RENATA PITA BOMFIM

ANÁLISE DE ASSOCIAÇÃO GENÔMICA PARA TOLERÂNCIA A SECA EM MILHO

LAVRAS-MG

2020

RENATA PITA BOMFIM

ANÁLISE DE ASSOCIAÇÃO GENÔMICA PARA TOLERÂNCIA A SECA EM MILHO

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-Graduação do Mestrado Profissional em Genética e Melhoramento de Plantas, área de concentração em Genética Quantitativa, para obtenção do título de Mestre

Prof. Dr. Welison Andrade Pereira

Orientador

Dr. Ivan Schuster

Coorientador

LAVRAS-MG

2020

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Bomfim, Renata Pita.
Análise de Associação Genômica para Tolerância a Seca em Milho / Renata Pita Bomfim. - 2020.
42 p.

Orientador(a): Welison Andrade Pereira.
Coorientador(a): Ivan Schuster.
Dissertação (Mestrado Profissional) - Universidade Federal de Lavras, 2020.
Bibliografia.

1. Milho. 2. Estresse Hídrico. 3. Mapeamento por Associação. I. Pereira, Welison Andrade. II. Schuster, Ivan. III. Título.

RENATA PITA BOMFIM

ANÁLISE DEASSOCIAÇÃO GENÔMICA PARA TOLERÂNCIA A SECA EM MILHO

GENOMIC ASSOCIATION ANALYSIS FOR DROUGHT TOLERANCE IN MAIZE

Dissertação apresentada à Universidade Federal de Lavras como parte das exigências do Programa de Pós-Graduação do Mestrado Profissional em Genética e Melhoramento de Plantas, área de concentração em Genética Quantitativa, para obtenção do título de Mestre

APROVADA em 26 de junho de 2020.

Dr. Welison Andrade Pereira UFLA

Dr. Ivan Schuster LPHT

Dr. Evandro Novaes UFLA

Prof. Dr. Welison Andrade Pereira

Orientador

Dr. Ivan Schuster

Coorientador

LAVRAS-MG

2020

RESUMO

O estresse hídrico é um dos fatores que mais limita o desenvolvimento e a produtividade das plantas. A identificação e o entendimento dos componentes genéticos associados à tolerância a seca são fundamentais para o desenvolvimento de novas variedades e híbridos tolerantes a este estresse. No presente estudo foram avaliadas 85 linhagens de milho em dois locais. Em cada local, as linhagens foram avaliadas com e sem estresse hídrico. As linhagens foram genotipadas com 8289 marcadores SNPs (*Single Nucleotide Polymorphism*), e os dados foram usados para análise de associação genômica entre marcadores e QTLs (*Quantitative Trait Loci*) de tolerância ao estresse hídrico. Foram identificados 10 QTLs em sete cromossomos. Dois QTLs estão localizados em regiões contendo os genes: *atg18f e ys3*. Oito QTLs estão em regiões genômicas contendo Modelos de Genes, sendo que a expressão de cinco deles estão descritas como relacionada a estresses bióticos e/ou abióticos. Os QTLs identificados neste trabalho têm potencial para serem usados em programas de Seleção Assistida por Marcadores Moleculares para tolerância a estresse hídrico, após validados em outros conjuntos de germoplasma.

Palavras-chave: Milho. Estresse hídrico. QTLs. Mapeamento por Associação. SNPs. Genotipagem por Sequênciamento.

ABSTRACT

Drought is one of the most limiting factors for the plant development and crop yield. The identification and understanding of the genetic components associated to drought tolerance are essential to develop new varieties and hybrids tolerant to such stress. In the present study 85 maize inbred lines were evaluated in two locations, under irrigated and water stress conditions. For Genomic Association Analysis between markers and QTLs related to drought tolerance, the lines were genotyped using 8,289 SNP markers. Ten QTLs were found on seven chromosomes related to the effect of drought in corn development. Two QTLs are located in regions including the *atg18f and ys3* genes. Eight QTLs are located in genomic regions comprising genetic models, five of them are described to related to biotic and abiotic stresses. The QTLs identified in this work have potential to be used in Marker-Assisted Selection programs for drought tolerance after validation across other sets of germplasm.

Keywords: Corn. Hydrical stress. QTLs. Association Mapping. SNPs. Genotyping by Sequencing.

SUMÁRIO

1 INTRODUÇÃO

O estresse pela baixa disponibilidade hídrica é um dos fatores que mais limitam o desenvolvimento do milho, tendo um impacto significativo na redução da produtividade (Araus et al., 2012; Nepolean et al., 2014). Em 2012, de acordo com um levantamento realizado pela FAOSTAT, estimou-se que de 15 a 20% da produtividade de milho são perdidos todos os anos em decorrência do estresse hídrico, e estas perdas podem aumentar em consequência das mudanças climáticas que vem ocorrendo (FAOSTAT., 2012).

O estresse hídrico afeta negativamente a cultura do milho durante todos os estágios de desenvolvimento, sendo que o estágio reprodutivo, particularmente entre a emergência do pendão e o enchimento de grãos, é o período de maior sensibilidade (Grant et al., 1989; Agrama e Moussa., 1996).

A tolerância do milho ao estresse hídrico pode reduzir as perdas em situações de seca, porém, a seleção de genótipos e o melhoramento genético em condições de estresse hídrico são atividades bastante complexas. A seleção de germoplasma tolerante ao estresse hídrico pode ser facilitada se for aplicada em características secundárias, cuja variância genética e herdabilidade são maiores, aumentando, neste contexto, a eficiência de seleção (Wang et al., 2016; Banziger e Laffite., 1997; Bolanõs e Edmeades., 1996; Ludlow and Muchow., 1990).

Diversos estudos têm demonstrado a eficiência do uso de características secundárias na seleção para tolerância a seca e sua correlação com a produtividade de grãos (Edmeades et al., 1993; Banzinger et al., 2000; Ziyomo et al., 2013). Como caracteres secundários, a altura de planta e espiga, peso de grãos, senescência foliar e stay green são exemplos de fácil medição e com alta herdabilidade, aumentando, assim, a resposta de seleção em condições de estresse hídrico (Nikolic et al., 2011; Nepolean et al., 2013; Xue et al., 2013; Sheikh et al., 2017).

Assim como utilizar caracteres secundários, outra alternativa que pode facilitar a seleção de germoplasma tolerante ao estresse hídrico é a seleção assistida por marcadores moleculares (SAM) associada ao rendimento de grãos e caracteres correlacionados, sob estresse (Cattivelli et al., 2008; Messmer et al., 2009; Wang et al., 2016; Shikha et al., 2017; Nepolean et al., 2018). Diversos estudos têm identificado QTLs associados a caracteres secundários que são afetados pelo estresse hídrico (Rahman et al., 2011; Sabadin et al., 2012; Almeida et al., 2013; Almeida et al., 2014; Zhang et al., 2016; Zhao et al., 2018). No entanto, estes estudos têm realizado mapeamento de QTLs em condições de estresse, mas não tem avaliado o impacto diferenciado do estresse sobre os genótipos.

Este trabalho teve como objetivo identificar QTLs associados à resposta diferencial de linhagens de milho ao estresse hídrico, avaliado pela diferença de desempenho em ambientes de estresse hídrico comparado com o desempenho em ambiente irrigado. Para isso, foram avaliados caracteres secundários relacionados com a tolerância ao estresse hídrico em germoplasma de milho adaptado ao cultivo no Brasil. As diferenças observadas nos caracteres secundários foram utilizadas para análise de associação genômica, e identificação QTLs associados a resposta diferencial do milho ao estresse hídrico.

2 MATERIAIS E MÉTODOS

2.1 Ambientes de avaliação e Delinemaneto Experimental

O trabalho foi conduzido no ano de 2019, com uma população de 85 linhagens de milho do banco de germoplasma da LongPing HighTech, em dois locais: Jardinópolis, em SP (Latitude S 20° 54’ 32.7” Longitude W 47° 53’ 26.2”), e Sorriso, em MT (Latitude S 12° 26’ 37.8” Longitude W 55° 49’ 37.7”).

Foi empregado o delineamento experimental em faixas, em que as faixas foram os regimes de irrigação, e em cada faixa os tratamentos foram distribuídos em blocos casualizados com três repetições. Foram utilizados dois regimes de irrigação. Para facilitar o manejo da irrigação, foi aplicada restrição no sorteio dentro de cada bloco, para que as linhagens pudessem ficar agrupadas por ciclo, dentro de cada bloco. Dentro dos blocos, as linhagens foram agrupadas em três grupos: Hiperprecoce, SuperPrecoce e Precoce. As parcelas foram constituídas por quatro linhas de 4,0 m, com espaçamento de 0,5 m entre linhas e 0,375 m entre plantas. A população inicial foi de 60 plantas por parcela, o que corresponde a uma densidade de 75.000 plantas ha^{-1}.

O preparo do solo, semeadura e tratos culturais seguiram as recomendações técnicas para a cultura do milho.

2.2 Manejo da Irrigação

Os experimentos foram irrigados pelo método de aspersão com barras. Por este método, uma lâmina de 10 mm de água foi aplicada semanalmente, de maneira a reproduzir o manejo de irrigação realizado pelos agricultores.

Nos tratamentos sob estresse hídrico, a irrigação foi suspensa na fase de pré-floração, e retomada após o florescimento. Devido a diferença de ciclo entre as linhagens, nas parcelas com linhagens Hiperprecoces, a irrigação foi interrompida aos 56 dias após o plantio (dap) e retomadas aos 89 dap; Para linhagens SuperPrecoces, a irrigação foi interrompida aos 59 dap e retomada aos 92 dap; e para as linhagens Precoces, a irrigação foi interrompida aos 60 dap e retomada aos 92 dap. No ensaio irrigado, grupo controle deste experimento, as irrigações foram conduzidas normalmente, até o término da fase reprodutiva.

2.3 Caracteres Avaliados

Nos dois locais de ensaio, foram avaliadas as seguintes características: Atura de Planta (AltPlant) em cm, Altura de Espiga (AltEsp) em cm, Peso de 100 Grãos (P100) em g, Peso de Grãos por Parcela (PGP) em Kg/ha e Umidade dos Grãos na colheita (UG) em %. No ensaio de Sorriso-MT, foi também avaliada Prolificidade (PROL) em número de espigas por planta. Os caracteres foram avaliados seguindo-se as recomendações do Manual de Fenotipagem para Tolerância de Estresse Abiótico em milho – Estresse Hídrico (Allah et al., 2016).

2.4 Dados Moleculares

A extração de DNA foi realizada a partir de quatro discos de tecido foliar, obtidos de quatro plantas no estágio V3, utilizando-se o Kit Fast ID Genomic DNA Extraction® (GENETIC ID., Fairfield, IA, EUA) de acordo com as instruções do fabricante. A genotipagem por sequenciamento foi realizada utilizando-se da plataforma Ion GeneStudio S5 Prime (Thermo Fisher Scientific., Waltham, MA, EUA), de acordo com as instruções do fabricante. Para genotipagem, foi utilizado um painel customizado de 8.289 marcadores SNPs. Após a genotipagem, marcadores com *Call Rate* (dados válidos) menor do que 90%, e MAF (Frequência do Alelo Menos Frequente) menor do que 5% forma excluídos antes das análises subsequentes.

2.5 Análise de Dados

Os dados fenotípicos avaliados em cada experimento foram analisados utilizando-se o método BLUP (*Best Linear Unbiased Prediction* - Melhor Predição Linear Não-viesada), procedimento PROC MIXED, do Software Selegen (Resende., 2016), para estimação dos valores genotípicos de cada linhagem, para cada característica. Foi utilizado o modelo:

$$y = X\beta + Zg + Wp + \varepsilon$$

Em que, y é o vetor de resposta observado para característica avaliada. X, Z e W são matrizes de incidência conhecidas, relacionando y a β, g e p, respectivamente. β é o vetor dos efeitos ambientais fixos para ambiente e bloco, g e p são vetores dos genótipos (aleatórios) e interação Genótipo (G) x Ambiente (E), respectivamente, e ε é o efeito residual. O modelo obtém os valores de BLUPs para os genótipos (efeitos aleatórios), resolvendo as equações do modelo misto de Henderson (1984).

Após obtidos os valores genotípicos preditos, foi estimada a diferença (Δ) entre os valores de cada característica quando avaliados sob estresse hídrico e sob tratamento controle,

em cada local. As características utilizadas para as análises posteriores foram ΔAltPlant, ΔAltEsp, ΔP100, ΔPGP, ΔUG, e ΔPROL. Os valores das diferenças (Δ) de cada característica foram submetidos a análise de variância utilizando-se o programa GENES (Cruz., 2016).

Os dados das diferenças entre os valores genotípicos para cada característica foram usados também como dados fenotípicos para a análise de Associação Genômica. Para poder incluir os efeitos da estrutura de populações e de parentesco genético entre as linhagens na análise de associação entre marcadores e QTLs, foi utilizado um modelo linear misto (MLM) para a análise de associação genômica. A matriz de estrutura de população (Q) foi estimada pelo agrupamento obtido pelo método Bayesiano generalizado, utilizando-se o modelo de máxima verossimilhança e o método de relaxamento em bloco, para acelerar a convergência, utilizando-se o software Admixture (Alexander et al., 2009).

A matriz de coeficientes de parentesco (Kinship - K), que explica a maior probabilidade de dois alelos serem idênticos em estado entre as linhagens, foi estimada pelo programa TASSEL (Bradbury et al., 2007). A análise de associação utilizando Modelos Lineares Mistos (MLM), considerando as matrizes Q e K como cofatores, também foi realizada no programa TASSEL. Os marcadores foram identificados como significativamente associados às características avaliadas quando $p < 0.001$. Os resultados de -log10(P) foram plotados em gráficos *Manhattan Plot* para visualização dos resultados.

Marcadores significativos a 0,1% de probabilidade na análise de MLM foram submetidos à análise de regressão linear múltipla com o método *Stepwise* de seleção do modelo. Esta análise tem por objetivo evitar marcadores redundantes no modelo, ou seja, marcadores ligados ao mesmo QTL, e excluir marcadores que contribuem pouco para a expressão da característica que está sendo avaliada. Para as análises de regressão múltipla foi utilizado o programa JMP (SAS Institute., 1990), com probabilidade de entrada e de saída de 5%, e procedimento Stepwise de seleção do modelo.

Uma análise de Desequilíbrio de Ligação foi realizada nos marcadores significativos obtidos a partir da análise de Regressão Múltipla utilizando-se o programa Haploview.

2.6 Associação com Genes ou Modelos de Genes Conhecidos

Os genes ou Modelos de Genes localizados nas regiões dos QTLs foram obtidos no *National Center for Biotechnology Information* (NCBI), utilizando-se a posição física nos cromossomos de cada marcador significativamente associado às diferenças observadas nas

características avaliadas sob estresse e sob irrigação. O genoma referência utilizado foi *Zea mays (assembly B73 RefGen_v4).*

3 RESULTADOS E DISCUSSÃO

3.1 Análise de Associação Genômica

Dos 8289 marcadores SNP utilizados, 1754 foram excluídos devido ao *call rate* inferior a 90%, e 509 foram excluídos por terem MAF menor do que 5%. Com a remoção destes marcadores, 6026 foram utilizados nas análises de associação genômica.

Apenas a característica ΔP100 em Sorriso não foi significativa na análise de variância (Tabela 1). Para todas as demais características, a diferença observada nos valores genotípicos das avaliações com e sem estresse hídrico foram significativas nos dois locais avaliados. Isso demonstra a variabilidade na população avaliada, o que é condição essencial para estudos de associação genômica.

Redução na altura de plantas, na altura de espiga, no peso de 100 grão e na produtividade são frequentemente relatadas como efeitos do estresse hídrico em milho (Almeida et al., 2014; Banziger et al., 2000; Ge et al., 2012; Zhao et al., 2018; Zhao et al., 2019).

Os valores das diferenças observadas nos ambientes com e sem estresse hídrico foram utilizados na análise de associação genômica com os marcadores moleculares. Em Jardinópolis, foi identificado na análise de MLM, um marcador significativamente associado à diferença na altura de espiga (ΔAltEsp), no cromossomo 1. Para a diferença no peso de grãos por parcela (ΔPGP), foram identificados dez marcadores nos cromossomos 1,2,3,5,6,7,8 e 9. Para a diferença na umidade de grãos (ΔUG), foram identificados quatro marcadores nos cromossomos 1,2,3 e 8. Em Sorriso, foi identificado um marcador significativamente associado à diferença na altura de planta (ΔAltPlant) e quatro marcadores associados nos cromossomos 1,3 e 4 para a diferença no peso de grãos por parcela (ΔPGP) (Figura 1).

Os marcadores significativos na análise de associação por modelos lineares mistos foram utilizados em uma análise de regressão múltipla. Uma vez que a análise de regressão múltipla com método Stepwise de seleção de modelo mantém apenas marcadores não redundantes no modelo final, cada marcador significativo nesta análise corresponde a um QTL. Três dos 10 marcadores associados a ΔPGP na análise de MLM em Jardinópolis, foram significativos na análise de regressão múltipla, localizados nos cromossomos 3, 6 e 7 (Tabela

2). Dos 4 marcadores associados a ΔUG em Jardinópolis, três foram significativos na análise de regressão múltipla, nos cromossomos 2, 3 e 8. Para ΔAltEsp apenas um marcador foi significativo na análise de associação. Esta característica não foi avaliada por regressão simples, para efeito de comparação dos coeficientes de determinação. Os marcadores significativos na análise de regressão múltipla explicaram 50,63, 43,51 e 9,23 % da variação nos valores genotípicos de ΔPGP, ΔUG e ΔAltEsp, respectivamente.

Em Sorriso, duas das cinco características avaliadas foram significativas na análise de regressão múltipla (Tabela 2). Os SNPs associados foram encontrados nos cromossomos 1 e 4 para ΔPGP e 4 para ΔAltPlant. Os marcadores significativos na análise de regressão múltipla explicaram 35,18% e 9,78% das variações em ΔPGP e ΔAltPlant, respectivamente.

Outros estudos de associação para tolerância a estresse hídrico em milho identificaram quantidades muito maiores de QTLs. Almeida et al. (2014) identificaram 203 QTLs associados aos caracteres, número de espigas por planta, *stay green* e altura de planta em regimes hídricos com e sem estresse.

Zhao et al. (2018) estudaram QTLs e Meta-QTLs para sete caracteres agronômicos, em múltiplas populações de milho sob condições de estresse e não estresse hídrico. Sessenta e nove regiões genômicas envolvidas na expressão fenotípica para altura de plantas e espiga, IFMF, peso de espiga e sabugo, peso de 100 grãos e comprimento de espigas foram identificadas. Estes QTLs explicaram de 4 a 17 % da variação fenotípica para o ambiente irrigado. Aproximadamente 52 dos 69 QTLs foram identificados em condições de estresse hídrico.

Todos estes trabalhos avaliaram a presença de QTLs associados às características secundárias relacionadas ao estresse hídrico em milho, em ambientes com estresse e sem estresse. No presente trabalho, foram consideradas as diferenças observadas entre os valores das características em condições de ausência ou presença de estresse hídrico, ou seja, a resposta das plantas ao estresse. Os QTLs identificados no presente trabalho, portanto, não estão associados às características avaliadas, e sim, à tolerância das linhagens de milho ao estresse hídrico. Sendo assim, independente da característica que foi avaliada, todos os QTLs identificados no presente trabalho estão relacionados a resposta do milho ao estresse hídrico.

Após a análise de Regressão Múltipla, foram identificadas sete regiões relevantes, em seis cromossomos, no ambiente de Jardinópolis, e três regiões, em dois cromossomos ambiente de Sorriso, totalizando 10 QTLs associados a tolerância do milho ao estresse hídrico. Em alguns cromossomos foram identificados mais de um QTL. Em Jardinópolis, foram identificados dois QTLs no cromossomo 3. Em Sorriso foram identificadas dois QTLs no cromossomo 4. Além disso, no cromossomo 1 foi identificada QTL em Jardinópolis e um em Sorriso.

Nenhuma das regiões genômicas contendo QTLs no mesmo cromossomo está em desequilíbrio de ligação, podendo-se considerar que se tratam de QTLs diferentes no mesmo cromossomo. Os 10 QTLs identificados estão localizados em sete cromossomos do milho.

As características que permitiram identificar o maior número de QTLs associados ao estresse hídrico em ambos locais experimentais foram ΔPGP e ΔUG, relacionadas a produtividade e maturidade respectivamente, indicando serem características com maior associação a resposta ao estresse hídrico nos genótipos avaliados neste trabalho.

3.2 Genes Candidatos

Dos 10 QTLs identificados para estresse hídrico neste trabalho, oito estão localizados em regiões já caracterizadas no genoma do milho. Dois estão em regiões que contém genes já descritos: *atg18f* e *ys3*. (Tabela 3).

O gene *atg18f* está localizado no cromossomo 3, na posição 192795015 até 192801094, associado ao marcador 10K13323. Este gene pertence à família de genes ATG (*AuTophaGy-related genes*), que fazem parte da coordenação de um processo celular altamente conservado, a autofagia (Su et al., 2020). O processo de degradação e reciclagem dos componentes celulares, pelos vacúolos e/ou lisossomos, através da autofagia, permite a manutenção da homeostase celular sob condições normais e a regulação do desenvolvimento celular em períodos de estresse (Su et al., 2020; Tejeda et al., 2020). Diversos estudos vêm demostrando o importante papel do processo da autofagia na adaptação de culturas como trigo, pimenta, tomate e milho a estresses ambientais, como deficiência de carbono e nitrogênio, estresse hídrico, térmico e osmótico (Zhai et al., 2016; Singorelli et al., 2019; Zhang et al., 2019; Tejeda et al., 2020).

Um estudo conduzido por Liu et al. (2009) demonstraram que o silenciamento do gene *AtATG18a*, em plantas de *Arabidopsis thaliana*, resulta em uma alta sensibilidade quando submetidas a condições de estresse hídrico e osmótico.

O gene *ys3* está localizado no cromossomo 3, na posição 97975251 até 97982164 associado ao marcador 10K03498, não foi ainda descrito na literatura como um gene relacionado à tolerância ao estresse hídrico em plantas. Este gene está descrito como responsável por conferir a coloração amarela entre as nervuras das plantas. Esta é a primeira vez que este gene está sendo associado ao estresse hídrico, e o estudo de seu papel nos mecanismos de resposta ao estresse hídrico pode ser mais bem avaliado em estudos futuros.

A região genômica localizada no cromossomo 6, na posição 133574533 até 133578647 associada ao marcador 10K09522, está próxima a um Modelo de Gene identificado como *LRR receptor-like serine/threonine-protein kinase,* comumente associado à tolerância das plantas a doenças.

A percepção e condução de sinais através de receptores localizados na superfície celular é um mecanismo comum entre os organismos vivos. Em plantas, a transdução de muitos destes sinais celulares é mediada por RLKs *(Receptor-Like Kinases).* RLK-RRL (*Leucine-rich repeat receptor-like protein kinase)* compreendem a maior subfamília de receptores RLKs encontrados em plantas, desempenhando um importante papel em processos celulares relacionados ao crescimento, desenvolvimento e resposta aos estresses ambientais (Shiu e Bleecker., 2001; Liu et al., 2017; Wei e Li., 2019).

Ouyang et al. (2010), estudaram a expressão de um gene putativo RLK com repetições ricas em leucina (*OsSIK1*) em arroz e sua correlação à melhora na tolerância ao estresse hídrico e de salinidade. Os autores identificaram que plantas transgênicas para alta expressão de *OsSIK1* se mostraram mais tolerantes aos estresses hídrico e salinidade do que plantas controle.

A região genômica localizada no cromossomo 8, na posição 67033198 até 67034788 associada ao marcador 10K07174, está próxima a um Modelo de Gene descrito como *indole-3-acetate beta-glucosyltransferase.* A *indole-3-acetate beta-glucosyltransferase (UDPG),* é uma enzima catalizadora envolvida na rota metabólica da auxina (AIA: ácido indolacético) nos vegetais. O processo de glicosilação da auxina é um dos mecanismos que contribuem para a homeostase hormonal, sendo a enzima UDPG, uma enzima catalizadora de uma reação química reversível que caracteriza o primeiro passo para a biossíntese de conjugados AIA-éster, (conjugação do fitormônio + açúcares) em plantas monocotiledôneas (Ostrowski et al., 2015). A auxina é um hormônio com papel importante na regulação de muitos processos fisiológicos das plantas (Woodward e Bartel., 2005). Ainda não existem na literatura, trabalhos que associem genes responsáveis pela expressão da enzima UDPG com a tolerância ao estresse hídrico em milho, podendo ser genes candidatos a futuros estudos para esta associação.

A região genômica localizada no cromossomo 2, na posição 197544698 até 197551487 associada ao marcador 10K02922, está próxima a um Modelo de Gene descrito como *ubiquitin carboxyl-terminal hydrolase 2.* O Sistema Ubiquitina (UPS- *Ubiquitin- Proteasome system*) estimula e permite às plantas se adaptarem perante a condições ambientais adversas, incluindo a exposição a estresses abióticos Diante a exposição aos estresses ambientais, há um aumento no nível de proteínas disfuncionais. A remoção destas proteínas é dada pela regulação de alguns

genes pertencentes ao UPS demostrando assim, seu papel crucial no desenvolvimento e crescimento da planta (Xu e Xue., 2019).

Outros trabalhos também vêm associando proteínas ubiquitina-ligases à tolerância ao estresse hídrico em *Arabidopsis thaliana, Oryza sativa* e *Camellia sinensis* (Min et al., 2016; Yang et al., 2018; Xie et al., 2019).

Li et al. (2019) estudaram a associação genômica, filogenética e expressão de 12 genes codificadores de Ubiquitina-ligases em milho, e revelaram uma correlação da expressão de todos os genes estudados em resposta a estresses abióticos incluindo salinidade, estresse hídrico e baixas temperaturas.

Ainda não há na literatura, trabalhos que associem diretamente enzimas conjugadoras da ubiquitina (UCHs - *ubiquitin C-terminal hydrolase*) a uma resposta ao estresse hídrico em milho.

A região genômica localizada no cromossomo 1, na posição 7980584 até 7991204 associada ao marcador 10K00004, está próxima a um Modelo de Gene identificado como *ANkyrin Repeat protein* (ANR). *ANkyrin Repeat Domains* é um dos domínios proteicos mais conservados e comumente encontrado nos procariotos, eucariotos e alguns vírus. Frequentemente mediam interações proteína-proteína, estando envolvidos em um número importante de processos fisiológicos tais como sinalização e crescimento (Sedgwick e Smerdon., 1999; Jiang et al., 2013).

Estudos vêm sendo conduzidos em arroz, pimenta, soja e *Arabidopsis thaliana* objetivando caracterizar e localizar genes codificadores de proteínas ANR, bem como investigar o seu perfil de expressão em resposta a estresses ambientais como salinidade e estresse hídrico (Sakamoto et al., 2013; Zhang et al., 2016).

Jiang et al. (2013) realizaram um estudo de Associação Genômica e do perfil de expressão da família de genes codificadores de proteínas ANR em milho. A proteína identificada pelo Modelo de Gene GRMZM2G092481, localizada no cromossomo 1, também identificada no presente estudo (Tabela 3), pertence à família ANK-A sendo codificada pelo gene *ZmANK2*, gene este com função ainda não descrita.

Dada a importância destas proteínas na resposta à estresses bióticos e abióticos, estudos ligados a associação genômica, relações filogenéticas e padrões de expressão gênica em milho, têm um grande potencial para o entendimento das bases genéticas relacionadas a diferentes estresses, incluindo o estresse hídrico.

A região genômica localizada no cromossomo 1, na posição 223350470 até 223352163 associada ao marcador 10K08801, está próxima a um Modelo de Gene descrito como *DNAJ*

heat shock family protein. *DNAJ heat shock protein* são chaperonas moleculares que regulam o enovelamento, localização, acúmulo e degradação de moléculas de proteínas em animais e plantas (Feder e Hofmann.; 1999). O estresse sofrido pelas plantas, é um dos principais fatores que leva a disfunção de proteínas importantes para manutenção da homeostase celular. Portanto, esta família de proteínas desempenha um importante papel para sobrevivência das plantas sob a condição de estresse através da manutenção da conformação funcional das proteínas e prevenção do acúmulo de proteínas disfuncionais (Hu et al., 2010).

Deif-Abou et al. (2019) realizaram um estudo de análise proteômica de HSPs (*heat shock proteins*) em milho, revelando que o nível de expressão de quatro delas desempenham um importante papel no combate ao estresse causado por altas temperaturas em milho. O estresse de altas temperaturas geralmente está relacionado a estresses hídricos, uma vez que situações de seca normalmente estão associadas a altas temperaturas.

A região genômica localizada no cromossomo 4, na posição 30459788 até 30474790 associada ao marcador 10K04283 está próxima a um Modelo de Gene descrito como *polyadenylation and cleavage factor homolog 4.*

Os fatores de poliadenilação e clivagem em eucariotos exercem um papel fundamental na criação da extremidade 3’ do mRNA, molécula responsável por guiar a síntese proteica de acordo com as instruções genéticas armazenadas no DNA (Alberts et al., 2010).

A poliadenilação também vem sendo relacionada, recentemente, à resposta a estresses abióticos. Ye et al. (2019) associaram o processo de poliadenilação alternativa (APA – *Alternative Polyadenylation*) como indiretamente ligado a regulação da resposta a estresse bióticos e abióticos em arroz.

Além das regiões encontradas e descritas no presente estudo, duas regiões genômicas ainda não caracterizadas, localizadas nos cromossomos 4 e 7 nas posições 178596028 até 178607782 e 88533046 até 8854960, respectivamente, foram encontradas em associação a resposta ao estresse hídrico nos genótipos avaliados. Estas regiões ainda precisam ser melhor caracterizadas para elucidação do seu papel fisiológico associado a tolerância do milho ao estresse abiótico.

4 CONCLUSÃO

No presente trabalho, foram identificados 10 QTLs associados a tolerância do milho ao estresse abiótico, através da avaliação do efeito do estresse nos caracteres secundários avaliados, Pelo menos sete destes QTLs estão localizados em regiões que contém genes ou Modelos de Genes com funções de proteção das plantas a estresses ambientais, incluindo estresse hídrico. Estes QTLs ainda precisam ser validados em um conjunto maior de linhagens, para avaliar a sua aplicação em programas de seleção assistida por marcadores moleculares para tolerância a seca no melhoramento de milho.

REFERÊNCIAS BIBLIOGRÁFICAS

ABOU-DEIF, Mahmoud Hussien et al. Proteomic analysis of heat shock proteins in maize (Zea mays L.). **Bulletin of the National Research Centre**, v. 43, n. 1, p. 199, 2019.

AGRAMA, Hesham AS; MOUSSA, Mounir E. Mapping QTLs in breeding for drought tolerance in maize (Zea mays L.). **Euphytica**, v. 91, n. 1, p. 89-97, 1996.

ALBERTS, Bruce et al. **Biologia molecular da célula**. 6. ed. Porto Alegre: Artmed Editora, 2017.

ALEXANDER, David H.; NOVEMBRE, John; LANGE, Kenneth. Fast model-based estimation of ancestry in unrelated individuals. **Genome research**, v. 19, n. 9, p. 1655-1664, 2009.

ALMEIDA, Gustavo Dias et al. Molecular mapping across three populations reveals a QTL hotspot region on chromosome 3 for secondary traits associated with drought tolerance in tropical maize. **Molecular breeding**, v. 34, n. 2, p. 701-715, 2014.

ALMEIDA, Gustavo Dias et al. QTL mapping in three tropical maize populations reveals a set of constitutive and adaptive genomic regions for drought tolerance. **Theoretical and Applied Genetics**, v. 126, n. 3, p. 583-600, 2013.

ARAUS, José Luis; SERRET, María Dolors; EDMEADES, Greg. Phenotyping maize for adaptation to drought. **Frontiers in physiology**, v. 3, p. 305, 2012.

BÄNZIGER, M.; LAFITTE, H. R. Breeding tropical maize for low N environments. II. The values of secondary traits for improving selection gains under low N. **Crop Sci**, v. 37, p. 1110-1117, 1997.

BÄNZIGER, Marianne et al. **Breeding for drought and nitrogen stress tolerance in maize: from theory to practice**. 1. ed. Mexico: Cimmyt, 2000.

BOLAÑOS, J.; EDMEADES, G. O. The importance of the anthesis-silking interval in breeding for drought tolerance in tropical maize. **Field Crops Research**, v. 48, n. 1, p. 65-80, 1996.

BRADBURY, Peter J. et al. TASSEL: software for association mapping of complex traits in diverse samples. **Bioinformatics**, v. 23, n. 19, p. 2633-2635, 2007.

CATTIVELLI, Luigi et al. Drought tolerance improvement in crop plants: an integrated view from breeding to genomics. **Field Crops Research**, v. 105, n. 1-2, p. 1-14, 2008.

CRUZ, C.D. Genes Software – **Extended and integrated with the R, Matlab and Selegen**. Acta Scientiarum. v. 38, n. 4, p. 547-552, 2016.

EDMEADES, G. O. et al. Causes for silk delay in a lowland tropical maize population. **Crop Science**, v. 33, n. 5, p. 1029-1035, 1993.

FAOSTAT. 2012. **Statistical database of the Food and Agriculture Organization of the United Nations**. FAO, Rome, Italy. Disponível em: http://faostat.fao.org. Acesso em: 02 jun. 2020.

FEDER, Martin E.; HOFMANN, Gretchen E. Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. **Annual review of physiology**, v. 61, n. 1, p. 243-282, 1999.

GE, Tida et al. Effects of water stress on growth, biomass partitioning, and water-use efficiency in summer maize (Zea mays L.) throughout the growth cycle. **Acta Physiologiae Plantarum**, v. 34, n. 3, p. 1043-1053, 2012.

GRANT, R. F. et al. Water deficit timing effects on yield components in maize. **Agronomy journal**, v. 81, n. 1, p. 61-65, 1989.

HENDERSON, Charles R. et al. **Applications of linear models in animal breeding**. Guelph: University of Guelph, 1984.Genes, 2019.

HU, Xiuli et al. Characterization of small heat shock proteins associated with maize tolerance to combined drought and heat stress. **Journal of plant growth regulation**, v. 29, n. 4, p. 455-464, 2010.

JIANG, Haiyang et al. Genome-wide identification and expression profiling of ankyrin-repeat gene family in maize. **Development genes and evolution**, v. 223, n. 5, p. 303-318, 2013.

LI, Yunfeng et al. Genome-wide identification, phylogenetic and expression analysis of the maize HECT E3 ubiquitin ligase genes. **Genetica**, v. 147, n. 5, p. 391-400, 2019.

LIU, Ping-Li et al. Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants. **BMC evolutionary biology**, v. 17, n. 1, p. 47, 2017.

LIU, Yimo; XIONG, Yan; BASSHAM, Diane C. Autophagy is required for tolerance of drought and salt stress in plants. **Autophagy**, v. 5, n. 7, p. 954-963, 2009.

LUDLOW, M. M.; MUCHOW, R. C. A critical evaluation of traits for improving crop yields in water-limited environments. In: **Advances in agronomy**. Academic Press, 1990. p. 107-153.

MESSMER, Rainer et al. Drought stress and tropical maize: QTL-by-environment interactions and stability of QTLs across environments for yield components and secondary traits. **Theoretical and Applied Genetics**, v. 119, n. 5, p. 913-930, 2009.

MIN, Hye Jo et al. CaPUB1, a hot pepper U-box E3 ubiquitin ligase, confers enhanced cold stress tolerance and decreased drought stress tolerance in transgenic rice (Oryza sativa L.). **Molecules and cells**, v. 39, n. 3, p. 250, 2016.

NEPOLEAN, T. et al. Molecular characterization and assessment of genetic diversity of inbred lines showing variability for drought tolerance in maize. **Journal of plant biochemistry and biotechnology**, v. 22, n. 1, p. 71-79, 2013.

NEPOLEAN, Thirunavukkarasu et al. Functional mechanisms of drought tolerance in subtropical maize (Zea mays L.) identified using genome-wide association mapping. **BMC genomics**, v. 15, n. 1, p. 1182, 2014.

NEPOLEAN, Thirunavukkarasu et al. Genomics-enabled next-generation breeding approaches for developing system-specific drought tolerant hybrids in maize. **Frontiers in plant science**, v. 9, p. 361, 2018.

NIKOLIĆ, Ana et al. Quantitative trait loci for yield and morphological traits in maize under drought stress. **Genetika**, v. 43, n. 2, p. 263-276, 2011.

OSTROWSKI, Maciej; HETMANN, Anna; JAKUBOWSKA, Anna. Indole-3-acetic acid UDP-glucosyltransferase from immature seeds of pea is involved in modification of glycoproteins. **Phytochemistry**, v. 117, p. 25-33, 2015.

OUYANG, Shou-Qiang et al. Receptor-like kinase OsSIK1 improves drought and salt stress tolerance in rice (Oryza sativa) plants. **The Plant Journal**, v. 62, n. 2, p. 316-329, 2010.

RAHMAN, H. et al. Molecular mapping of quantitative trait loci for drought tolerance in maize plants. **Genet Mol Res**, v. 10, n. 2, p. 889-901, 2011.

RESENDE, MDV de et al. **O software selegen-Reml/Blup**. 2006.

SABADIN, P. K. et al. Studying the genetic basis of drought tolerance in sorghum by managed stress trials and adjustments for phenological and plant height differences. **Theoretical and Applied Genetics**, v. 124, n. 8, p. 1389-1402, 2012.

SAKAMOTO, H.; NAKAGAWARA, Y.; OGURI, S. The expression of a novel gene encoding an ankyrin-repeat protein, DRA1, is regulated by drought-responsive alternative splicing. **Int. J. Biol. Veterinary Agr. Food Engin**, v. 7, p. 12, 2013.

SAS INSTITUTE. **SAS/STAT user's guide: version 6**. SAS Institute Incorporated, 1990.

SEDGWICK, Steven G.; SMERDON, Stephen J. The ankyrin repeat: a diversity of interactions on a common structural framework. **Trends in biochemical sciences**, v. 24, n. 8, p. 311-316, 1999.

SHEIKH, F. M. et al. Recent advances in breeding for abiotic stress (drought) tolerance in Maize. **Int J Curr Microbiol App Sci**, v. 6, n. 4, p. 2226-2243, 2017.

SHIKHA, Mittal et al. Genomic selection for drought tolerance using genome-wide SNPs in maize. **Frontiers in plant science**, v. 8, p. 550, 2017.

SHIU, Shin-Han; BLEECKER, Anthony B. Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. **Proceedings of the National Academy of Sciences**, v. 98, n. 19, p. 10763-10768, 2001.

SIGNORELLI, Santiago et al. Linking autophagy to abiotic and biotic stress responses. **Trends in plant science**, v. 24, n. 5, p. 383-478, 2019.

SU, Tong et al. Autophagy: An Intracellular Degradation Pathway Regulating Plant Survival and Stress Response. **Frontiers in Plant Science**, v. 11, p. 164, 2020.

TEJEDA, Luis Herminio Chairez et al. Abiotic stress and self-destruction: ZmATG8 and ZmATG12 gene transcription and osmotic stress responses in maize. **Biotechnology Research and Innovation**, 2020.

WANG, Nan et al. Identification of loci contributing to maize drought tolerance in a genome-wide association study. **Euphytica**, v. 210, n. 2, p. 165-179, 2016.

WANG, Nan et al. Identification of loci contributing to maize drought tolerance in a genome-wide association study. **Euphytica**, v. 210, n. 2, p. 165-179, 2016.

WEI, Kaifa; LI, YiXuan. Functional genomics of the protein kinase superfamily from wheat. **Molecular Breeding**, v. 39, n. 11, p. 141, 2019.

WOODWARD, Andrew W.; BARTEL, Bonnie. Auxin: regulation, action, and interaction. **Annals of botany**, v. 95, n. 5, p. 707-735, 2005.

XIE, Hui et al. Global ubiquitome profiling revealed the roles of ubiquitinated proteins in metabolic pathways of tea leaves in responding to drought stress. **Scientific reports**, v. 9, n. 1, p. 1-12, 2019.

XU, Fa-Qing; XUE, Hong-Wei. The ubiquitin-proteasome system in plant responses to environments. **Plant, cell & environment**, v. 42, n. 10, p. 2931-2944, 2019.

XUE, Yadong et al. Genome-wide association analysis for nine agronomic traits in maize under well-watered and water-stressed conditions. **Theoretical and applied genetics**, v. 126, n. 10, p. 2587-2596, 2013.

YANG, Liang et al. Overexpression of the maize E3 ubiquitin ligase gene ZmAIRP4 enhances drought stress tolerance in Arabidopsis. **Plant Physiology and Biochemistry**, v. 123, p. 34-42, 2018.

YE, Congting et al. Genome-wide alternative polyadenylation dynamics in response to biotic and abiotic stresses in rice. **Ecotoxicology and environmental safety**, v. 183, p. 109485, 2019.

ZAMAN-ALLAH, M. et al. **Phenotyping for abiotic stress tolerance in maize: drought stress: a field manual.** 1. ed. Mexico: Cimmyt, 2016.

ZHAI, Yufei et al. Autophagy, a conserved mechanism for protein degradation, responds to heat, and other abiotic stresses in Capsicum annuum L. **Frontiers in plant science**, v. 7, p. 131, 2016.

ZHANG, Dayong et al. Genome-wide characterization of the ankyrin repeats gene family under salt stress in soybean. **Science of the Total Environment**, v. 568, p. 899-909, 2016.

ZHANG, Jiazi et al. The Responses of Wheat Autophagy and ATG8 Family Genes to Biotic and Abiotic Stresses. **Journal of Plant Growth Regulation**, p. 1-10, 2019.

ZHANG, Xuehai et al. Genome-wide association studies of drought-related metabolic changes in maize using an enlarged SNP panel. **Theoretical and applied genetics**, v. 129, n. 8, p. 1449-1463, 2016.

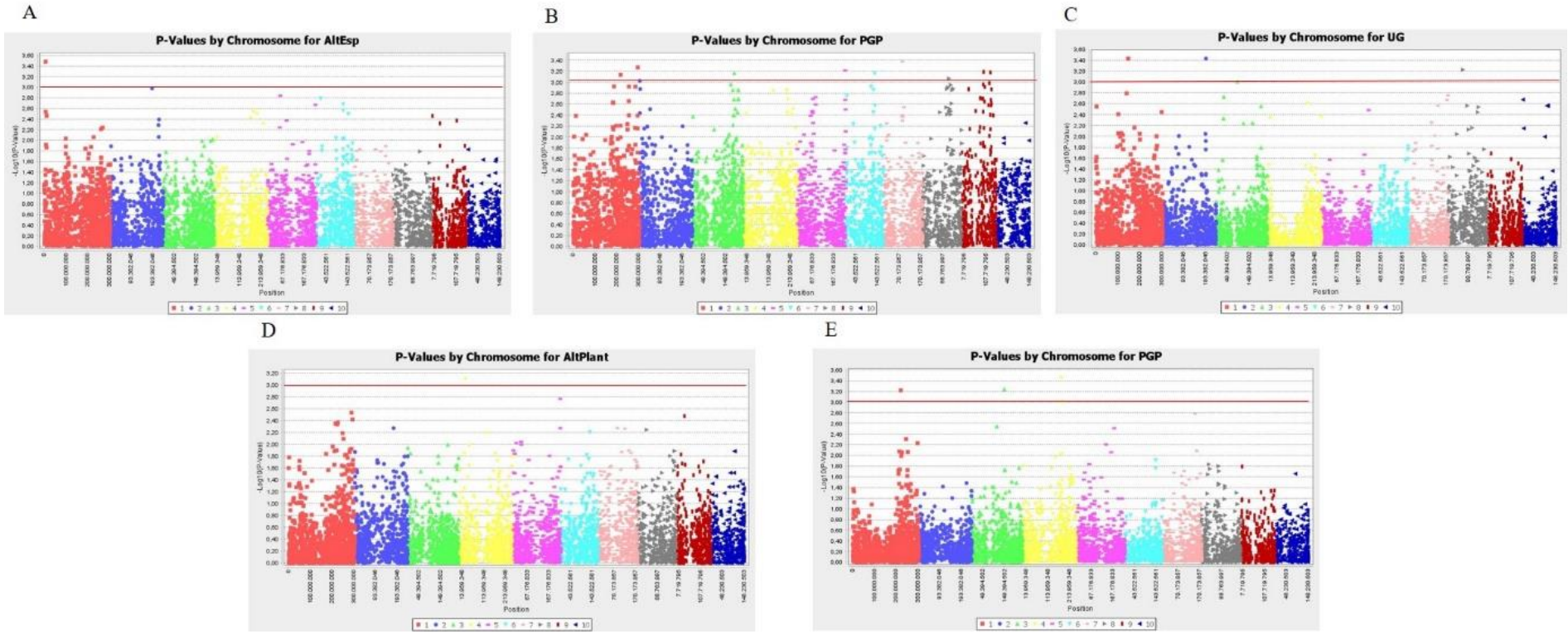
ZHAO, Xiaoqiang et al. Comparative QTL analysis for yield components and morphological traits in maize (Zea mays L.) under water-stressed and well-watered conditions. **Breeding Science**, v. 69, n. 4, p. 621-632, 2019.

ZHAO, Xiaoqiang et al. Identification of QTLs and meta-QTLs for seven agronomic traits in multiple maize populations under well-watered and water-stressed conditions. **Crop Science**, v. 58, n. 2, p. 507-520, 2018.

ZIYOMO, Cathrine; BERNARDO, Rex. Drought tolerance in maize: Indirect selection through secondary traits versus genomewide selection. **Crop Science**, v. 53, n. 4, p. 1269-1275, 2013.

ANEXO I

Figura 1 - *Manhattan Plot* do estudo de Associação Genômica para diferença nos valores características avaliadas com e sem estresse hídrico onde A, B e C representam respectivamente: Diferença na Altura de Espiga, Diferença no Peso de Grãos por Parcela e Diferença na Umidade dos Grãos para Jardinópolis-SP e D e E representam respectivamente: Diferença na Altura de Planta, e Diferença no Peso de Grãos por Parcela para Sorriso-MT. A linha horizontal vermelha indica o ponto de corte para 0,1% de probabilidade.



Fonte: Tassel (2020).

ANEXO II

Tabela 1 - Resultados da Análise de Variância das diferenças entre as características avaliadas com e sem estresse hídrico.

Local	Características	GL Genótipo	GL Resíduo	QM Genótipo	QM Resíduo	PROB
Sorriso	ΔP100	73	118	16.71	22.65	100.0
	ΔPGP	73	141	992,920.87	275,366.19	< 0.01
	ΔPROL	73	136	0.1036	0.0366	< 0.01
	ΔAlt Plant	73	145	734.20	90.09	< 0.01
	ΔAlt Esp	73	145	317.44	52.39	< 0.01
Jardinópolis	ΔP100	72	108	100,484.26	16,030.51	< 0.01
	ΔPGP	72	106	1,686,073,529.94	514,670,289.09	< 0.01
	ΔUG	61	79	33.00	4.92	< 0.01
	ΔAlt Plant	72	141	1,125.74	122.12	< 0.01
	ΔAlt Esp	72	141	406.97	57.50	< 0.01

Fonte: Do autor (2020).

GL Genótipo=graus de liberdade do genótipo; GL Resíduo=graus de liberdade do resíduo; QM Genótipo=quadrado médio do genótipo; QM Resíduo=quadrado médio do resíduo; PROB=probabilidade.

ANEXO III

Tabela 2 - Resultados da Análise de Regressão Múltipla dos marcadores moleculares associados a 0.1% de probabilidade na análise de Modelos Lineares Mistos e a diferença nos valores das características avaliadas com e sem estresse hídrico.

Local	Característica	Marcador	Cromossomo	Posição	R2 (%)	P-valor
Jardinópolis	ΔPGP	10K06573	7	88548758	0,2054	<0,0001
		10K13323	3	192799311	0,4223	<0,0001
		10K09522	6	133575442	0,5063	0,0012
	ΔUG	10K07174	8	67034962	0,2505	0,0025
		10K02922	2	197548923	0,3446	0,0006
		10K03498	3	97976555	0,4351	0,0031
	ΔAlt Esp	10K00004	1	7987927	0,0923	0,0081
Sorriso	ΔPGP	10K04703	4	178596640	0,2164	0,0002
		10K08801	1	223351538	0,3518	0,0003
	ΔAlt Plant	10K04283	4	30467312	0,0978	0,0063

Fonte: Do autor (2020).

R^2=coeficiente de determinação; PROB=probabilidade.

ANEXO IV

Tabela 3 - Genes candidatos ou região genômica vinculada aos polimorfismos de SNPs mais fortemente associados a tolerância ao estresse hídrico.

Local	Caract. Avaliada	Cromossomo	Início	Fim	Modelo de Gene	Gene	Descrição do Gene
Jardinópolis	ΔPGP	7	88533046	88549640	GRMZM2G126170		uncharacterized
		3	192795015	192801094	GRMZM2G116700	*atg18f*	autophagy-related protein 18f [*Zea Mays*]
		6	133574533	133578647	GRMZM2G131609		LRR receptor-like serine/threonine-protein kinase *[Zea Mays]*
	ΔUG	8	67033198	67034788	GRMZM2G022101		indole-3-acetate beta-glucosyltransferase *[Zea Mays]*
		2	197544698	197551487	GRMZM2G014917		ubiquitin carboxyl-terminal hydrolase 2 *[Zea Mays]*
		3	97975251	97982164	GRMZM2G063306	*ys3*	protein zinc induced facilitator-like1 *[Zea Mays]*
	ΔAlt Esp	1	7980584	7991204	GRMZM2G092481		ankyrin repeat protein *[Zea Mays]*
Sorriso	ΔPGP	4	178596028	178607782	GRMZM2G053766		uncharacterized
		1	223350470	223352163	GRMZM2G047153		DNAJ heat shock family protein *[Zea Mays]*
	ΔAlt Plant	4	30459788	30474790	GRMZM2G007734		polyadenylation and cleavage factor homolog 4 *[Zea Mays]*

Fonte: Do Autor (2020).

Caract. Avaliada= Característica Avaliada.

ANEXO V

Título

Tutorial para Análise de Associação Genômica Ampla (GWAS) por meio da ferramenta TASSEL na rotina da Empresa LongPing HighTech

INTRODUÇÃO

A utilização de ferramentas da bioinformática vem se tornando uma prática comum no contexto de ensino, pesquisa e empresarial. A importância da GWAS para a obtenção de genótipos superiores é uma realidade e há muita expectativa em relação aos ganhos efetivos que esta tecnologia pode proporcionar para esta empresa. Entre as ferramentas utilizadas, optou-se pela análise de associação utilizando Modelos Lineares Mistos (MLM) no programa TASSEL. Uma das grandes precauções associadas à sua utilização é que exista uma rotina bem delineada e livre de possíveis efeitos imprevistos nas análises. Para isto, o Laboratório Integrado de Análises Genômicas de Cravinhos (iCGA) vem trabalhando sobre um workflow que tem se mostrado bastante adequado para os interesses e princípios do grupo LongPingHighTech. Este produto técnico, atrelado ao trabalho de conclusão de curso de Mestrado Profissional, é elaborado em um bom momento, uma vez que a padronização das atividades e a sequência lógica das execuções têm grande relevância para a otimização dos resultados alcançados. A partir da próxima seção, este tutorial será apresentado de forma bastante objetiva e sintética, favorecendo, aos olhos desta equipe, a melhor maneira de apresentação encontrada. O objetivo central deste tutorial é fornecer um caminho seguro e eficaz para que o time de pesquisa possa percorrer, chegando com precisão ao melhor destino em suas análises de rotina. Assim, uma vez que este documento já se encontra implantado e em pleno funcionamento na empresa, espera-se que siga propiciando boas diretrizes para o planejamento no âmbito da pesquisa e desenvolvimento de genótipos superiores

Tutorial

ANÁLISE DE ASSOCIAÇÃO GENÔMICA AMPLA (GWAS), UTILIZANDO TASSEL

Elaborado por:	**Revisado por:**	**Aprovado por:** *Ivan Schuster em 09/06/2020*
Renata Bomfim	*Ivan Schuster*	

*Escopo:

Este é um tutorial que exemplifica uma rotina de trabalho de bioinformática para análise de associação entre marcadores SNPs e QTLs, em uma população de mapeamento, para uma característica desejada utilizando-se MLM no programa TASSEL. Este material é utilizado pelo Laboratório Integrado de Análises Genômicas de Cravinhos (iCGA) em análises de rotina.

Termos/Abreviações e suas definições:

Termo	Definição
GWAS	*Genome-Wide Association Study (*Estudo de Associação Genômica Ampla)
iCGA	*Integrated Cravinhos Genome Analysis* (Laboratório Integrado de Análises Genômicas de Cravinhos)
EPI	Equipamento de Proteção Individual
SNPs	*Single Nucleotide Polymorphism (*Polimorfismo de Nucleotídeo Único)
QTLs	*Quantitative Trait Loci* (Locos controladores de características quantitativas).
MLM	Modelos Lineares Mistos
TASSEL	*Trait Analysis by Association, Evolution and Linkage* (Programa utilizado para análises de Associação Genômica, Coeficientes de Parentesco e Desequilíbrio de Ligação).
t-GBS	*Target genotyping-by-sequencing* (Genotipagem por Sequenciamento Alvo-Direcionado)
DMS	*Document Management System (*Sistema de Gerenciamento de Documentos)

Passo a Passo

Para análise de GWAS no programa Tassel, é necessário preparo de três arquivos de dados, no formato (txt) – texto separado por tabulação:

- ✓ Arquivo de dados Genotípicos.
- ✓ Arquivo de dados Fenotípicos.
- ✓ Arquivo com a estrutura de Populações.

1

Preparo dos dados Genotípicos

i. Abrir o arquivo contendo os dados genotípicos obtidos a partir da análise de sequênciamento (t-GBS).

O arquivo contém as seguintes informações: nome do marcador (*Marker*), cromossomo em que este marcador está localizado (Chr), posição no cromossomo (Pos), identificação das amostras genotipadas (Gen 1, Gen2…) e o genótipo de cada marcador (Figura 1):

	A	B	C	D
1	Marker	Marker 1	Marker 2	Marker 3
2	Chr	chr1	chr1	chr1
3	Pos	0	1	2
4	Geno 1	AA	TT	GG
5	Geno 2	AA	TT	GG
6	Geno 3	AA	./.	AA

Figura 1: Tabela obtida a partir do sequênciamento (t-GBS).

ii. Preparar o arquivo **hapmap** a partir do arquivo obtido no passo (i).

Colar transposto as informações acima em um novo arquivo Excel.

O novo arquivo precisa ter todas as colunas identificadas, como no exemplo abaixo (Figura 2), seguindo-se o número de colunas até finalizar o número de genótipos.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panel	QCcode	Geno 1	Geno 2	Geno 3	Geno 4
2	Marker 1	A/T	chr1	0	+	NA	NA	NA	NA	NA	NA	AA	AA	AA	AA
3	Marker 2	T/G	chr1	1	+	NA	NA	NA	NA	NA	NA	TT	TT	./.	TT
4	Marker 3	G/A	chr1	2	+	NA	NA	NA	NA	NA	NA	GG	GG	AA	GG

Figura 2: Tabela formatada para entrada no programa TASSEL.

Identificar a coluna que contém o nome dos marcadores como “rs#” (sem espaço).

A coluna “*alleles”* deve ser preenchida com os alelos correspondentes a cada marcador utilizado na análise. Esta informação está contida em um documento localizado no *DMS*. Para associação correta entre as colunas de cada documento, utilizar a fórmula “=PROCV“ do Excel.

O cabeçalho da coluna “Chr” deve ser substituído por “chrom”. As informações da coluna podem ser mantidas conforme documento anterior.

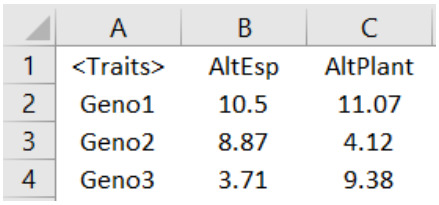
A coluna “pos” pode ser mantida com cabeçalho e informações do documento anterior.

A coluna *strand* (fita) deve ser preenchida, para plantas, sempre na mesma orientação (+) para todos os marcadores.

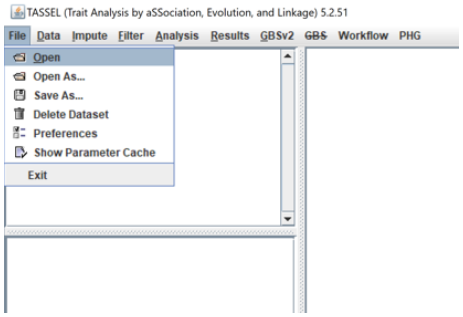
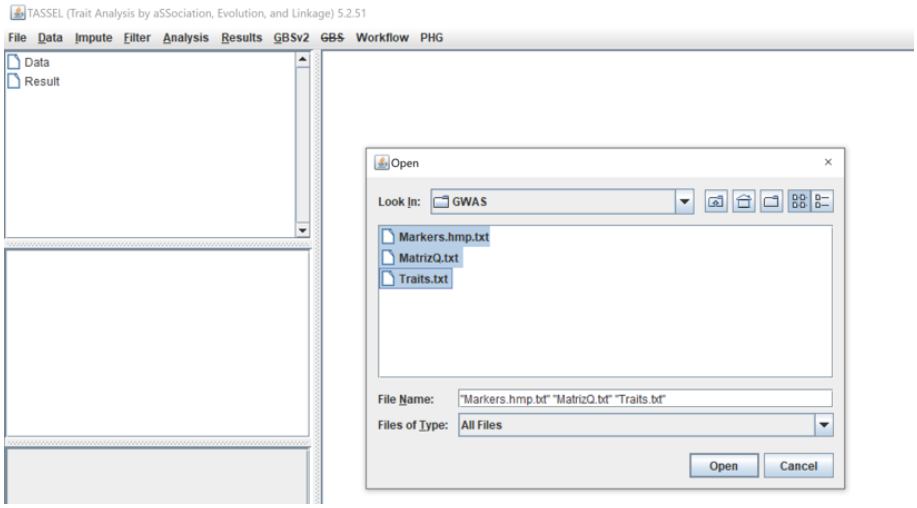
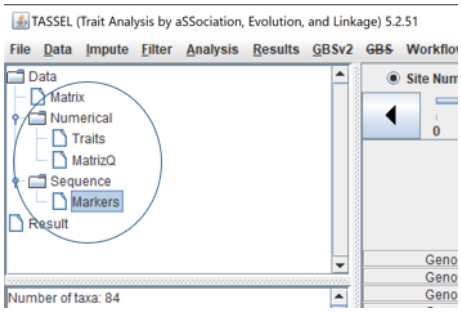
Para as colunas: *“assembly#, “center”, “protLSID”, “assayLSID”, “panel”* e “*QCcode”* preencher com NA.

Dados perdidos (./.) devem ser codificados como “??”.

Salvar o Arquivo como “nome.hmp”, no formato (txt); exemplo: “Markers.hmp.txt”.

2 **Preparo dos dados Fenotípicos**	Os dados fenotípicos obtidos devem ser planilhados seguindo a formatação da Figura 3:  **Figura 3**: Tabela formatada com dados Fenotípicos. A primeira linha da primeira coluna deve ser identificada como <trait> para o programa entender que trata-se do dado fenotípico. A primeira coluna identifica os nomes das linhagens/variedades (sem espaço no nome). As colunas seguintes contêm os dados fenotípicos, com a identificação da característica avaliada na primeira linha (sem espaço no nome). De caráter obrigatório, utilizar como separador decimal o ponto. Não considerar a unidade de medida da característica na tabela; exemplo: cm. Salvar o arquivo no formato (txt); exemplo: “Traits.txt”.

3 **Arquivo contendo a Estrutura de Populações**	O arquivo “Q” com a estrutura da população deve ser obtido previamente usando o programa *Structure*, *Instruct ou Admixture*. Neste contexto, deve-se seguir o procedimento para Análise de Estrutura de Populações localizado no *DMS*. O arquivo “Q” (Obtido pela análise de Estrutura de Populações) para entrada no programa TASSEL deve ter o formato exemplificado na Figura 4: A primeira linha deve conter <Covariate> na primeira coluna para o programa entender que se trata de uma covariável. A primeira coluna contém os nomes dos genótipos, identificados na segunda linha do arquivo com <Trait>. Pode-se usar o nome das linhagens/variedades (sem espaço no nome). As demais colunas contêm o número de grupos em que a população está estruturada, e a probabilidade de cada indivíduo pertencer a cada grupo. Os cabeçalhos de cada coluna contêm a identificação “Qn”, onde ‘n” é o número do grupo. A B C D E F 1 <Covariate> 2 <Traits> Q1 Q2 Q3 Q4 Q5 3 Geno1 0.00001 0.00001 0.99996 0.00001 0.00001 4 Geno2 0.280231 0.154537 0.450986 0.00001 0.114236 5 Geno3 0.00001 0.00001 0.99996 0.00001 0.00001 **Figura 4**: Tabela formatada com dados do arquivo “Q”. O arquivo deve ser salvo no formato (txt); exemplo: “MatrizQ.txt”.

4 **Análise MLM utilizando-se TASSEL**	Abrir o programa TASSEL.  **i- <u>Leitura dos Dados</u>** Clicar em *File/Open* (Figura 5).  **Figura 5** Selecionar os três arquivos gerados a partir dos passos 1, 2 e 3 deste tutorial, pode-se fazer a seleção dos três arquivos simultaneamente (Figura 6).  **Figura 6** Após a seleção os arquivos aparecerão na primeira janela do lado esquerdo do programa (Figura 7).  **Figura 7**

ii- **Filtro dos Dados**

Filtrar o Arquivo com o Fenótipo, para analisar uma característica por vez (Figura 8).

Selecionar o arquivo que contém os fenótipos *"Traits"*, clicar em *Filter*, clicar em *Traits*, desmarcar as características que não serão analisadas, mantendo apenas uma e a informação "Taxa" marcadas (Figura 9).

Clicar em OK, o arquivo temporário: "*Filtered_Traits*" com os dados fenotípicos foi criado (Figura 10).

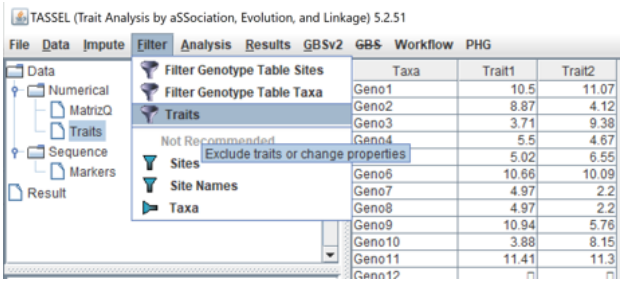


Figura 8

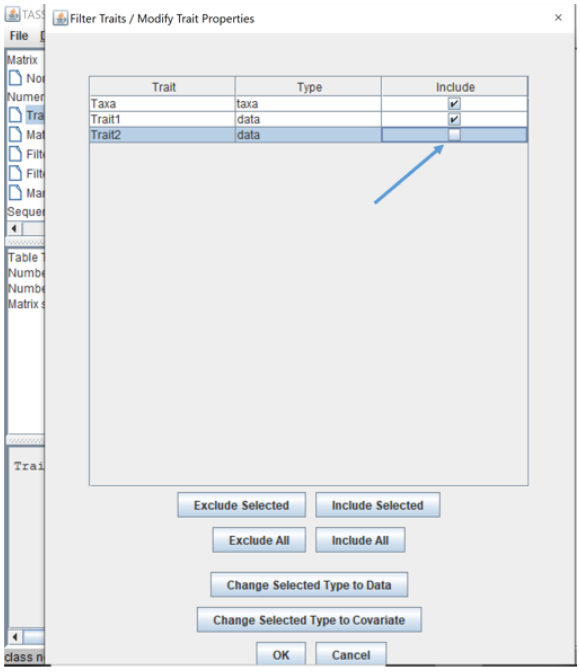


Figura 9

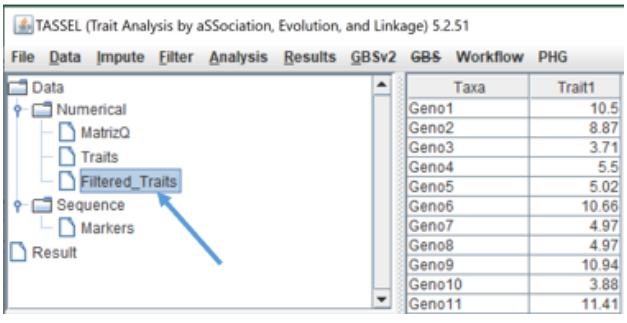


Figura 10

Filtrar o Arquivo de Estrutura de Populações.

Selecionar o arquivo que contém a estrutura da população *"MatrizQ",* clicar em *Filter*, clicar em *Traits*, desmarcar o último grupo da Matiz Q e clicar em OK. (Figuras 11 e 12).

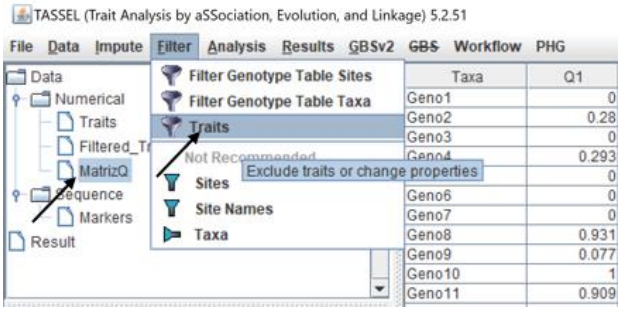


Figura 11

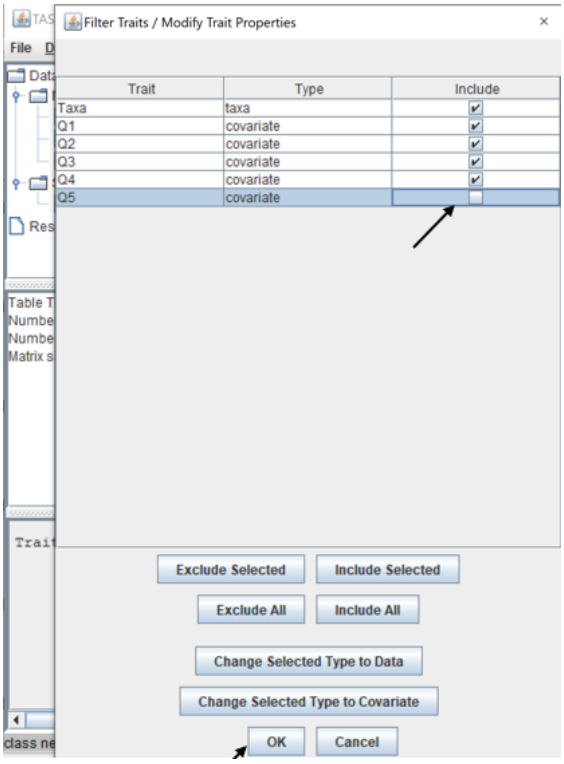


Figura 12

O Arquivo temporário: "*Filtered_MatrizQ*" com os dados da estrutura de população foi criado (Figura 13).

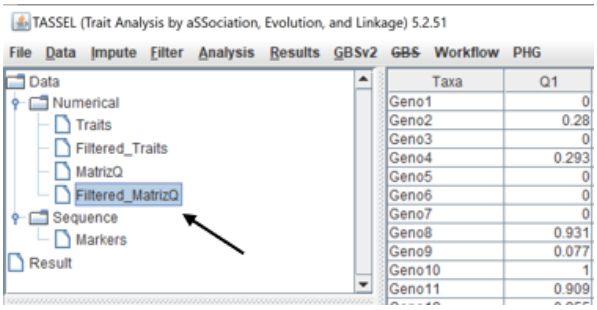


Figura 13

iii- Geração da Matriz de Parentesco *Kinship*

Marcar o arquivo com os dados genotípicos "*Markers*", clicar em *Analysis*, clicar em *Relatedness* clicar em *Kinship* (Figura 14).

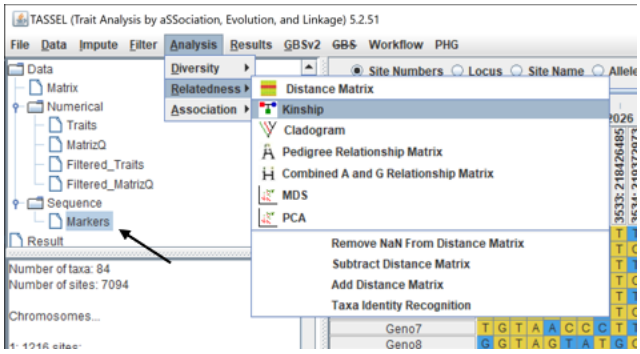


Figura 14

Selecionar o método de análise "NORMALIZED_IBS" (Identidade por semelhança normalizada) e clicar OK, a matriz temporária de parentesco foi criada com o nome "*Normalized_IBC_Markers*" (Figuras 15 e 16).

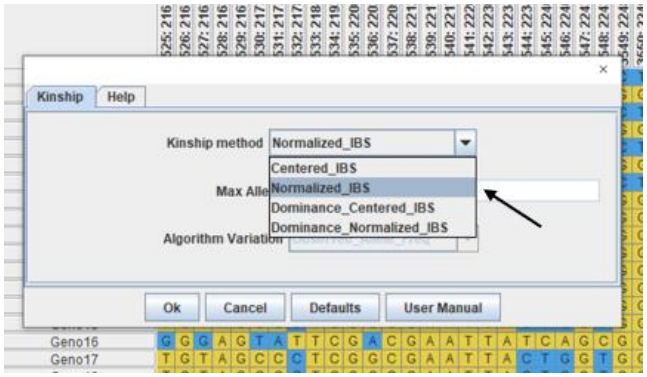


Figura 15

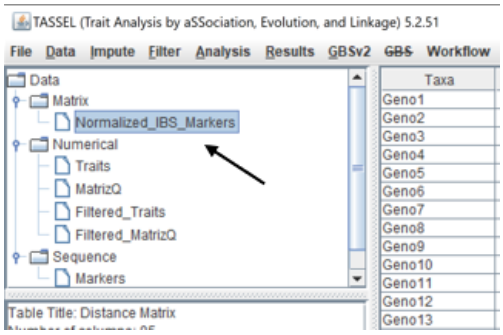


Figura 16

iv- União dos dados para Análise MLM

Unir os arquivos com os dados genotípicos, fenotípicos e de estrutura de populações em um único arquivo.

Selecionar os arquivos filtrados de dados fenotípicos e estrutura de população, e o arquivo de marcadores, com a tecla *Ctrl* pressionada. Clicar em *Data* e clicar em *Union Join* (Figuras 17 e 18).

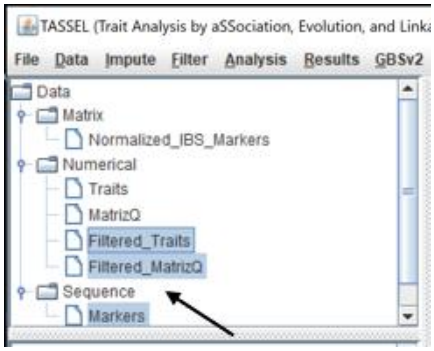


Figura 17

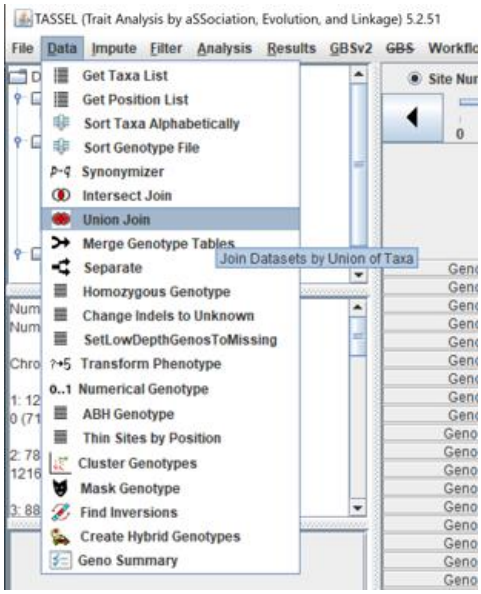


Figura 18

Um novo arquivo com todas as informações será criado, com o nome dos arquivos originais ligados pelo sinal de + (Figura 19).

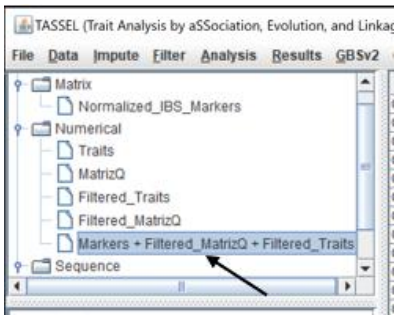


Figura 19

v- <u>Análise MLM</u>

Com a Tecla *Ctrl* pressionada, selecione o arquivo criado no passo acima e o arquivo *Kinship* - "*Normalized_IBC_Markers*" - Clicar em *Analysis*, clicar em *Association* e clicar em MLM (Figuras 20 e 21).

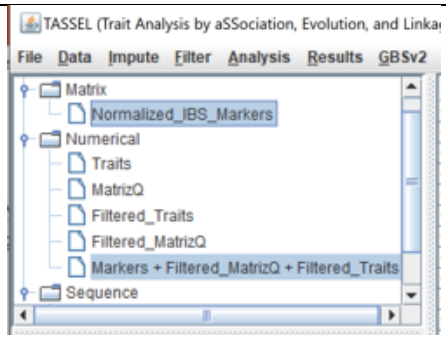


Figura 20

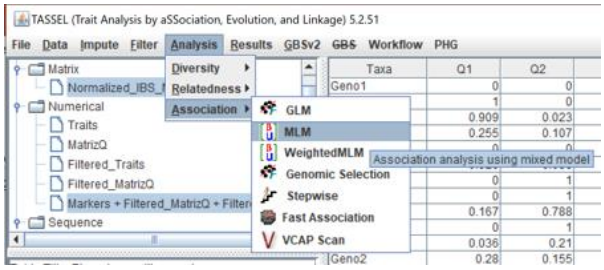


Figura 21

Nas duas telas seguinte clicar em *Run* e *Okay* (Figuras 22 e 23).

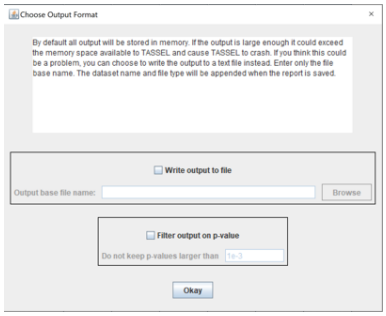


Figura 22 **Figura 23**

Os arquivos de Resíduo, Estatística e Efeitos serão criados (Figura 24).

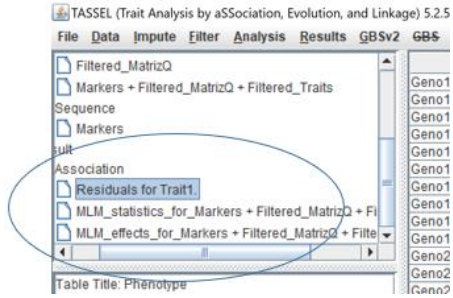
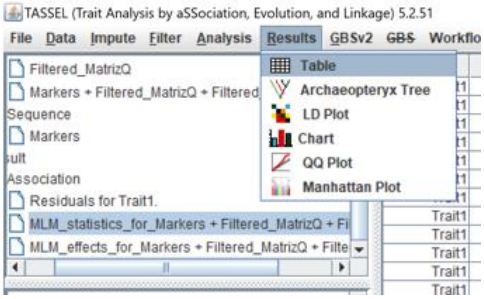
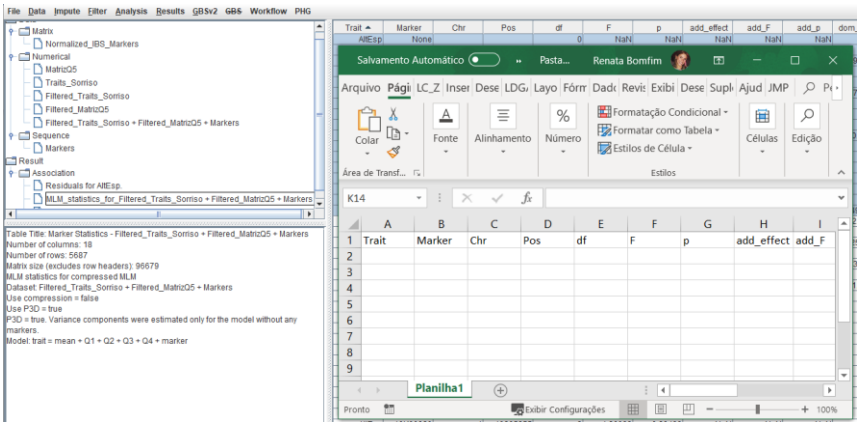
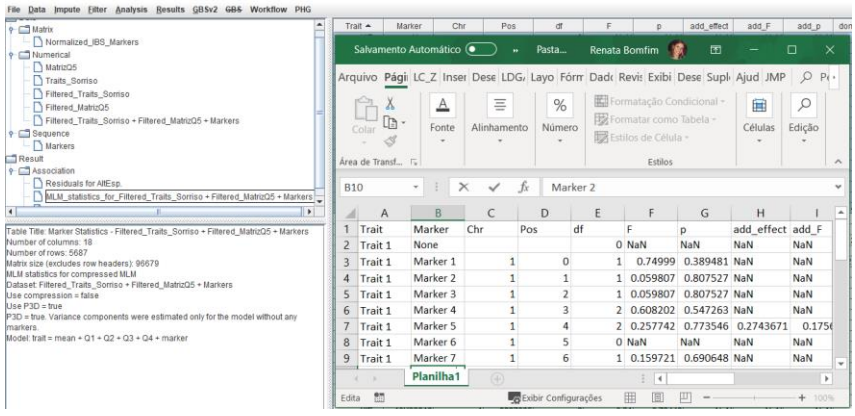


Figura 24

5 **Exportação dos dados gerados**	**i-** **<u>Para exportação dos dados Estatísticos, Efeitos e Resíduos seguir o passo abaixo:</u>** Selecionar cada um dos arquivos gerados, clicar em *Results*, clicar em *Table*, exportar os arquivos no formato desejado e salvar no local desejado (Figura 25).  **Figura 25** **OBS**: Em alguns casos, quando a configuração do computador usa vírgula como separador de decimais, ao abrir o arquivo salvo desta forma no Excel, a configuração dos números fica errada. Uma forma simples de contornar sem alterar a configuração do computador é copiar os dados diretamente na tela do Tassel e colar em um arquivo Excel. Neste caso, os cabeçalhos não podem ser copiados e precisam ser digitados no Tassel (Figura 26). Após o preenchimento manual do cabeçalho, copia-se o restante da tabela e cola-se na mesma planilha do Excel (Figura 27).  **Figura 26**  **Figura 27**

ii- Descrição dos dados estatísticos obtidos.

Trait1	Marker	Chr	Pos	df	F	p	add_effect	Add_F	add_p	dom_effec	dom_F	dom_p	errordf	MarkerR2	GeneticVar	ResidualVar	-2LnLikelihood
Trait1	None			0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	71	NaN	1.97E-04	19.703253	407.3073113
Trait1	Marker 1	1	0	1	0.74999	0.38948	NaN	NaN	NaN	NaN	NaN	NaN	75	0.01019	1.97E-04	19.703253	407.3073113
Trait1	Marker 2	1	1	1	0.05981	0.80753	NaN	NaN	NaN	NaN	NaN	NaN	75	8.13E-04	1.97E-04	19.703253	407.3073113
Trait1	Marker 3	1	2	1	0.05981	0.80753	NaN	NaN	NaN	NaN	NaN	NaN	75	8.13E-04	1.97E-04	19.703253	407.3073113
Trait1	Marker 4	1	3	2	0.6082	0.54726	NaN	NaN	NaN	NaN	NaN	NaN	75	0.01653	1.97E-04	19.703253	407.3073113
Trait1	Marker 5	1	4	2	0.25774	0.77355	0.27436709	0.1757	0.67644	2.80015715	0.38342	0.53785	75	0.007	1.97E-04	19.703253	407.3073113
Trait1	Marker 6	1	5	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	75	0	1.97E-04	19.703253	407.3073113
Trait1	Marker 7	1	6	1	0.15972	0.69065	NaN	NaN	NaN	NaN	NaN	NaN	75	0.00217	1.97E-04	19.703253	407.3073113
Trait1	Marker 8	1	7	1	0.18946	0.66473	NaN	NaN	NaN	NaN	NaN	NaN	75	0.00257	1.97E-04	19.703253	407.3073113

Figura 28 – Arquivo "Statistics" obtido pela Análise MLM – TASSEL**.**

Coluna	Descrição
Trait1	Nome do Trait
Marker	Nome do Marcador
Chr	Número do Cromossomo
Pos	Posição do marcador no cromossomo
df	Graus dre liberdade
F	Valor da estatística F
p	Probabilidade associada ao teste de F
add_effect	Effeito aditivo. Neste exemplo a maior parte dos marcadores nao apresenta valor, pois trata-se de linhagens homozigotas na maioria.
Add_F	Valor da estatística F para o efeito aditivo
add_p	Probabilidade associada ao teste de F para o efeito aditivo
dom_effec	Effeito de dominância. Neste exemplo a maior parte dos marcadores nao apresenta valor, pois trata-se de linhagens homozigotas na maioria.
dom_F	Valor da estatística F para o efeito de dominância
dom_p	Probabilidade associada ao teste de F para o efeito de dominância
errordf	Graus de liberdade do resíduo
MarkerR2	Valor do R2 de cada marcador
GeneticVar	Variância Genética da popjulação
ResidualVar	Variância residual
-2LnLikelihood	Valor de -2 vezes o Ln da verossimilhança

Figura 29 – Descrição do significado dos dados contidos nas colunas do Arquivo.

iii- **Para geração do *Mahattan Plot,* selecionar o arquivo *"MLM_statistics…',* clicar em *Results* e clicar em *Manhattan Plot** (*Figura 30).

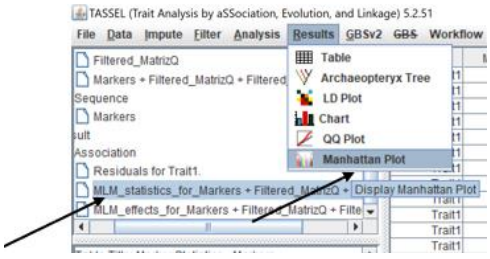


Figura 30

Após a geração do *"Manhattan Plot*", este, pode ser salvo preferencialmente em formato "JPG", clicando-se em *Save* e selecionando o local desejado (Figura 31).

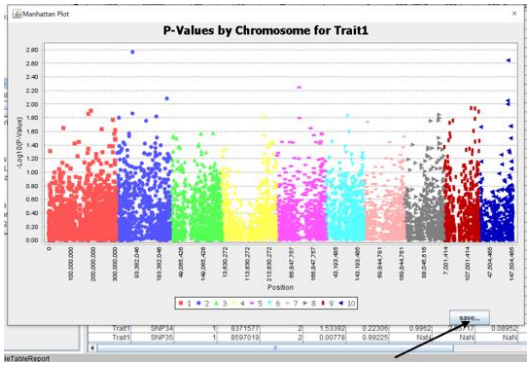


Figura 31

	Após o procedimento aplicado, com os resultados obtidos, realizar a interpretação dos dados conforme os objetivos do projeto em análise.

CONSIDERAÇÕES FINAIS

É importante destacar que este Tutorial vem sendo utilizado, desde 2019, pelo Laboratório de Pesquisa de Cravinhos/SP como ferramenta de análise de bioinformática, com o objetivo de realizar Análise de Associação Genômica (MLM -TASSEL) em características de interesse para a Companhia – LongPingHighTech. A percepção de utilidade deste documento tem sido altamente produtiva, uma vez que otimiza o tempo necessário para as análises de rotina da empresa e assegura uma execução padronizada, bem como os resultados obtidos.

Para os próximos anos, acredita-se que a sua adoção seja implementada de forma definitiva e que o treinamento de novos usuários da empresa seja facilitado, bastando o seguimento dos passos enumerados. Adicionalmente, controle de qualidade dos dados poderia ser facilmente adotado, uma vez que análises possam ser realizadas em mão-dupla, envolvendo tanto o aprendiz quanto o treinador, e os dados finais comparados, com a finalidade de detectar possíveis equívocos no procedimento.

Por fim, é relevante ressaltar que a revisão deste documento tem sido realizada anualmente, com os ajustes necessários para a melhoria das saídas de dados.