160, 2006.

YU, J.; HOLLAND, J. B.; McMULLEN, M. D.; BUCKLER, E. S. Genetic design and statistical power of nested association mapping in maize. **Genetics**, New York, v. 178, p. 539-551. 2008.

ZHANG, J.; ZHENG, H. G.; AARTI, A.; PANTUWAN, G.; NGUYEN, T. T.; TRIPATHY J. N.; SARIAL, A. K.; ROBIN, S.; BABU, R. C.; NGUYEN, B. D.; SARKARUNG, S.; BLUM, A.; NGUYEN, H. T. Locating genomic regions associated with components of drought resistance in rice: comparative mapping within and across species. **Theoretical and Applied Genetics**, Springer: Berlim, v. 103, n. 1, p. 19-29, 2001.

ZHANG, N.; XU, Y.; AKASH, M.; MCCOUCH, S.; OARD, J. H. Identification of candidate markers associated with agronomic traits in rice using discriminant analysis. **Theoretical and Applied Genetics**, Baton Rouge, v. 110, p. 721–729, 2005.

ZHANG, P., LI, J., LI, X., LIU, X., ZHAO, X., LU, Y. Population Structure and Genetic Diversity in a Rice Core Collection (*Oryza sativa* L.) Investigated with SSR Markers. P**los One**, v. 6, n.12, dec. 2011.

ZHU, J. K. Salt and drought stress signal transduction in plants. **Annual Review of Plant Biology**, Tucson, v. 53, n. 1, p. 247-273, jun. 2002.

ZHU, C.; GORE, M.; BUCKLER, E. S.; YU, J. Status and Prospects of Association Mapping in Plants. **The Plant Genome**, Manhattan, v. 1, n. 1, p. 5–20, 2008.