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

Expressão alelo-específica em genótipos, órgãos, tempos e condições de estresse hídrico em cana-de-açúcar

Guilherme Bovi Ambrosano

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

Piracicaba
2022

Guilherme Bovi Ambrosano
Engenheiro Agrônomo

Expressão alelo-específica em genótipos, órgãos, tempos e condições de estresse hídrico em cana-de-açúcar

versão revisada de acordo com a Resolução CoPG 6018 de 2011

Orientador
Prof. Dr. **GABRIEL RODRIGUES ALVES MARGARIDO**

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

Piracicaba
2022

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

Ambrosano, Guilherme Bovi

Ocorrência em nível genômico de expressão alelo-específica em diferentes genótipos, órgãos, tempos e condições de estresse hídrico em cana-de-açúcar / Guilherme Bovi Ambrosano. - - versão revisada de acordo com a Resolução CoPG 6018 de 2011. - - Piracicaba, 2022.

50 p.

Dissertação (Mestrado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Expressão gênica diferencial 2. Dose alélica 3. Poliploidia 4. *Saccharum* 5. RNA-seq I. Título

AGRADECIMENTOS

Sou grato ao Prof. Dr. Gabriel Rodrigues Alves Margarido pelos seis anos de orientação, ao longo da graduação e do mestrado. A ele e a todos que, nesse período, fizeram parte do Laboratório de Bioinformática Aplicada à Bioenergia, principalmente pela imensa contribuição para meu crescimento pessoal e profissional. Também agradeço, pois neste trabalho foram usados dados de pesquisas realizadas por outros integrantes do Laboratório. Amostras de folha foram coletadas, tiveram o RNA extraído e sequenciado em 2016, pelo Dr. Fernando Henrique Correr. Da mesma forma, os dados de colmo foram coletados em 2016 pelo Dr. Guilherme Kenichi Hosaka. Finalmente, os dados de raízes, coletados em 2017, foram coletados pela Dra. Ana Letycia Basso Garcia. Estes experimentos e a genotipagem do painel foram realizados graças ao trabalho também dos docentes e pesquisadores da UFSCar, principalmente da Profa. Dra. Monalisa Sampaio Carneiro e Dr. Thiago Balsalobre. Para o sequenciamento desses dados, foram utilizadas ferramentas, serviço e a expertise do Laboratório de Biotecnologia Animal do departamento de Zootecnia da ESALQ, liderado pelo Prof. Luiz Lehmann Coutinho. Também contribuíram para a construção dos modelos utilizados neste trabalho pesquisadores da Universidade de Queensland, da Austrália. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. A obtenção dos dados de RNA-Seq foi possível graças ao financiamento da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo nº 2015/22993-7.

SUMÁRIO

RESUMO

Expressão alelo-específica em genótipos, órgãos, tempos e condições de estresse hídrico em cana-de-açúcar

A cana-de-açúcar é uma das principais culturas do estado de São Paulo e do Brasil, fato que motiva estudos visando a compreensão do funcionamento do seu genoma. O conhecimento dos mecanismos moleculares que levam ao aparecimento de determinada característica é de grande valor para futuros esforços de melhoramento genético. Entretanto, este grupo de plantas apresenta uma complexa estrutura genômica, intrínseca ao gênero *Saccharum* e amplificada por anos de melhoramento genético buscando aumentar o teor de sacarose. As cultivares comerciais modernas de cana-de-açúcar são poliploides aneuploides, apresentando de oito a 14 cópias de cada cromossomo, provenientes de cruzamentos interespecíficos entre *S. officinarum* e *S. spontaneum*. Essa complexidade dificulta o estudo e a compreensão dos processos que ocorrem nas células destas plantas, visto que a grande maioria das metodologias utilizadas na área de genômica para obtenção e análise de dados foram desenvolvidas tendo em mente organismos diploides. Um mecanismo pouco compreendido que afeta a expressão gênica da cana-de-açúcar é a expressão diferencial alelo-específica. Por esse mecanismo, os alelos de um loco polimórfico podem ser expressos em taxas diferentes em condições distintas. No caso dos organismos poliploides, os locos heterozigóticos podem apresentar diferentes doses alélicas. Um sítio heterozigótico octaploide pode apresentar qualquer número entre uma e sete cópias de um alelo alternativo, por exemplo. Até o momento, pouco se sabe sobre como a dose alélica variável em organismos poliploides influencia a expressão alelo-específica. Este trabalho teve como objetivo investigar a expressão alelo-específica em dados de colmo, raiz e folhas de cana-de-açúcar. Utilizando uma abordagem Bayesiana, foi modelada a contagem dos alelos expressos de determinado SNV, baseando-se no conhecimento da dose alélica para esse mesmo sítio variante de um único nucleotídeo (SNV, do inglês *Single Nucleotide Variant*), verificando-se a ocorrência de desbalanceamento alélico. Além de representar três órgãos vegetais distintos, o conjunto de dados utilizado englobou híbridos comerciais e representantes de *S. officinarum* e de *S. spontaneum*, agrupados quanto à tolerância à seca, ao teor de sólidos solúveis e ao teor de fibra. Os dados de colmo foram amostrados em diferentes tempos, enquanto que as amostras de raiz foram obtidas de plantas submetidas ou não a estresse hídrico. O genótipo IN84-58 teve folhas e colmos amostrados, enquanto o genótipo SP80-3280 teve amostras de raízes, folhas e colmos. Isso permitiu a comparação entre órgãos desses genótipos. Observou-se que, apesar da presença de genes que variaram quanto à expressão alelo-específica entre órgãos, tempos e condições de estresse hídrico, os genótipos tiveram maior influência sobre este fenômeno, diferenciando principalmente representantes de *S. spontaneum* dos demais.

Palavras-chave: Expressão gênica diferencial, Dose alélica, Poliploidia, *Saccharum*, RNA-seq

ABSTRACT

Allele-specific expression in genotypes, organs, times and water stress conditions in the highly polyploid sugarcane

Sugarcane is one of the main crops in the state of São Paulo and in Brazil, which motivates studies seeking to understand the functioning of its genome. Knowledge about molecular mechanisms that lead to the appearance of specific traits is of great value for future genetic breeding efforts. However, this group of plants has a complex genomic structure, intrinsic to the genus *Saccharum* and further enhanced by years of breeding to increase the sucrose content. Modern commercial sugarcane cultivars are polyploids with frequent aneuploidy, with eight to 14 copies of each chromosome, resulting from interspecific crossings between *S. officinarum* and *S. spontaneum*. This complexity hinders the study and understanding of the processes that take place in the cells of these plants, because the vast majority of techniques used in genetics and genomics were developed with a diploid mindset. One of the poorly understood mechanisms that affects sugarcane gene expression is allele-specific expression. This corresponds to situations where the alleles of a polymorphic locus are expressed at different rates under different conditions. In the case of polyploid organisms, heterozygous loci can present different allele doses. For example, a heterozygous octaploid site can have any number between one and seven copies of one of the alleles. So far, little is known about how the varying dose in polyploid organisms influences allele-specific expression. We aimed to model the occurrence of allele-specific expression in sugarcane stalks, roots and leaves. The expressed allele counts of a given Single Nucleotide Variant (SNV) were modeled using a Bayesian approach based on the knowledge of the allele dose for that same SNV, allowing us to assess the occurrence of allelic imbalance. In addition to the three organs, the dataset comprises commercial hybrids, accessions representing *S. officinarum* and *S. spontaneum*, grouped in terms of drought tolerance, soluble solids and fiber content. The stalk data were sampled at different time points, while root samples were taken from plants subjected or not to water stress. Genotype IN84-58 had leaves and culms sampled, while SP80-3280 had samples of roots, leaves and culms. This allowed comparisons between organs for these genotypes. We noted that, despite observing genes whose presence of allele-specific expression was not consistent between organs, times and water stress conditions, genotypes had a larger influence on the occurrence of this phenomenon, differentiating mainly representatives of *S. spontaneum* from the others.

Keywords: Differential gene expression, Allele dose, Polyploidy, *Saccharum*, RNA-seq

1. INTRODUÇÃO

1.1. Aspectos evolutivos, genômicos e de melhoramento genético da cana-de-açúcar

A cana-de-açúcar (*Saccharum spp.* L.) é uma das culturas mais importantes do estado de São Paulo e do Brasil. Seu colmo é a matéria prima na produção de açúcar e etanol. Além disso, os subprodutos também podem ser aproveitados, seja como fertilizantes, na alimentação animal ou ainda na produção de etanol de segunda geração. Segundo os dados da Conab, na safra de 2020/21 foram colhidos 654,5 milhões de toneladas de cana-de-açúcar e produzidos 41,2 milhões de toneladas de açúcar e 29,7 bilhões de litros de etanol. As estimativas para a safra de 2021/22 são de uma colheita de 568,4 milhões de toneladas de matéria prima (356,7 milhões só na região Sudeste) e produção de 33,9 milhões de toneladas de açúcar e de 24,8 bilhões de litros de etanol (CONAB, 2021).

Apesar da importância desta cultura, o seu estudo em nível molecular é difícil. Isso é devido à sua grande complexidade genômica. As cultivares modernas resultam de décadas de melhoramento, visando aumentar o seu teor de açúcar e combater problemas fitossanitários (DINARDO-MIRANDA; VASCONCELLOS; ANDRADE LANDELL, 2010). Esse melhoramento se deu a partir do cruzamento de espécies poliploides do gênero *Saccharum*, principalmente *S. officinarum* L. (cana-nobre, com alto teor de sacarose) e *S. spontaneum* L. (espécie mais resistente a doenças e a estresses abióticos).

A evolução do próprio gênero *Saccharum* foi marcada por uma série de eventos de autopoliploidização e alopoliploidização (KIM et al., 2014). Com isso, todos esses organismos apresentam uma estrutura genômica poliploide complexa, de difícil compreensão, ainda mais considerando-se que a maioria das tecnologias para estudos de genômica funcional foram desenvolvidas visando organismos diploides. As cultivares modernas de cana-de-açúcar são poliploides ainda mais complexos, tendo sido submetidas a vários anos de melhoramento genético com cruzamentos interespecíficos, apresentando aneuploidia e elementos transponíveis ativos (VILELA et al., 2017; SOUZA et al., 2019).

Indivíduos da espécie *S. officinarum* são octaploides (2n = 8x = 80), enquanto representantes de *S. spontaneum* variam de penta (2n = 5x = 40) a decahexaploide (2n = 16x = 128). As canas-de-açúcar cultivadas atualmente apresentam aproximadamente 120 cromossomos, tendo de oito a 14 cópias de cada representante do conjunto básico. Estudos apontam que 10% a 20% dos cromossomos de representantes de cultivares modernas de cana-de-açúcar são oriundos de *S. spontaneum*, 70% a 80% de *S. officinarum*, e aproximadamente 10% são recombinantes entre os cromossomos das duas espécies (PIPERIDIS; D'HONT, 2020).

1.2. Desafios em estudos de genômica funcional em cana-de-açúcar

Os fatores que permitem a manifestação de determinadas características fenotípicas ou fisiológicas de um organismo precisam não apenas estar presentes no genoma, mas precisam ser expressas. A expressão gênica tem início com a transcrição da informação contida nas fitas de DNA. Em diferentes condições, como células de órgãos distintos, condições ambientais distintas e em diferentes genótipos, espera-se encontrar padrões de expressão gênica distintos (ALBERTS et al., 2017; HARTWELL et al., 2011).

Uma melhor compreensão dos diversos fenômenos que afetam a expressão gênica considerando a complexa estrutura genômica da cana-de-açúcar poderia agilizar os ganhos em programas de melhoramento desta cultura. Uma complicação adicional é que regiões heterozigóticas no genoma de organismos poliploides apresentam

mais configurações alélicas possíveis (doses alélicas) que em diploides. Sendo assim, é mais difícil diferenciar transcritos referentes a genes parálogos, homeólogos, alelos e isoformas nestes organismos.

Devido à alta complexidade, não há ainda uma sequência de referência genômica completa e de alta qualidade para a cana-de-açúcar. Diversos trabalhos têm empregado diferentes estratégias para entender este conjunto de genomas (THIRUGNANASAMBANDAM; HOANG; HENRY, 2018). Algumas dessas iniciativas levaram à recente publicação de sequências genômicas de alguns genótipos.

Primeiro, Garsmeur et al. (2018) construíram uma montagem do genoma monoploide do genótipo R570, híbrido amplamente utilizado para estudos genômicos dessa cultura. Zhang et al. (2018) montaram o genoma do clone haploide AP85-0441 (1n = 4x = 32), obtido a partir do genótipo octaploide SES208, representante de *S. spontaneum*. Usando leituras longas e técnicas de captura de conformação da cromatina, esses autores produziram uma montagem de cromossomos completos (pseudomoléculas) dessa espécie selvagem. Além disso, foram geradas quatro sequências distintas de cada cromossomo, cada uma delas referente a uma cópia alélica específica. Já Souza et al. (2019) apresentaram uma montagem do genoma poliploide do híbrido SP80-3280. Neste último estudo foi possível, pela primeira vez, identificar diferentes haplótipos e suas possíveis regiões regulatórias, isso é, homólogos ou homeólogos de um genótipo híbrido da cana-de-açúcar.

Cada uma destas ferramentas, bem como os novos estudos possibilitados por elas, representa um passo importante para uma compreensão completa do genoma desta cultura. Porém, as sequências genômicas atualmente existentes ainda fornecem uma visão incompleta dos genomas híbridos de maior relevância. Além disso, não é simples modelar adequadamente a dose alélica e o nível de expressão alelo-específica. Estes fenômenos são bastante relevantes em nível celular e têm um papel importante em muitas características fenotípicas, mas dificilmente são estudados (ZHAO et al., 2019).

Além disso, comumente não se dispõe de informação da ploidia nem da dose para cada sítio variante de um único nucleotídeo (SNV, do inglês *Single Nucleotide Variant*) no genoma da cana-de-açúcar. Para contornar isto, modelos estatísticos podem ser empregados visando fornecer estimativas, como é feito no SuperMASSA (SERANG; MOLLINARI; GARCIA, 2012) e no updog (GERARD et al., 2018). Utilizando um modelo estatístico Bayesiano, o SuperMASSA é capaz de estimar a ploidia e a dose alélica mais prováveis em uma determinada região de um genótipo com base nas abundâncias relativas de dois alelos de um SNV (GARCIA et al., 2013).

1.3. Modelagem da expressão alélica diferencial no genoma da cana-de-açúcar

Para um SNV em dada região do genoma poliploide podem existir múltiplas configurações distintas. Por exemplo, se em uma região octaploide existirem duas cópias com um determinado nucleotídeo, enquanto as demais apresentarem um segundo alelo, a dose alélica deste indivíduo para este SNV pode ser denotada por 6:2.

Além da ocorrência de diferentes doses alélicas no genoma, também há o fenômeno da expressão alelo-específica, que pode alterar o número de cópias de determinado alelo presentes no transcriptoma. Com isso, determinado alelo pode ser expresso em detrimento dos demais, tornando a proporção alélica relativa no transcriptoma diferente daquela observada no genoma. Esse desbalanceamento pode se alterar em determinadas condições ou permanecer constante, apesar de outros fatores. Por exemplo, se o produto de um alelo tivesse impacto positivo na retenção de água, a sua expressão seria favorecida em condições de seca, mas não necessariamente em outras condições (ALBERT et al., 2018). Este fenômeno tem especial importância em organismos poliploides, já que o grande número

de cópias cromossômicas aliado à expressão alelo-específica conferiria a eles maior plasticidade para se adaptar a diferentes ambientes (SALMAN-MINKOV; SABATH; MAYROSE, 2016).

Existem diferentes técnicas que permitem analisar o transcriptoma de um organismo, obtendo-se as abundâncias relativas dos alelos. Uma delas é o chamado RNA-Seq (NAGALAKSHMI et al., 2008), que consiste na transcrição reversa do RNA para cDNA (DNA complementar) e o posterior sequenciamento dessas fitas de DNA com métodos de NGS (do inglês *Next Generation Sequencing*). Esta abordagem pode ser aplicada mesmo na ausência de conhecimento genômico prévio (METZKER, 2010), o que é particularmente relevante no caso da cana-de-açúcar, já que seu genoma complexo dificulta a obtenção deste conhecimento.

Rotinas de modelagem estatística apropriadas devem ser empregadas para compreender esta complexa característica da expressão gênica, que sofre influência da dose alélica e de condições externas à célula. Por exemplo, Rao et al. (2021) utilizaram modelos lineares generalizados de efeitos mistos para modelar a interação entre os efeitos fixos de tratamento e do tipo do alelo (alternativo ou referência). Por sua vez, os indivíduos foram modelados como efeitos aleatórios. Esse tipo de estudo baseia-se no paradigma frequentista da estatística, que utiliza testes de hipótese ou intervalos de confiança para fazer inferência sobre uma população a partir de uma amostra. Fazendo isso, considera-se que as estatísticas convergem na direção de um valor real do parâmetro quando o ensaio é repetido com um número grande o suficiente de observações.

Já o paradigma Bayesiano se baseia no Teorema de Bayes de probabilidades condicionais:

$$\Pr(A|B) = \frac{\Pr(A) \times \Pr(B|A)}{\Pr(B)},$$

em que $\Pr(A|B)$ representa a probabilidade condicional de o evento A acontecer dado que o evento B já aconteceu, $\Pr(A)$ é a probabilidade marginal do evento A, $\Pr(B|A)$ indica a probabilidade condicional de B dado A e, por fim, $\Pr(B)$ é a probabilidade marginal de B. Pode-se demonstrar que $\Pr(\boldsymbol{B}) = \int \mathbf{Pr}(\boldsymbol{A}) \times \mathbf{Pr}(\boldsymbol{B}|\boldsymbol{A})\, \boldsymbol{dA}$, de forma que o denominador desta equação funciona como uma constante de normalização do numerador (KRUSCHKE, 2015).

A fórmula do Teorema de Bayes envolve quatro probabilidades: a probabilidade marginal de uma determinada distribuição, chamada probabilidade *a priori* – $\Pr(A)$; a probabilidade de a distribuição hipotetizada ser verdadeira, após a coleta de dados, conhecida como probabilidade *a posteriori* – $\Pr(A|B)$; a probabilidade de se obter os dados tal que eles vieram da distribuição hipotetizada, chamada de verossimilhança – $\Pr(B|A)$; e a probabilidade marginal de se obter os dados, também conhecida como evidência – $\Pr(B)$. O Teorema de Bayes permite calcular a probabilidade *a posteriori* a partir da probabilidade *a priori*, da verossimilhança e da evidência. Além disso, os parâmetros que definem a probabilidade *a priori* podem eles mesmos apresentar suas próprias distribuições de probabilidade *a priori*. Assim, tem-se um modelo hierárquico que fornece a distribuição de probabilidade de um parâmetro para estabelecer a probabilidade *a priori* de outro, mais elevado na hierarquia (KRUSCHKE, 2015).

Como a evidência ($\Pr(\boldsymbol{B}) = \int \mathbf{Pr}(\boldsymbol{A}) \times \mathbf{Pr}(\boldsymbol{B}|\boldsymbol{A})\, \boldsymbol{dA}$) pode resultar em uma integral muito complexa, foram desenvolvidos algoritmos que fornecem uma estimativa da probabilidade *a posteriori* sem a necessidade do cálculo deste termo. Por exemplo, existem os métodos de cadeias de Markov Monte Carlo (MCMC, do inglês *Markov Chain Monte Carlo*) (KRUSCHKE, 2015). O *software* Stan (STAN DEVELOPMENT TEAM, 2019) utiliza um método de amostragem chamado *No-U-Turn Sampling* (NUTS), uma variante do algoritmo de MCMC *Hamiltonian Monte* Carlo (HMC), baseado na dinâmica Hamiltoniana e no princípio de conservação de energia.

Muitos estudos de expressão gênica em nível de alelos modelam os dados usando a distribuição beta-binomial (EDSGÄRD et al., 2016; HARVEY et al., 2015; PICKRELL et al., 2010; RAGHUPATHY et al., 2018; SUN; HU, 2013). A distribuição beta-binomial surge de um modelo hierárquico em que são combinadas as distribuições binomial e beta. Mais especificamente, ela surge quando a probabilidade θ de sucesso da distribuição binomial não é fixa, mas é uma variável aleatória distribuída conforme uma distribuição beta. Essa distribuição possui dois parâmetros, α e β, de modo que quando $\alpha = \beta = 1$, ela toma a forma de uma distribuição uniforme variando de 0 a 1. Conforme aumentam α e β, desde que ambos os parâmetros sejam iguais, a distribuição beta será simétrica, em formato de sino, com variância cada vez menor. No entanto, se $\alpha < \beta$, teremos uma distribuição com assimetria positiva e, se $\alpha > \beta$, uma distribuição com assimetria negativa, de forma que o valor esperado será igual a $\frac{\alpha}{\alpha+\beta}$.

Skelly et al. (2011) propuseram a construção de um modelo Bayesiano hierárquico com três estágios. No primeiro estágio, considerou-se que o número de leituras mapeadas em cada SNV i e gene g (y_{ig}) estava distribuído de acordo com $y_{ig} \sim \text{Binomial}(n_{ig}, \theta_{ig})$, em que n_{ig} é a cobertura total do SNV e θ_{ig} é o nível de desbalanceamento alélico (para um organismo diploide). No segundo estágio, considerou-se que θ_{ig} e ε_{ig} (o parâmetro que quantifica a dispersão) estavam distribuídos cada um conforme uma distribuição beta. Finalmente, o modelo é completado atribuindo distribuições de probabilidade a priori para os parâmetros das distribuições beta. Aplicando este modelo em dados de levedura e de humanos provenientes de diferentes plataformas de sequenciamento, os autores concluíram que o modelo apresentou boa reprodutibilidade e uma alta resolução na identificação de sítios com expressão diferencial alelo-específica.

Xie et al. (2019) desenvolveram um modelo capaz de identificar ao mesmo tempo os genes com expressão alelo-específica e a variação na expressão diferencial de alelos entre SNVs de um mesmo gene. Este modelo considera que $y_{gij} \sim \text{Binomial}(n_{gij}, \theta_{gij})$, em que y_{gij} é a contagem de transcritos provenientes do alelo de referência, n_{gij} é a cobertura total e θ_{gij} é o desbalanceamento alélico, para um SNV i, gene g e réplica j. Então, são definidos os efeitos fixos β_g para genes e os efeitos aleatórios S_{gi} para SNVs e R_{gj} para réplicas, em que $S_{gi} \sim \text{N}(0, \sigma^2_{sg})$ e $R_{gj} \sim \text{N}(0, \sigma^2_{rg})$, de modo que $\log \frac{\theta_{gij}}{1-\theta_{gij}} = \beta_g + S_{gi} + R_{gj}$. Por fim, são atribuídas as probabilidades a priori para os parâmetros β_g, σ^2_{sg} e σ^2_{rg}, de modo que $\beta_g \sim \text{N}(0, \sigma^2)$; ${}^{1}/_{\sigma^2_{sg}} \sim \text{Gamma}(a_s, b_s)$ e ${}^{1}/_{\sigma^2_{rg}} \sim \text{Gamma}(a_r, b_r)$. Com este modelo, é possível testar $H_0: \beta_g = 0$ (se o gene g apresenta desbalanceamento alélico) e $H_0: \sigma^2_{sg} = 0$ (se o nível de desbalanceamento é uniforme entre os SNVs de um mesmo gene g). Os autores demonstraram a aplicabilidade deste modelo em dados reais de bovinos e em dados simulados, concluindo que ele possui maior poder e acurácia que outros métodos existentes.

Em cana-de-açúcar, existem trabalhos com o objetivo de investigar a expressão alelo-específica, mas com foco em genes específicos ou famílias de genes. Por exemplo,Vilela et al. (2017) sequenciaram BACs (*Bacterial Artificial Chromosome*) de cana de açúcar para investigar as duplicações do genoma que ocorreram durante a evolução do gênero *Saccharum*. Para isso, foram analisados haplótipos de R570 dos genes LFY (*Leafy*), PHYC (*Phytochrome-C*) e TOR (*Target of Rapamycin*), essenciais pelo desenvolvimento da planta e componentes de redes complexas de regulação gênica na célula. Esses dados foram contrastados com dados de expressão gênica dos genes PHYC e TOR, revelando à primeira vista uma alta correlação entre a proporção no genoma e o nível de expressão de SNVs presentes em haplótipos do gene TOR. Entretanto, uma avaliação mais minuciosa mostrou que haplótipos diferentes deste gene apresentavam padrões de expressão gênica específicos, sem apresentar a transcrição de determinados alelos dos SNVs. Além disso,

não houve correlação entre a dose alélica e a expressão gênica dos SNVs presentes em haplótipos de PHYC. Já Sforça et al. (2019) deram atenção aos genes CENP-C (*Centromere Protein C*) e HP600 de SP80-3280, visando estudar a complexidade genômica de um conjunto de cultivares modernas de cana-de-açúcar. Comparando as dosagens genômicas com a proporção dos alelos expressos, concluiu-se que somente a proporção no genoma não é capaz de explicar as razões observadas do gene HP600 no transcriptoma. Cai et al. (2020) também estudaram a expressão alelo-específica em um conjunto de genes do genótipo SES208 de *S. spontaneum*. Foram estudados 29 fatores de transcrição SsDof (*DNA-binding with one finger*), envolvidos no desenvolvimento das plantas e na resposta a hormônios e ao estresse abiótico. Observando-se a expressão dos alelos individuais, demonstrou-se que há expressão alelo-específica nesses genes e que ela responde a estímulos, como a presença de hormônios. Isso sugere que a variação na proporção expressa de cada alelo está associada a diferentes papéis que cada um dos alelos pode desempenhar na célula.

Um modelo Bayesiano hierárquico foi desenvolvido por Correr et al. (2022), visando verificar se a proporção relativa da expressão de um alelo apresenta algum desbalanceamento com relação à dose estimada para esse mesmo alelo. Visto que este modelo considera a ploidia de um sítio específico, ele é apropriado no contexto de organismos poliploides.

Esse modelo assume que o número de leituras do RNA-Seq do alelo de referência para cada SNV, réplica e genótipo é distribuído conforme uma distribuição binomial. Por sua vez, considera-se que a proporção esperada do alelo de referência em um determinado SNV para dado genótipo (θ) segue uma distribuição beta com parâmetros α igual à dose do alelo de referência e β igual à dose do alelo alternativo. Assim, diferentes réplicas de um mesmo genótipo compartilham esta informação para cada SNV.

Combinando informações provenientes de dados do genoma e da expressão gênica, foi obtido um intervalo de máxima densidade (HDI, do inglês *Highest Density Interval*) para θ, em que os valores que este parâmetro pode assumir apresentam maior densidade de probabilidade. A partir das estimativas de α e β obtidas com base nos dados genômicos, calcula-se o valor de $\frac{\alpha}{\alpha+\beta}$. Assim, é possível verificar se este valor se encontra dentro do HDI obtido. Caso contrário, há uma indicação de que os dados de expressão gênica diferem significativamente da proporção esperada com base nas dosagens genômicas, sugerindo a ocorrência de expressão diferencial alelo-específica.

Correr et al. (2022) buscaram uma abordagem global, analisando a expressão alelo-específica comparando dados de GBS e de RNA-Seq entre folhas das cultivares IN84-58, RB72454, SES205A, SP80- 3280, US85-1008 e White Transparent, contrastantes quanto ao teor de fibra. Observou-se que os genótipos híbridos interespecíficos (RB72454, SP80-3280 e US85-1008) apresentaram um maior número de SNVs com expressão diferencial alelo-específica. Além disso, os genes contendo SNVs que apresentaram expressão alélica diferencial estavam relacionados principalmente a processos gerais do metabolismo, à composição de organelas e à resposta ao estresse e a outros estímulos. Foram destacados genes associados à resistência a doenças (*enhanced disease resistance 2*), ao metabolismo de carbono (*UTP-glucose-1-phosphate uridylytransferase*, *Sucrose-phosphate synthase*, *Sucrose transport protein SUT4*) e ao processo de fotossíntese (*RuBisCO large subunit-binding protein* e *phosphoenolpyruvate carboxylase 3*) que apresentaram expressão diferencial de alelos. Entretanto, não foram encontrados termos GO enriquecidos com genes diferencialmente expressos, provavelmente pela quantidade relativamente baixa de sítios variantes que passaram pelos filtros aplicados. Esses resultados sugerem que, em cana-de-açúcar, há um excesso de SNVs com expressão alélica diferencial associada a genótipos oriundos de cruzamentos interespecíficos quando comparados aos selvagens.

Margarido et al. (2022) também fizeram uma análise da expressão alelo-específica em escala genômica de genótipos de cana-de-açúcar, comparando dados de WGS *(Whole Genome Sequencing)* e de RNA-Seq. Foram selecionados

sete genótipos híbridos: KQ228, SRA5, Q155, KQB09-20432, SRA1, MQ239 e Q186, contrastantes quanto à produtividade, teor de fibra e teor de sólidos solúveis. Os termos GO enriquecidos contendo genes com expressão alelo-específica incluíram resposta de defesa (GO:0006952), atividade UDP-glycosyltransferase (GO:0008194), associada à biossíntese da parede celular, atividade sulfotransferase (GO:0008146) e sulfatação (GO:0051923). Foram encontrados genes cuja expressão alelo-específica pode estar relacionada às diferenças fenotípicas dos genótipos, como genes associados à formação da parede celular no genótipo rico em fibras SRA5, mas não no genótipo rico em açúcares KQ228.

Este trabalho teve como objetivo estudar, em nível genômico, a ocorrência da expressão alelo-específica em cana-de-açúcar em diferentes genótipos, em células de diferentes órgãos, em diferentes pontos no tempo e em diferentes condições de estresse hídrico. Para determinar a ocorrência da expressão diferencial de alelos, foi utilizado o modelo Bayesiano hierárquico desenvolvido por Correr et al. (2022). Foram contrastados os dados genômicos, obtidos pela técnica de GBS, e os de transcriptoma, obtidos por RNA-Seq. Com isso, foi possível ter uma visão geral de como esse fenômeno se comporta no genoma complexo da cana-de-açúcar, em quais genes ele se manifesta e quais condições levam à sua ocorrência.

Referências

ALBERT, E. et al. Allele-specific expression and genetic determinants of transcriptomic variations in response to mild water deficit in tomato. **The Plant Journal**, v. 96, n. 3, p. 635–650, nov. 2018.

ALBERTS, B. et al. **Fundamentos da Biologia Celular**. 4. ed. Porto Alegre: Artmed, 2017.

CAI, M. et al. Allele specific expression of Dof genes responding to hormones and abiotic stresses in sugarcane. **PLOS ONE**, v. 15, n. 1, p. e0227716, 16 jan. 2020.

CONAB. **Acompanhamento da Safra Brasileira (Cana-de-Açúcar - Safra 2021/22 3° Levantamento)**. Brasília: [s.n.].

CORRER, F. H. et al. Allele expression biases in mixed-ploid sugarcane accessions. **Scientific Reports**, 2022.

DINARDO-MIRANDA, L. L.; VASCONCELLOS, A. C. M. DE; ANDRADE LANDELL, M. G. DE. **Cana-de-Açúcar**. Campinas: Instituto Agronômico, 2010.

EDSGÄRD, D. et al. GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. **Scientific Reports**, v. 6, n. 1, p. 21134, 18 fev. 2016.

GARCIA, A. A. F. et al. SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. **Scientific Reports**, v. 3, n. 1, p. 3399, 2 dez. 2013.

GARSMEUR, O. et al. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. **Nature Communications**, v. 9, n. 1, p. 2638, 6 dez. 2018.

GERARD, D. et al. Genotyping Polyploids from Messy Sequencing Data. **Genetics**, v. 210, n. 3, p. 789–807, 1 nov. 2018.

HARTWELL, L. H. et al. **Genetics: From Genes to Genomes**. 4. ed. Nova Iorque: McGraw-Hill, 2011.

HARVEY, C. T. et al. QuASAR: quantitative allele-specific analysis of reads. **Bioinformatics**, v. 31, n. 8, p. 1235–1242, 15 abr. 2015.

KIM, C. et al. Comparative Analysis of Miscanthus and Saccharum Reveals a Shared Whole-Genome Duplication but Different Evolutionary Fates. **The Plant Cell**, v. 26, n. 6, p. 2420–2429, jun. 2014.

KRUSCHKE, J. **Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan**. Second edition ed. [s.l.] Elsevier Science, 2015.

MARGARIDO, G. R. A. et al. Limited allele-specific gene expression in highly polyploid sugarcane. **Genome Research**, v. 32, n. 2, p. 297–308, fev. 2022.

METZKER, M. L. Sequencing technologies — the next generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31–46, 8 jan. 2010.

NAGALAKSHMI, U. et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. **Science**, v. 320, n. 5881, p. 1344–1349, 6 jun. 2008.

PICKRELL, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. **Nature**, v. 464, n. 7289, p. 768–772, 10 abr. 2010.

PIPERIDIS, N.; D'HONT, A. Sugarcane genome architecture decrypted with chromosome-specific oligo probes. **The Plant Journal**, v. 103, n. 6, p. 2039–2051, 12 set. 2020.

RAGHUPATHY, N. et al. Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. **Bioinformatics**, v. 34, n. 13, p. 2177–2184, 1 jul. 2018.

RAO, X. et al. Allele-specific expression and high-throughput reporter assay reveal functional genetic variants associated with alcohol use disorders. **Molecular Psychiatry**, v. 26, n. 4, p. 1142–1151, 2 abr. 2021.

SALMAN-MINKOV, A.; SABATH, N.; MAYROSE, I. Whole-genome duplication as a key factor in crop domestication. **Nature Plants**, v. 2, n. 8, p. 16115, 1 ago. 2016.

SERANG, O.; MOLLINARI, M.; GARCIA, A. A. F. Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. **PLoS ONE**, v. 7, n. 2, p. e30906, 17 fev. 2012.

SFORÇA, D. A. et al. Gene Duplication in the Sugarcane Genome: A Case Study of Allele Interactions and Evolutionary Patterns in Two Genic Regions. **Frontiers in Plant Science**, v. 10, 7 maio 2019.

SKELLY, D. A. et al. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. **Genome Research**, v. 21, n. 10, p. 1728–1737, out. 2011.

SOUZA, G. M. et al. Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. **GigaScience**, v. 8, n. 12, 1 dez. 2019.

STAN DEVELOPMENT TEAM. **Stan Modeling Language Users Guide and Reference Manual**. , 2019.

SUN, W.; HU, Y. eQTL Mapping Using RNA-seq Data. **Statistics in Biosciences**, v. 5, n. 1, p. 198–219, 9 maio 2013.

THIRUGNANASAMBANDAM, P. P.; HOANG, N. V.; HENRY, R. J. The Challenge of Analyzing the Sugarcane Genome. **Frontiers in Plant Science**, v. 9, 14 maio 2018.

VILELA, M. DE M. et al. Analysis of three sugarcane homo/homeologous regions suggests independent polyploidization events of *Saccharum officinarum* and *Saccharum spontaneum*. **Genome Biology and Evolution**, p. evw293, 12 jan. 2017.

XIE, J. et al. Modeling allele-specific expression at the gene and SNP levels simultaneously by a Bayesian logistic mixed regression model. **BMC Bioinformatics**, v. 20, n. 1, p. 530, 28 dez. 2019.

ZHANG, J. et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. **Nature Genetics**, v. 50, n. 11, p. 1565–1573, 8 nov. 2018.

ZHAO, C. et al. Quantification of allelic differential expression using a simple Fluorescence primer PCR-RFLP-based method. **Scientific Reports**, v. 9, n. 1, p. 6334, 19 dez. 2019.

2. A EXPRESSÃO ALELO-ESPECÍFICA NO GENOMA POLIPLOIDE DE CANA-DE-AÇÚCAR É INFLUENCIADA PELO GENÓTIPO, ÓRGÃO, TEMPO E CONDIÇÃO DE ESTRESSE HÍDRICO

Resumo

A cana-de-açúcar apresenta diversas camadas de complexidade genômica, reunindo poliploidia, aneuploidia, hibridação interespecífica, elevada heterozigosidade e elementos transponíveis ativos. Esses fenômenos dificultam a aplicação de ferramentas para seu estudo visando o melhoramento genético. Diversos estudos avaliam e quantificam a expressão alelo-específica (EA), mas poucos buscam estudá-la em poliploides complexos, como a cana-de-açúcar. Desses, poucos apresentam uma abordagem global ao longo do genoma, focando em vez disso em pequenos grupos de genes. Esse trabalho busca avaliar em nível genômico o fenômeno da EA na cana-de-açúcar, comparando diferentes genótipos, órgãos, pontos no tempo e em condição de estresse hídrico. Foram encontrados genes que mostram diferenças quanto à EA em todos os fatores estudados. Foi demonstrado que, além desse fenômeno ser pouco frequente no genoma de cana-de-açúcar, ele é pouco consistente em diferentes órgãos, ao longo do tempo e em condições diferentes. Além disso, o genótipo foi o principal fator que diferenciou as amostras quanto à EA, corroborando observações de que esse fenômeno é herdável. Em primeiro lugar, foi possível separar genótipos representantes de *S. spontaneum* dos demais. Isso revela uma diferença na EA entre genótipos selvagens com baixo teor de açúcar em relação aos demais, a qual é menos pronunciada entre genótipos híbridos e *S. officinarum*.

Abstract

Sugarcane presents several layers of genomic complexity, including polyploidy, aneuploidy, interspecific hybridization, high heterozygosity and active transposable elements. These phenomena hinder the application of tools for its study and breeding. Several studies have assessed allele-specific expression (ASE), but few involve complex polyploids, such as sugarcane. Among the studies in sugarcane, few aimed at a genome-wide approach, focusing instead on small groups of genes. This work seeks to evaluate the phenomenon of ASE in sugarcane at a genomic level, comparing different genotypes, organs, points in time and under water stress conditions. Genes were found to have differences in ASE for all factors studied. We showed that, in addition to the low occurrence of this phenomenon in the sugarcane genome, ASE is not very consistent in different organs, over time and under different water stress conditions. In addition, more diferences in ASE were observed when comparing the different genotypes than in other comparisons, corroborating observations that this phenomenon is heritable. First, it was possible to separate genotypes representing *S. spontaneum* from the others. This shows that there is a difference in ASE between wild-type low-sugar genotypes compared to others, but less between hybrid genotypes and *S. officinarum*.

2.1 Introdução

A cana-de-açúcar é uma importante cultura, matéria prima para a produção de açúcar e etanol, essencial na produção de alimento e de energia. Estima-se que em 2020 houve uma colheita de 1,87 bilhão de toneladas de cana de açúcar em 26,5 milhões de ha plantados em todo o mundo. Isso corresponde à maior produção dentre todos os 162 itens presentes no banco de dados e à segunda maior produtividade, atrás somente de cogumelos e trufas (Figura 1). Além disso, subprodutos como a vinhaça, o bagaço e folhas que restam no campo após encaminhados os colmos para

a moagem também podem ser aproveitados, seja no campo, como fertilizantes, na alimentação animal ou ainda na produção de etanol de segunda geração.

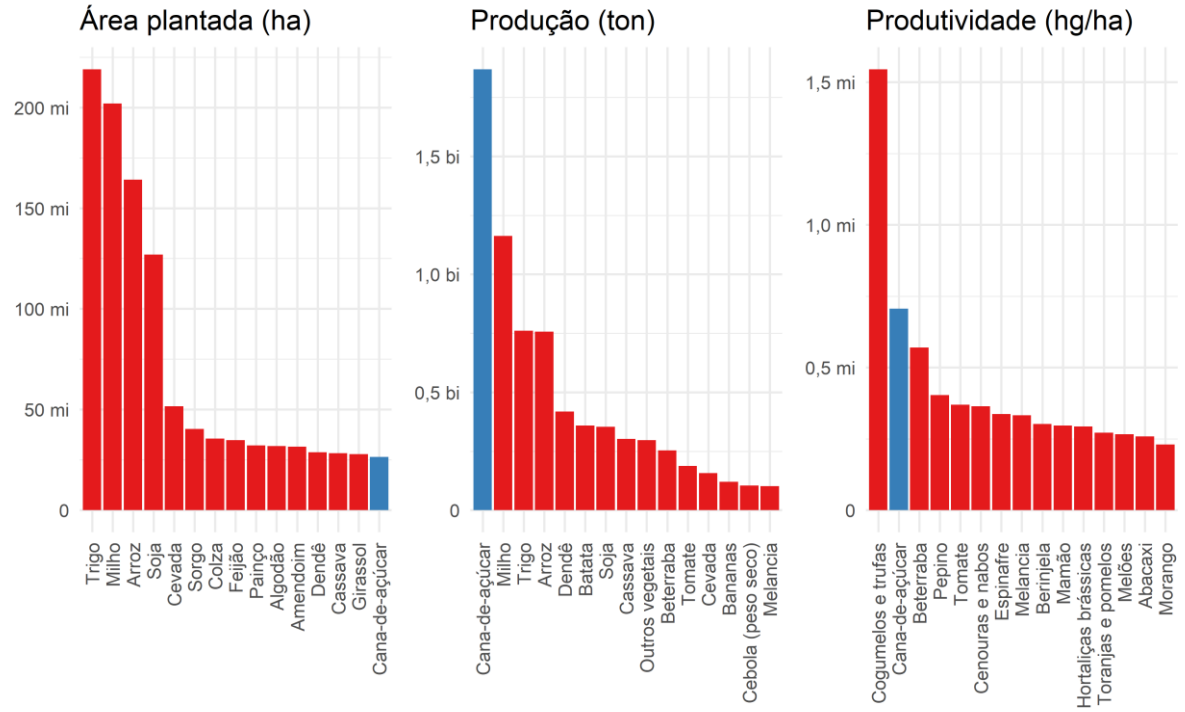

Figura 1. Quinze culturas com maiores valores de área plantada (ha), produção (ton) e produtividade (hg/ha) (FAOSTAT, 2022).

O grupo de organismos denominados canas-de-açúcar é formado por poliploides complexos. Houve diversos eventos de poliploidização ao longo da evolução do gênero *Saccharum* (KIM et al., 2014) de modo que indivíduos de *S. spontaneum* variam de penta (2n = 5x = 40) a decahexaploide (2n = 16x = 128) e *S. officinarum* são octaploides (2n = 8x = 80). Essas duas espécies foram primordialmente utilizadas em cruzamentos para gerar as cultivares modernas, que apresentam uma complexidade genômica ainda maior. Para aumentar os teores de açúcar, foram realizados retrocruzamentos com *S. officinarum*. Indivíduos representantes das cultivares de cana-de-açúcar cultivadas atualmente apresentam aproximadamente 120 cromossomos, tendo de oito a 14 cópias de cada representante do conjunto básico. De 10% a 20% dos cromossomos destas plantas são oriundos de *S. spontaneum*, 70% a 80% de *S. officinarum* e aproximadamente 10% são recombinantes (PIPERIDIS; D'HONT, 2020).

A cana-de-açúcar apresenta uma complexa arquitetura genômica, sendo poliploide aneuploide, inclusive com elementos transponíveis ativos (SOUZA et al., 2019; VILELA et al., 2017). Essa estrutura complexa traz consequências para a expressão gênica, principalmente a expressão alelo-específica. A expressão gênica é um fenômeno importante já que, para exercer algum papel na célula, a informação presente no genoma precisa ser convertida em um determinado produto gênico. A função deste produto gênico depende, em um primeiro nível, da sequência gênica do alelo expresso. Portanto, se mais de um alelo está presente, a determinação de qual deles será expresso é essencial para garantir que o produto gênico exerça determinada função. Isto é particularmente importante nos casos em que essa função é alterada drasticamente devido a sítios variantes entre as sequências dos diferentes alelos (ZHAO et al., 2019).

As regiões heterozigóticas no genoma de organismos poliploides apresentam mais configurações genotípicas possíveis (doses alélicas) que em diploides. No entanto, apesar de avanços recentes, ainda não há uma sequência de referência genômica completa e de alta qualidade para a cana-de-açúcar (GARSMEUR et al., 2018; SOUZA et al., 2019; ZHANG et al., 2018). Sendo assim, é difícil determinar com exatidão as doses alélicas em um sítio específico no momento da chamada dos SNVs. O SuperMASSA é um *software* capaz de estimar a ploidia e a dose alélica em SNVs de organismos autopoliploides (SERANG; MOLLINARI; GARCIA, 2012). Devido à complexidade do genoma da cana-de-açúcar, também é difícil diferenciar transcritos oriundos de genes parálogos, homeólogos, alelos e isoformas.

A presença de múltiplas cópias cromossômicas pode conferir aos organismos poliploides uma maior plasticidade, permitindo que se adaptem a diferentes condições ambientais ao expressar alelos distintos. Diversos estudos investigaram a ocorrência deste fenômeno em conjuntos de genes específicos (CAI et al., 2020; SFORÇA et al., 2019; VILELA et al., 2017), mas poucos buscaram observá-lo em todo o genoma (CORRER et al., 2022; MARGARIDO et al., 2022). A compreensão de como este mecanismo se comporta em escala genômica, em quais genes e sob quais condições se manifesta, é importante para os programas de melhoramento genético dessa cultura.

2.2 Material e Métodos

2.2.1 Material biológico

Os dados de RNA-Seq utilizados provêm de três experimentos com cana-de-açúcar. O primeiro deles utilizou dados de folhas (CORRER et al., 2020b) e analisou triplicatas dos genótipos SES205A e IN84-58, ambos representantes de *S. spontaneum*, e US85-1008, híbrido comercial com alto teor de fibra; RB72454 e SP80-3280, híbridos comerciais com teor de sólidos solúveis elevado, e White Transparent, representante de *S. officinarum*. Outro estudo avaliou dados de sequenciamento de colmos (HOSAKA et al., 2021) em quatro tempos: abril, junho, agosto e outubro de 2016. Para isto, foram utilizados os seguintes genótipos (em ordem crescente de teor de sólidos solúveis): IN84-58, F36-819, R570 e SP80-3280. Por fim, um experimento com raízes estudou plantas em condições de estresse hídrico. Foram utilizados os genótipos tolerantes à seca SP79-1011 e RB867515, os genótipos RB92579 e SP80-3280, sensíveis à seca e os genótipos RB855113 e RB855536, de comportamento intermediário. Nota-se que há dados de folhas e de colmos para o genótipo IN84-58, enquanto o genótipo SP80-3280 esteve presente nos três experimentos. Um resumo das características dos genótipos utilizados pode ser encontrado na Tabela 1.

Tabela 1. Características dos genótipos utilizados neste estudo.

Genótipo	Genoma	Classificação	Órgão	Níveis
SP79-1011	Híbrido	Tolerante à seca	Raiz	2 (seca e controle)
RB867515	Híbrido			
RB855113	Híbrido	Intermediária		
RB855536	Híbrido			
RB92579	Híbrido	Sensível à seca		
SP80-3280	Híbrido			
SP80-3280	Híbrido	Brix muito alto	Colmo	4 tempos
R570	Híbrido	Brix alto		
F36-819	Híbrido	Brix baixo		
IN84-58	*S. spontaneum*	Brix muito baixo		
IN84-58	*S. spontaneum*	Fibra alta	Folha	–
SES205A	*S. spontaneum*			
US85-1008	Híbrido			
RB72454	Híbrido	Fibra baixa		
SP80-3280	Híbrido			
White transparent	*S. officinarum*			

Esses genótipos fazem parte do Painel Brasileiro de Genótipos de Cana-de-Açúcar (PBGCA), localizado no Centro de Ciências Agrárias da Universidade Federal de São Carlos (UFSCar), Araras, São Paulo, e delineado pelos pesquisadores do Programa de Melhoramento Genético da Cana-de-Açúcar (PMGCA). Este painel contém 266 genótipos, incluindo espécies ancestrais do gênero *Saccharum*, genitores dos programas brasileiros de melhoramento, cultivares de importância histórica e cultivares comerciais modernas. Os genótipos do PBGCA foram plantados em campo, seguindo um delineamento em blocos ao acaso com quatro repetições. Além disso, para o experimento de resposta à seca foi instalado um experimento em casa de vegetação de acordo com um delineamento inteiramente ao acaso.

2.2.2 Extração do RNA e sequenciamento

Foram coletadas as raízes, colmos imaturos (+1) e folhas +1. As amostras foram imediatamente congeladas em nitrogênio líquido e armazenadas em *freezer* a -80°C. O RNA total das amostras foi extraído utilizando o *RNeasy Plant Mini Kit* (Qiagen). Antes do sequenciamento, foram realizados alguns passos para verificar a qualidade e a concentração de RNA, pela visualização em gel de agarose e avaliação da integridade (RIN, do inglês *RNA integrity number*) em equipamento Bioanalyzer (Agilent). Foram selecionadas amostras com valores de RIN acima de 8,0 para o preparo das bibliotecas. Por fim, as bibliotecas de sequenciamento foram preparadas segundo o protocolo do *kit Illumina TruSeq Stranded mRNA* e sequenciadas em plataforma Illumina HiSeq 2500, obtendo-se leituras pareadas de 100 pares de base. Foram sequenciadas três réplicas biológicas e duas réplicas técnicas de folhas de cada genótipo. Para o sequenciamento, foram utilizadas duas *lanes*, cada uma contendo uma amostra de cada réplica biológica de cada genótipo. As 48 amostras de colmo (4 genótipos x 4 tempos x 3 réplicas) foram sequenciadas em seis *lanes*, o que corresponde a uma cobertura final de oito amostras por *lane*. As bibliotecas foram distribuídas de forma que em cada

lane fossem sequenciadas 24 amostras diferentes, de modo que foram obtidas três réplicas técnicas para cada réplica biológica de cada genótipo e de cada tempo. Por fim, as raízes foram sequenciadas em três *lanes*, sendo que cada *lane* recebeu 12 das 36 amostras (6 genótipos x 2 tratamentos x 3 réplicas), resultando em uma réplica técnica de cada. Obteve-se uma média de 68,6 milhões de leituras por amostra de colmos, 43,4 milhões de leituras por amostra de folhas e 39 milhões de leituras por amostra de raízes. Foi feita a soma das contagens de leituras das réplicas técnicas para cada SNV de cada amostra e, na modelagem, foram consideradas somente as três réplicas biológicas.

2.2.3 Processamento dos dados

Para controle de qualidade e pré-processamento, foram utilizados o *software* FastQC (v0.11.9)(ANDREWS et al., 2010) e o Trimmomatic (v0.39) (BOLGER; LOHSE; USADEL, 2014), removendo-se bases com escore de qualidade Phred inferior a 20 em uma janela de quatro pares de bases e leituras com menos de 75 pb. Além disso, foram filtradas as bases das pontas das leituras até que apresentassem um escore mínimo de três. Por fim, as primeiras 12 bases de cada leitura também foram removidas, independentemente dos escores de qualidade.

Utilizando-se o *software* Trinity (v2.11.0) (GRABHERR et al., 2011), foi montado um transcriptoma *de novo* com os dados dos três órgãos, de todos os genótipos e todos os tratamentos. Para isso, foi utilizado o tamanho de *kmer* padrão igual a 25 e um tamanho mínimo de *contig* (*--min_contig_length*) igual a 300. Estes valores foram definidos em análises anteriores para minimizar a fragmentação de transcritos (CORRER et al., 2020a). Foi selecionada apenas a isoforma mais longa de cada transcrito montado, a fim de evitar sequências redundantes (GRABHERR et al., 2011). Este transcriptoma foi utilizado como referência para o alinhamento das leituras, usando o *software* bowtie2 (v2.3.5.1) (LANGMEAD; SALZBERG, 2012), com a opção *--very-sensitive*. Também foram modificadas a função de escore mínimo (*--score-min L,0.0,-0.8*) e a penalidade para indels (*--RDG 4,1 --RFG 4,1*), e foram utilizados fragmentos entre 50 e 800 pares de base. Estes valores também foram definidos conforme resultados de estudos anteriores (MARGARIDO et al., 2022).

Os genótipos do PBGCA tinham sido previamente genotipados com a técnica de genotipagem por sequenciamento (GBS). Durante o protocolo do GBS, as amostras de SES205A e US85-1008 eram as únicas sem réplicas nas bibliotecas de sequenciamento. Por outro lado, as amostras de SP80-3280, RB855536 e RB92579 apresentaram mais réplicas. Foi feita a descoberta de SNVs para o conjunto de dados de GBS com todos os genótipos do painel utilizando o mesmo transcriptoma montado *de novo* como sequência de referência e o *software* Tassel4-Poly. Foi usado um valor mínimo da MAF *(minor allele frequency)* igual a 0,01 e um valor mínimo de MAC *(minor allele count*) igual a 40. Para estimar as dosagens gênicas para a região de cada SNV, foram utilizados o SuperMASSA (SERANG; MOLLINARI; GARCIA, 2012) e o VCF2SM (PEREIRA; GARCIA; MARGARIDO, 2018). Foi usado um modelo de inferência de Hardy-Weinberg com taxa mínima de chamada de 50%, o *naive posterior threshold* utilizado foi igual a 0,5, e a probabilidade mínima *a posteriori* para que uma variante fosse mantida foi igual a 0,5. Foram mantidos no conjunto de dados os sítios que apresentavam valores estimados para a ploidia entre 6 e 14. Estes parâmetros também foram ajustados em estudos anteriores (CORRER et al., 2022). Por fim, a quantificação da expressão gênica foi feita para os dados de RNA-Seq usando o módulo ASEReadCounter, do GATK (v4.2.0.0) (VAN DER AUWERA; O'CONNOR, 2020). Nesse estudo, foi considerado como alelo de referência o alelo presente na sequência de referência (a montagem *de novo* do transcriptoma), sendo o outro o alelo alternativo.

A anotação funcional foi feita com base no transcriptoma montado *de novo* e o pipeline do *software* Trinotate (v3.2.1) (BRYANT et al., 2017), com auxílio dos *softwares* TransDecoder (v5.5.0), BLASTp e BLASTx do NCBI, HMMER (v3.3.1), RNAMMER (v1.2), signalP (v4.1) e tmhmm (v2.0c). Foi utilizado o pacote *goseq* (v1.44) (YOUNG et al., 2010), do *software* R (v4.1.3) (R CORE TEAM, 2022) para verificar se havia associação da ocorrência de expressão alelo-específica com determinadas categorias de GO (*Gene Ontology*). Para essa análise, foi usado um nível de significância igual a 0,05 e os p-valores foram ajustados para múltiplos testes pelo método de Benjamini-Hochberg (BENJAMINI; HOCHBERG, 1995), considerando o total de termos GO testados.

2.2.4 Análise estatística

Foram observados os agrupamentos das amostras quanto à proporção relativa do alelo de referência (proporção nos dados de GBS) e à expressão relativa do alelo de referência (proporção nos dados de RNA-Seq). Para isso, foram feitos agrupamentos hierárquicos utilizando-se as distâncias euclidianas e o método de Ward.

Foi usado um modelo Bayesiano hierárquico (CORRER et al., 2022) para buscar sítios variantes com expressão alelo-específica. Esse modelo considera uma distribuição binomial para as leituras do alelo de referência y_{kijr}, referente ao k-ésimo genótipo, i-ésimo SNV, j-ésima categoria do fator e r-ésima repetição. Os parâmetros desta distribuição são o número total de leituras, n_{kijr}, e a proporção da expressão dos alelos de referência, θ_{kij}.

O parâmetro θ_{kij} tem como distribuição *a priori* uma beta, com parâmetros α_{ki} e β_{ki}, que correspondem diretamente às dosagens do alelo de referência e do alelo alternativo, respectivamente, valores estimados pelo SuperMASSA de acordo com os dados de GBS. Nesse caso, as leituras para um mesmo SNV de amostras provenientes de um mesmo genótipo compartilham a distribuição beta da qual é amostrada o parâmetro θ_{kij}. As análises foram realizadas considerando um nível de significância de 5% e corrigindo os p-valores pelo método de Bonferroni, a partir de uma aproximação para o número de testes independentes realizados.

2.3 Resultados e Discussão

2.3.1 Análises descritivas dos dados

Foram identificados na montagem *de novo* do transcriptoma 424.296 genes e 984.181 transcritos. Considerando apenas a isoforma mais longa, os contigs apresentaram um N50 de 910 pb, uma mediana de comprimento de 484 pb e uma média de 754 pb. Dos 425 grupos BUSCO (*Benchmarking Universal Single-Copy Orthologs*) analisados (MANNI et al., 2021), 330 (77,6%) estavam completos no conjunto de dados, dos quais 313 apresentavam cópia única. Dos 425 BUSCO, 70 (16,5%) estavam fragmentados e 25 (5,9%) não estavam presentes. As taxas de alinhamento variaram de 88,8% a 95,2%, com uma média de 94,28% e mediana de 94,51%.

Foram identificados 72.231 sítios polimórficos nos dados genômicos. Destes, 47.475 também foram verificados nos dados de RNA-Seq, ou seja, estavam presentes em genes efetivamente expressos. Após um filtro de densidade máxima igual a 10 sítios variantes em uma janela de 60 pares de bases, profundidade de cobertura igual ou maior que 5 nos dados de RNA-Seq e igual ou maior que 10 nos dados de GBS, restaram 25.830 SNVs distribuídos em 5.350 transcritos. A maioria dos SNVs encontrados tiveram a ploidia estimada igual a 12 (Figura 2). Na sequência,

foram observadas ploidias de seis, 10 e 14, sendo que a ploidia menos observada nestes SNVs foi igual a oito. Como apontado por Correr et al. (2022) nos dados de folha, essas observações estão de acordo com estudos citológicos em cana-de-açúcar (PIPERIDIS; D'HONT, 2020).

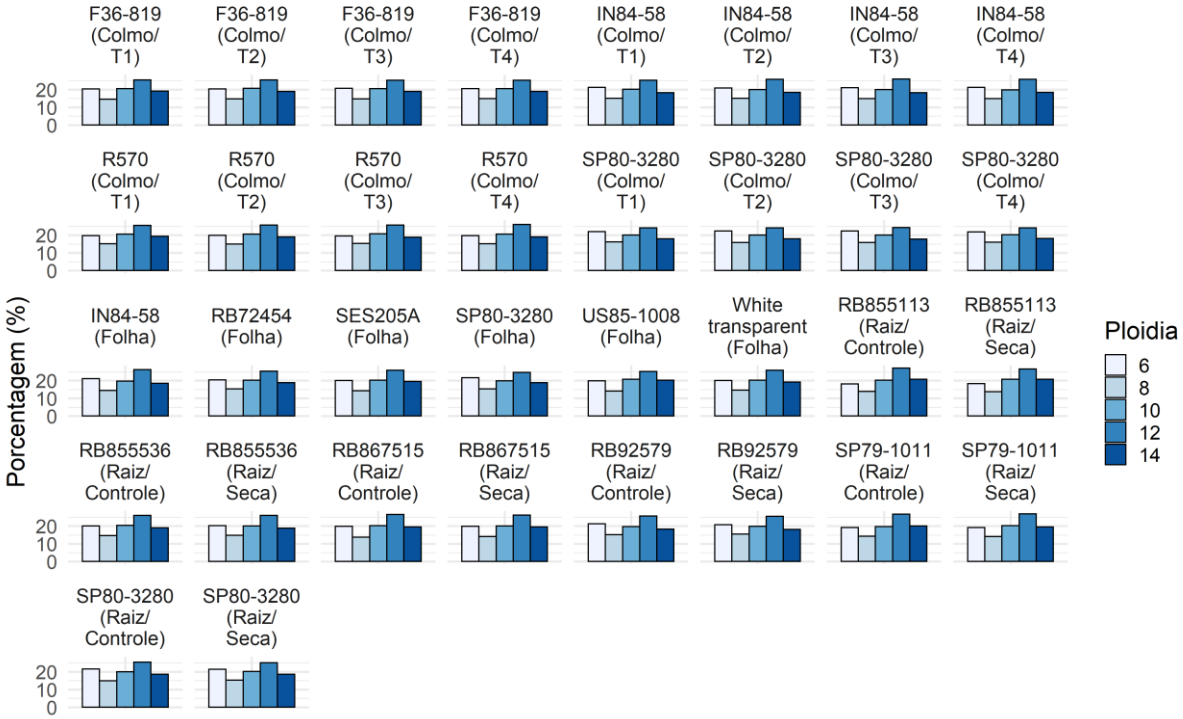

Figura 2. Proporção das ploidias para cada órgão, genótipo e categoria (tempo em colmos ou condição de estresse hídrico em raízes).

Nota-se que as amostras de colmo apresentaram maior número de sítios variantes após a filtragem do que raízes e folhas (Figura 3). A maior profundidade de cobertura obtida em dados de colmo resultou em um maior número de sítios passando pelos filtros, o que explica a diferença entre órgãos observada (Figura 3). Ainda, para folhas, notam-se mais SNVs nos genótipos IN84-58, RB72454 e SP80-3280. Nota-se que o genótipo SP80-3280 se destacou em número de sítios variantes descobertos nos dados de colmos, folhas e raízes. Para raízes, notam-se mais sítios variantes para os genótipos RB92579, SP80-3280 e RB855536. A diferença entre genótipos no número de sítios variantes encontrados pode ser explicada pelo número diferente de réplicas de cada genótipo no protocolo do GBS. Deste total de 25.830 SNVs, 7.442 eram homozigóticos para todos os genótipos deste conjunto de dados. Assim, restaram na análise 18.388 SNVs heterozigóticos, distribuídos em 4.598 transcritos.

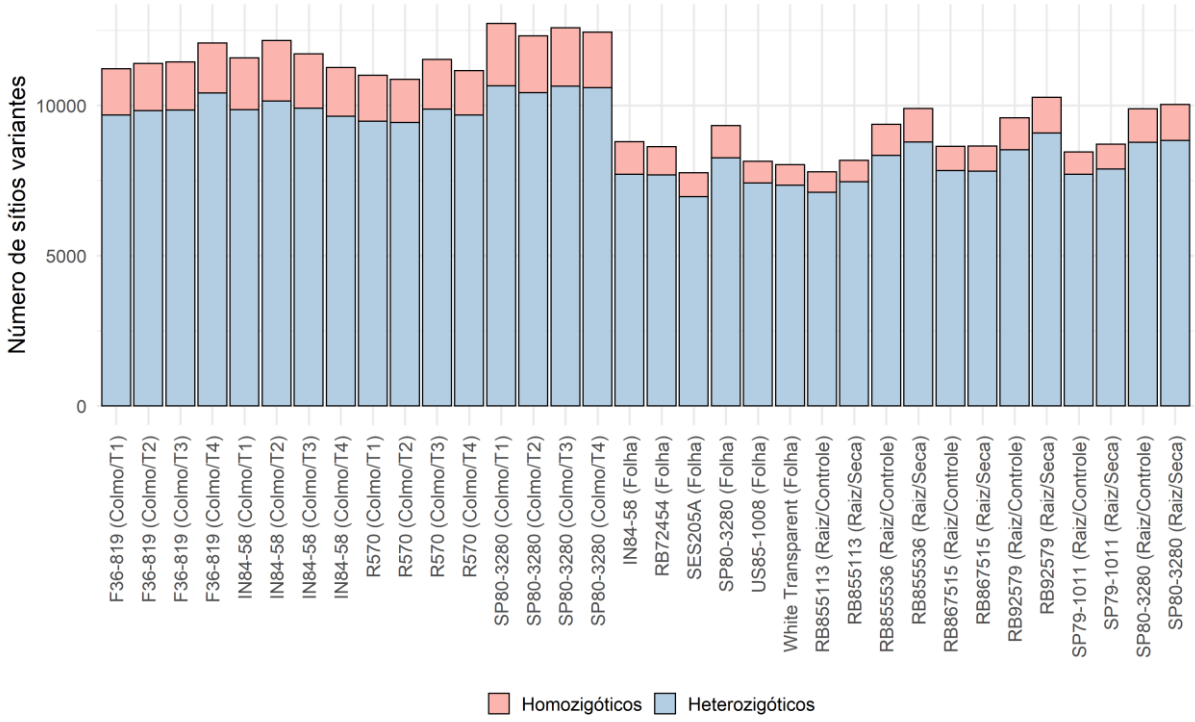

Figura 3. Número de SNVs heterozigóticos e homozigóticos observados em função do órgão, genótipo e categoria (tempo em colmos ou condição de estresse hídrico em raízes).

O experimento com colmos apresentou um total de 12.335 SNVs efetivamente investigados, o experimento com folhas 9.746, e o experimento com raízes 9.929 sítios variantes. Dos 4.598 transcritos que apareceram no conjunto de dados total, 3.809 estavam presentes nos dados de colmo, 2.907 nos dados de folha e 3.003 nos dados de raiz. Com isso, no total havia cerca de três SNVs amostrados por transcrito. Os 4.598 transcritos analisados apresentavam em média 2.423 pb. Transcritos mais longos não necessariamente apresentaram um maior número de SNVs (Figura 4). Isto provavelmente se deve às limitações intrínsecas da técnica de GBS.

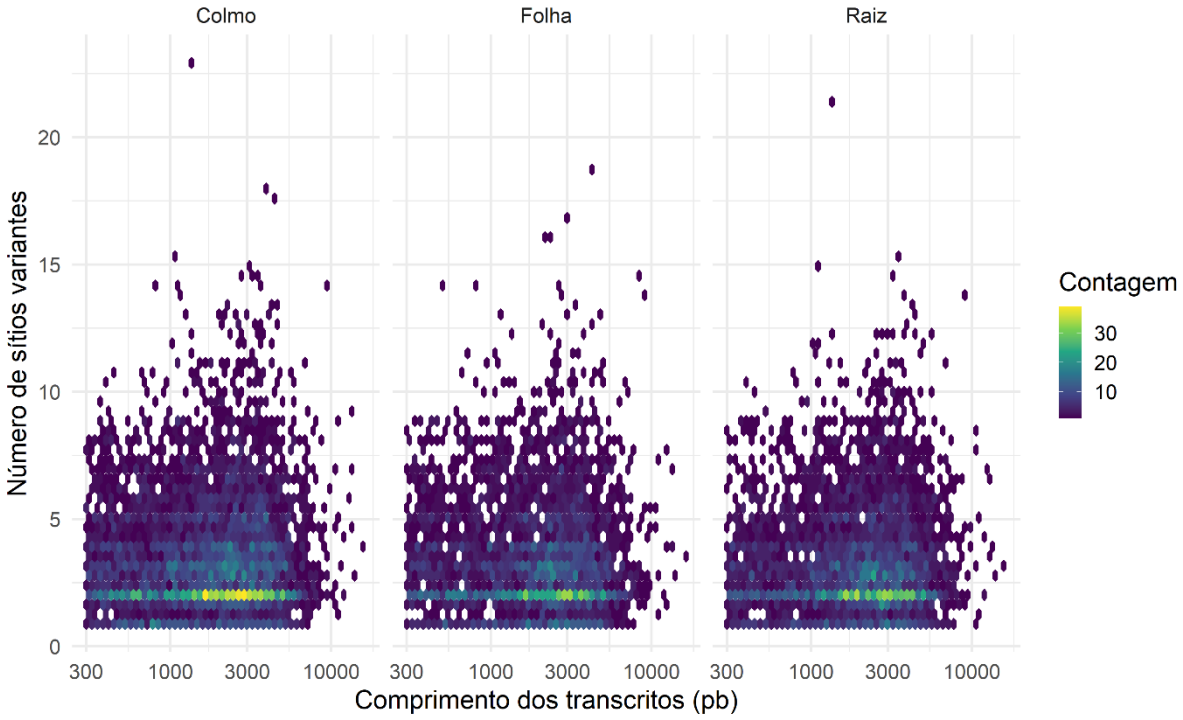

Figura 4. Número de SNVs em função do comprimento do transcrito (pb) para dados de colmos, folhas e raízes.

Dos 18.388 SNVs com pelo menos um genótipo heterozigótico, 6.629 SNVs (em 2.348 transcritos) apresentaram expressão diferencial alelo-específica (EA). Observa-se que os dados de raiz apresentaram um pico um pouco mais pronunciado de sítios variantes com expressão alelo-específica (VEA) próximo à região central dos transcritos (Figura 5). Os dados de colmo foram os que apresentaram a menor tendência de viés em função da posição relativa do sítio variante no transcrito. A natureza desses dados, inclusive o protocolo utilizado no sequenciamento, forneceu uma maior confiança na detecção de EA em colmos.

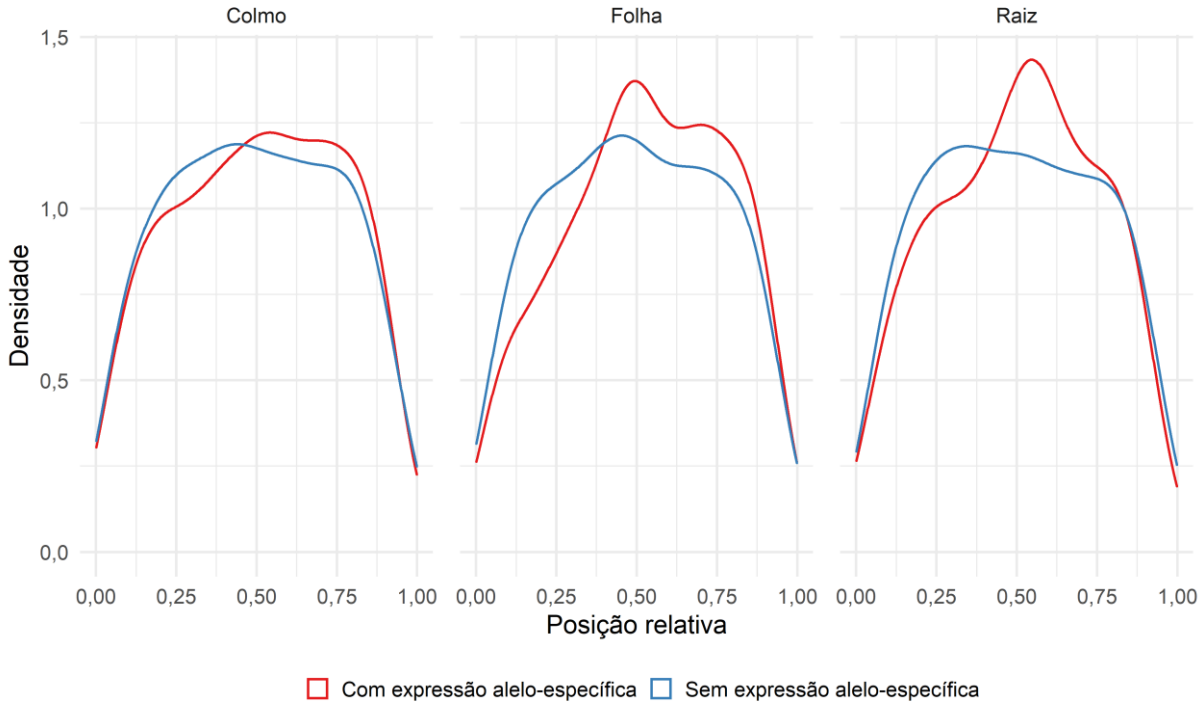

Figura 5. Densidade de SNVs em função da posição relativa no transcrito para dados de colmos, folhas e raízes.

A relação entre a proporção relativa do alelo de referência nos dados de expressão e a proporção genômica é apresentada na Figura 6, para cada genótipo, órgão e nível do fator estudado. Observa-se que alguns SNVs, apesar de apresentarem um grande desvio, não foram considerados PEA devido ao número pequeno de leituras de RNA-Seq. Isso é esperado, já que um número maior de leituras no RNA-Seq confere maior poder estatístico para o teste de hipótese nula, já que pode afastar da proporção genômica o intervalo de credibilidade da distribuição *a posteriori*.

Nota-se também um maior acúmulo de pontos no quadrante superior direito dos gráficos, indicando mais SNVs cuja proporção do alelo de referência no GBS foi maior e houve maior expressão do alelo de referência. O acúmulo de pontos neste quadrante é maior que no quadrante inferior esquerdo, indicando que houve menos SNVs com maior proporção do alelo alternativo que expressaram este alelo com maior intensidade. Além disso, há mais pontos no quadrante superior esquerdo comparado ao inferior direito, indicando uma tendência de expressar o alelo de referência de SNVs que no GBS têm uma maior proporção do alelo alternativo, e não o contrário. Isso mostra um provável artefato técnico: como foi usada a montagem *de novo* do transcriptoma para a chamada de SNVs nos dados de GBS, muitos dos alelos mais expressos foram tomados como referência. Há também uma grande quantidade de SNVs em que foi expresso somente o alelo de referência, mesmo sendo heterozigóticos (linha horizontal onde Y = 1).

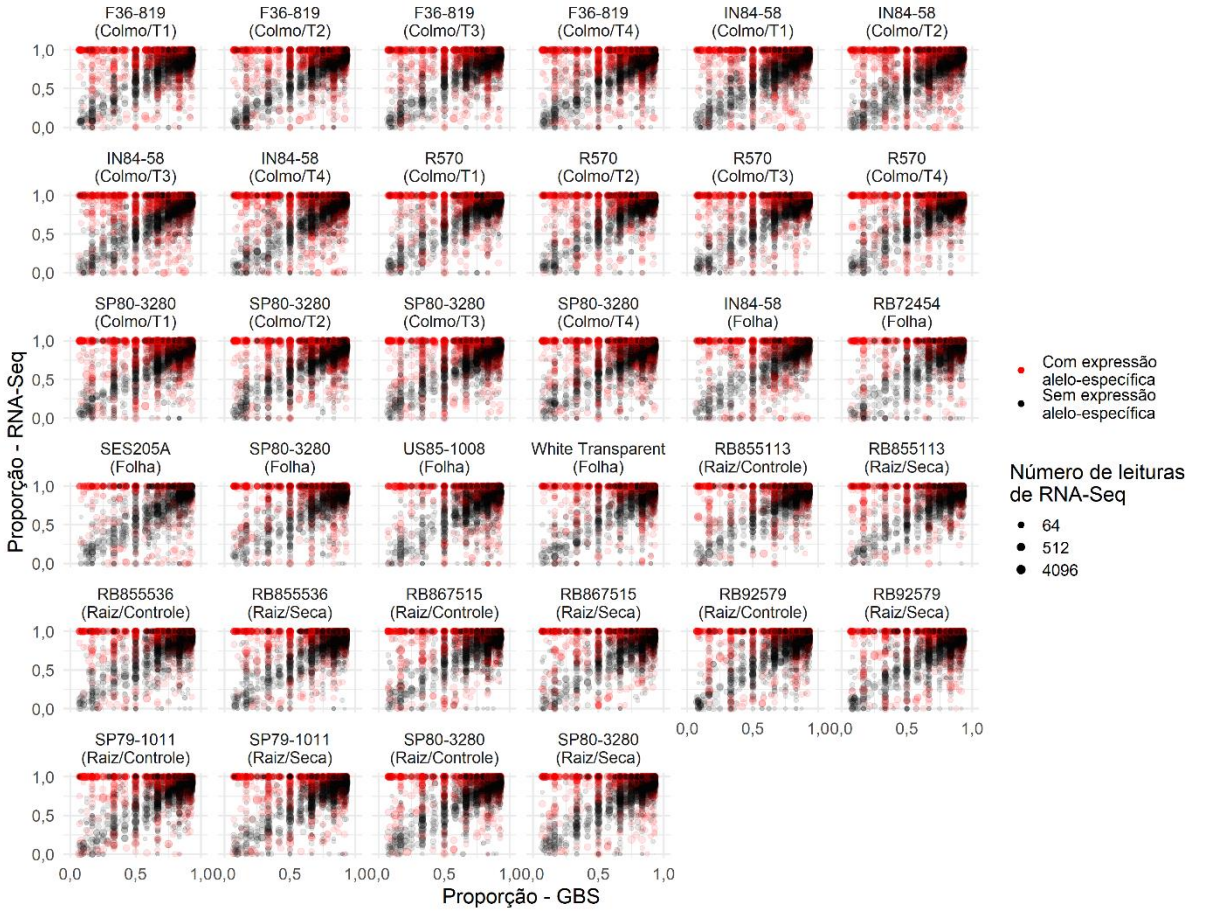

Figura 6. Relação entre a proporção do alelo de referência no RNA-Seq e a proporção do alelo de referência no GBS. Pontos vermelhos representam SNVs em que foi rejeitada a hipótese nula e o tamanho dos pontos é proporcional ao número de leituras de RNA-Seq para determinado SNV.

2.3.2 Análise de agrupamentos hierárquicos

Quando considerada a proporção relativa do alelo de referência em todos os SNVs, foi possível separar, principalmente, SES205A e IN84-58 (ambos representantes de *S. spontaneum*) e RB855113 das demais (Figura 7A). Por outro lado, quando utilizada a proporção relativa do alelo de referência e os PEA, não foi possível separar claramente os genótipos representantes de *S. spontaneum* dos demais (Figura 7B). A expressão relativa do alelo de referência separou, em primeiro lugar, a amostra de folha de SES205A das demais, tanto na análise com todos os SNVs quanto na análise somente dos PEA (Figuras 7C e 7D). Observando-se todos os SNVs, há uma separação dos representantes de *S. spontaneum* dos demais (Figura 7C). Quando se observam apenas os PEA (Figura 7D), as amostras de folhas dos genótipos SES205A, IN84-58 e US85-1008 (este último descendente de *S. spontaneum* e *S. robustum*) formam três grupos distintos, e então há um grande grupo com as amostras restantes. Em ambos os casos em que se utilizou a expressão relativa do alelo de referência, o genótipo White Transparent (*S. officinarum*) esteve agrupado com os genótipos híbridos. A diferença entre híbridos comerciais e genótipos selvagens é explicada pelos esforços de melhoramento visando aumentar principalmente o teor de açúcar (HOSAKA et al., 2021). Também, genótipos com teores elevados de sólidos solúveis tendem a ser mais homogêneos, já que apresentam uma base genética estreita, frequentemente incluindo retrocruzamentos e cruzamentos com os mesmos genitores. Nota-se que, quando utilizados somente PEA, juntamente

com amostras do genótipo IN84-58, agruparam-se também raízes de RB867515, raízes de RB855113 em condição de seca e folhas de SP80-3280. Não foi possível perceber agrupamentos quanto ao órgão, ao tempo ou à condição de seca em detrimento do genótipo, sugerindo que o genótipo é o fator mais importante para separar estas amostras tanto em relação à proporção quanto em relação à expressão relativa do alelo de referência.

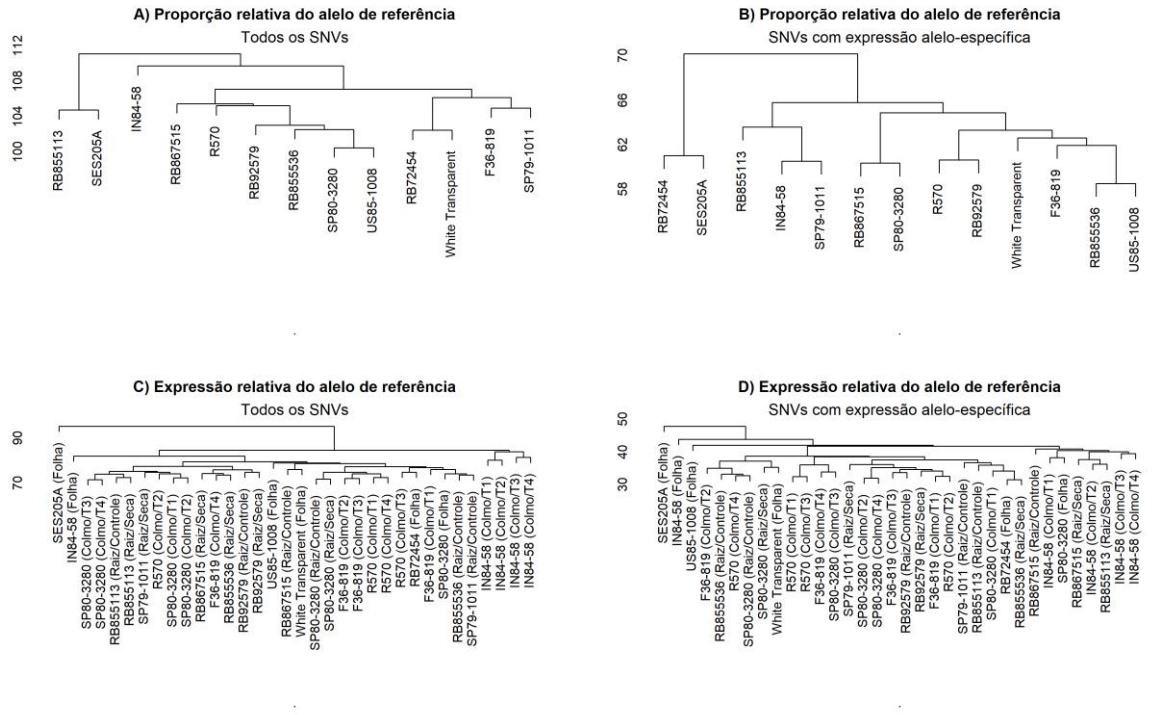

Figura 7. Agrupamentos feitos com as distâncias entre as amostras considerando-se a proporção genômica e a expressão relativa do alelo de referência, com todo os SNVs e apenas com aqueles que apresentaram expressão alelo-específica significativa.

Foi calculado o viés da expressão alélica, isso é, a proporção do alelo de referência nos dados de expressão subtraída da proporção genômica. Com base nesse valor, foi feito um dendrograma com todos os SNVs estudados (Figura 8). Nota-se como as amostras de folha de SES205A ficaram isoladas das demais genótipos, bem como um agrupamento com folhas e colmos de IN84-58. Isso mostra que foi possível diferenciar os genótipos, pelo menos os representantes de *S. spontaneum* dos demais, pela intensidade da EA, considerando-se todos os SNVs. Nota-se que, dentro do grupo formado pelos híbridos não houve um agrupamento claro de genótipos, órgãos ou níveis. Por exemplo, há um grupo que contém colmos dos tempos T1, T3 e T4 de SP80-3280, mas nesse grupo também está presente raiz de SP79-1011 em condição de seca. Além disso, as folhas de SP80-3280 e raízes de SP80-3280 estão em grupos distintos, ambos contendo raízes, folhas e colmos de outros genótipos.

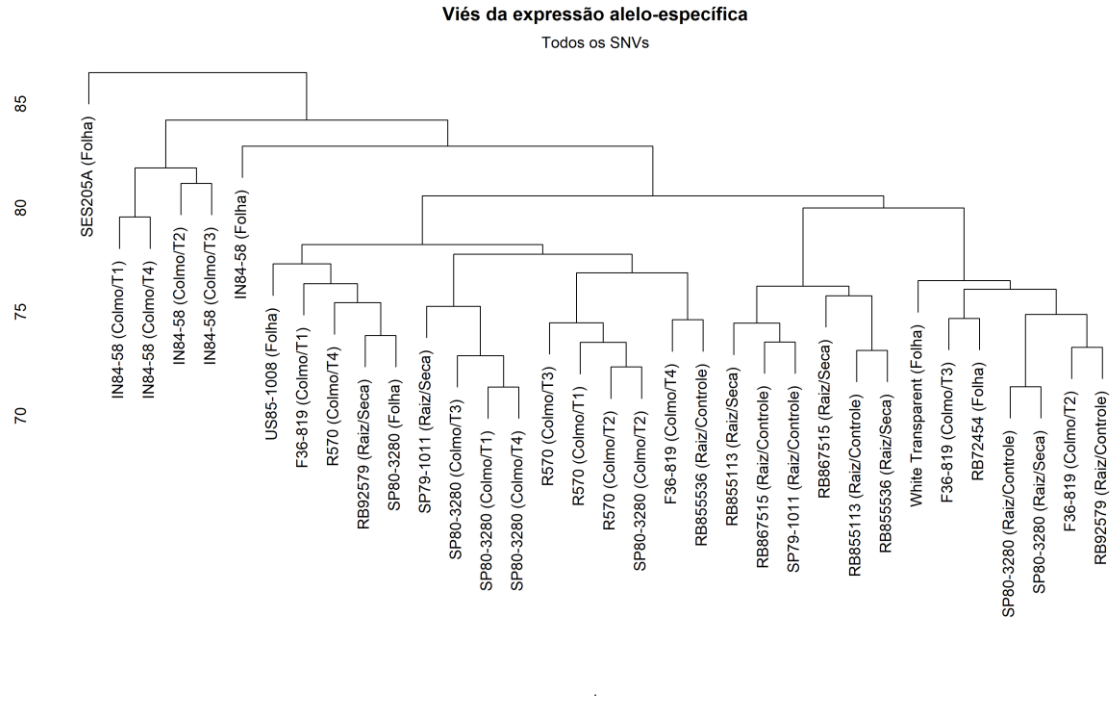

Figura 8. Agrupamentos com base nas distâncias entre as amostras considerando-se o viés de expressão alelo específica (proporção do alelo de referência nos dados de expressão subtraída da proporção genômica), com todo os SNVs analisados.

2.3.3 Perfil do conjunto de dados quanto à ocorrência de expressão alelo-específica

Observa-se, nas Figuras 9 a 11, que a minoria (de 20% a 30%) dos SNVs apresentaram EA significativa em cada órgão, genótipo e nível do fator. As amostras de colmo apresentaram de 27% a 30% de SNVs com expressão alelo-específica (Figura 9). O genótipo IN84-58 apresentou os maiores valores, chegando a apresentar 31,5% de SNVs com EA no tempo T2. De modo geral, o genótipo R570 apresentou menores valores em todos os tempos. Os dados de folha apresentaram entre 23% e 24% de SNVs com EA, com exceção de IN84-58 e US85-1008, que apresentaram 26,8% e 21,3%, respectivamente. Por fim, os dados de raiz apresentaram valores entre 22% e 23%, exceto pelos genótipos RB855113 e SP79-1011, que apresentaram valores próximos a 24% nos dois tratamentos, e SP80-3280, que apresentou 24,9% de SNVs com EA em situação de controle.

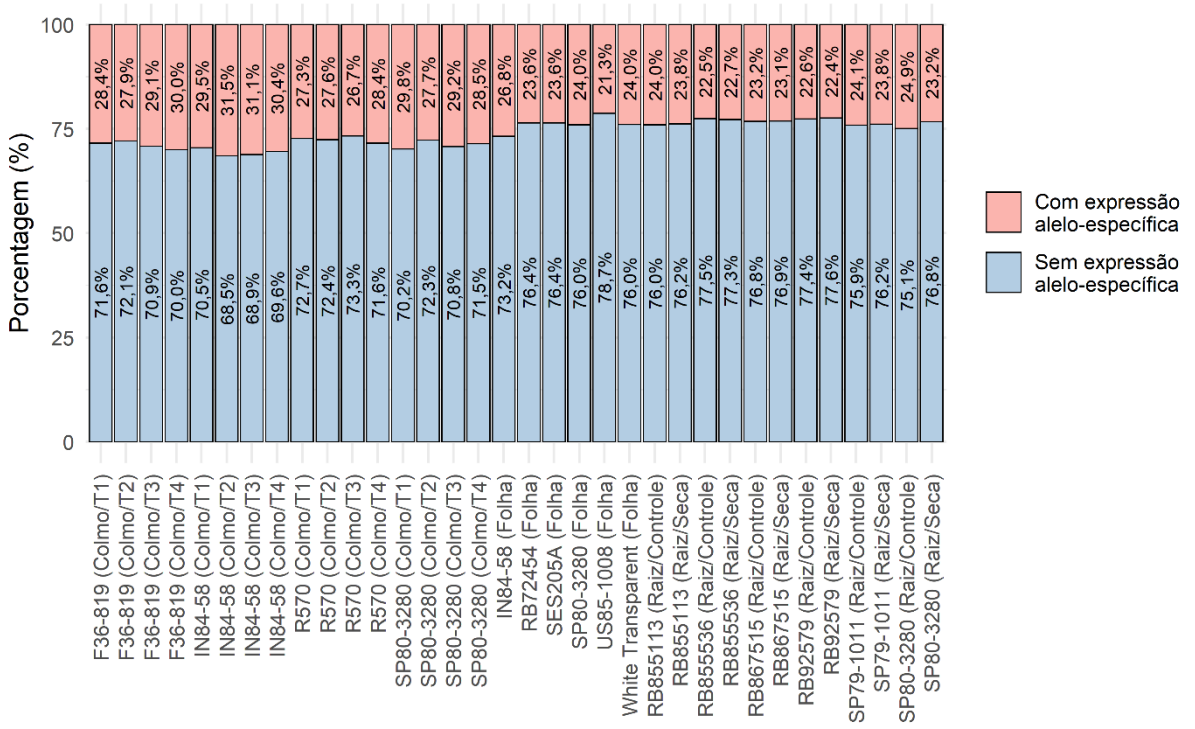

Figura 9. Porcentagens de SNVs com e sem expressão alelo-específica para cada órgão, genótipo e nível do fator.

Na Figura 10 observam-se as porcentagens de significância da EA para os níveis dos fatores estudados dentro de órgãos e genótipos. Para os dados de colmo (quatro níveis), de 15,7% (R570) a 18,5% (F36-819) dos SNVs apresentaram EA em todos os níveis, enquanto de 13,7% (F36-819) a 17% (R570) apresentaram em pelo menos um nível, mas não em todos. Nos dados de raiz, 17,8% (RB92579) a 19,8% (SP80-3280) apresentaram EA em ambas as condições (seca e controle), enquanto 5% (SP80-3280) a 5,9% (RB855113) apresentaram EA em somente uma das condições. Essa maior porcentagem de SNVs com resultados contrastantes em diferentes níveis nos dados de colmo pode ser parcialmente explicada por haver mais comparações sendo feitas. Além disso, o poder estatístico para detecção da EA foi maior no experimento de colmo, já que houve uma maior cobertura no RNA-Seq e mais sítios passaram pelos filtros. Isto é consistente também com a observação de maior frequência de SNVs sem EA no experimento de raízes, para todos os tratamentos (Figura 9). As porcentagens de SNVs de um mesmo genótipo que não apresentaram EA significativa (Figura 9) são próximas das porcentagens de SNVs que não apresentaram diferença na EA em todos os níveis (Figura 10). A diferença é maior quando comparadas as porcentagens de SNVs de um mesmo genótipo que apresentaram EA (Figura 9) com as porcentagens de SNVs que apresentaram EA em todos os níveis (Figura 10). Essas observações sugerem que a EA é pouco consistente entre tratamentos.

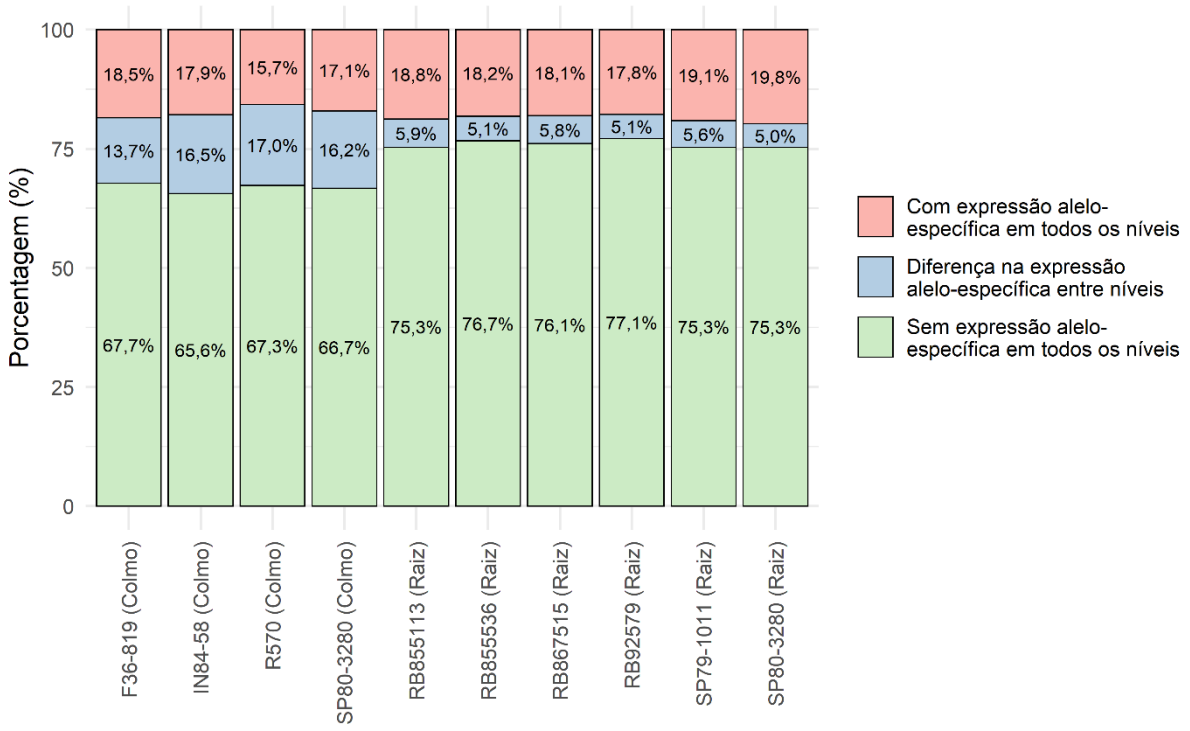

Figura 10. Porcentagem de SNVs com ou sem expressão alelo-específica em todos os níveis e porcentagem de SNVs com diferença entre pelo menos dois dos níveis do fator estudado, em função do órgão e do genótipo.

Comparando-se os genótipos com amostras de múltiplos órgãos, observa-se que 12,4% dos SNVs de IN84-58 apresentaram expressão alelo-específica em todos os níveis e órgãos (folhas e colmos em quatro tempos), enquanto que 64,3% não apresentaram expressão alelo-específica em qualquer nível ou órgão (Figura 11). Já para o genótipo SP80-3280, 9,2% dos SNVs apresentaram EA em todos os níveis e órgãos (folhas, colmos em quatro tempos e raízes em duas condições), enquanto 64,5% não apresentaram EA em nível ou órgão algum. Comparando-se esses genótipos nas Figuras 9 e 10, nota-se que o IN84-58 consistentemente apresentou uma maior proporção de PEA em todos os níveis do que SP80-3280. Em contrapartida, 14,1% dos SNVs que aparecem em SP80-3280 apresentaram diferença quanto à presença de expressão alelo-específica tanto entre níveis quanto entre órgãos, enquanto este número em IN84-58 foi de 10,8%. Comparado à Figura 10, observam-se menos SNVs com expressão alelo-específica em todos os casos. Por outro lado, foi semelhante o número de SNVs sem expressão alelo-específica em nenhum caso.

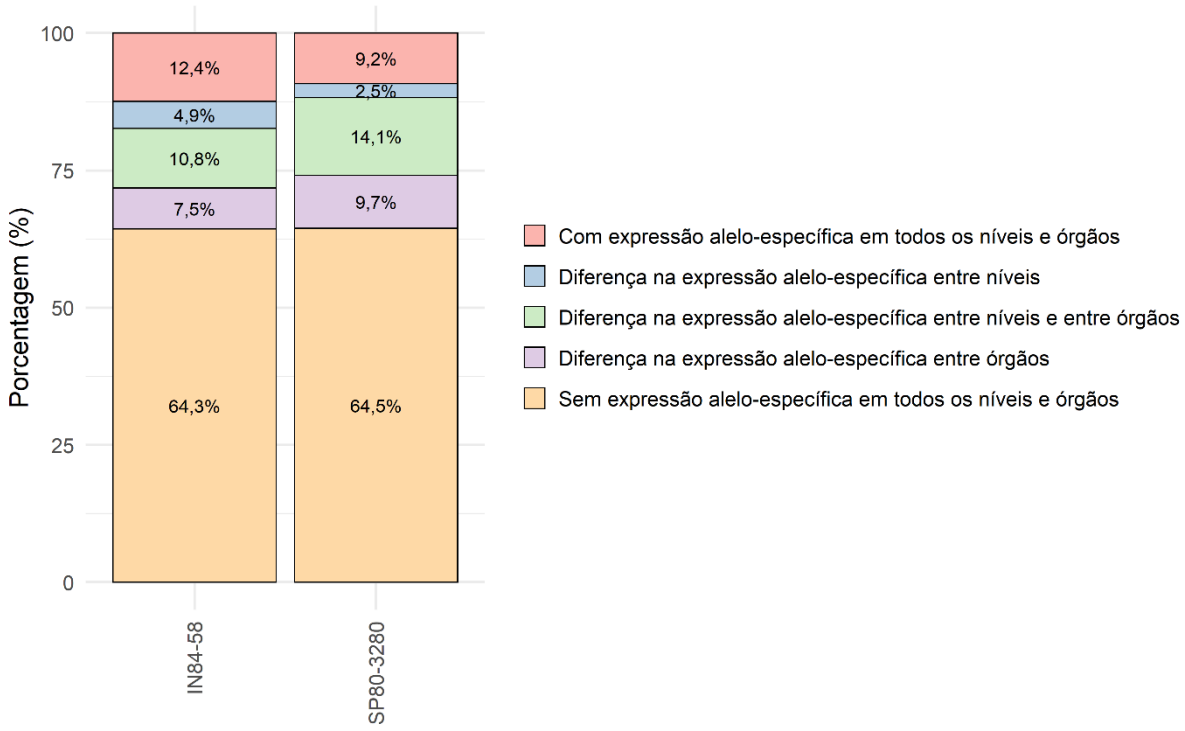

Figura 11. Porcentagem de SNVs com ou sem expressão alelo-específica em todos os níveis do fator estudado e todos os órgãos e porcentagem de SNVs com diferença entre pelo menos dois dos níveis ou dois órgãos, em função do genótipo.

Nas Figuras 12 e 13, observa-se o número de sítios variantes e de transcritos exclusivos por genótipo ou por grupo de genótipos. Considerou-se um transcrito com expressão alelo-específica (TEA) quando pelo menos um SNV contido nele apresentou EA significativa. Visto que foram observadas 2.649 diferentes combinações de genótipos com SNVs exclusivos e 1.828 combinações com transcritos exclusivos, são mostrados nas figuras os 40 grupos com maior número de SNVs ou de transcritos sem expressão alelo-específica em todos os casos. Considerou-se que houve diferença entre genótipos, órgãos ou níveis quando foi detectada EA significativa em uma das categorias dos fatores estudados, mas não em outra.

Observa-se que a maior parte de sítios variantes e de transcritos apresentaram expressão alelo-específica exclusivamente em um único genótipo. Separadamente, cada um dos 13 genótipos estudados apresentou mais SNVs com expressão alelo-específica que quaisquer grupos de genótipos. Apesar de os genótipos IN84-58 e SP80-3280 aparecerem em mais de um experimento, mesmo quando considerados os experimentos isoladamente estes ainda foram os genótipos em que mais aparecem transcritos e sítios variantes exclusivos, como notado por Correr et al. (2022). Nota-se ainda que IN84-58 apresentou um número maior de SNVs com alguma diferença quanto à EA (seja entre órgãos, entre genótipos ou entre os níveis dos fatores estudados). Considerando apenas diferenças entre níveis (entre tempos ou entre condições de estresse hídrico), o genótipo IN84-58 apresentou 282 (13,6%) SNVs nessa categoria, enquanto SP80-3280 apresentou 134 (11%), F36-819 apresentou 116 (11,9%) e R570 apresentou 77 (11,1%). Dos SNVs exclusivos do genótipo IN84-58, 14,9% apresentaram diferenças entre os órgãos, enquanto para SP80-3280 esse número foi de 14,3%. Por fim, 176 (8,5%) SNVs exclusivos de IN84-58 apresentaram simultaneamente diferença entre níveis e órgãos, enquanto 103 (7,9%) dos SNVs exclusivos de SP80-3280 apresentaram essa diferença.

Não parece haver uma tendência de os genótipos híbridos apresentarem maior número de PEA que os selvagens. Os quatro genótipos para os quais há dados de colmo apareceram primeiro, tanto em relação ao número de

sítios variantes quanto em relação ao número de transcritos. Como observado na Figura 3, houve um maior número de SNVs encontrados nos dados de colmo. Em nível de sítios (Figura 12), depois dos primeiros quatro genótipos vêm os genótipos US85-1008 (híbrido), RB92579 (híbrido), SP79-1011 (híbrido), SES205A (*S. spontaneum*), RB855536 (híbrido), White Transparent (*S. officinarum*), RB867515 (híbrido) e RB72454 (híbrido). O grupo de genótipos F36-819 e SP80-3280 apresenta mais sítios variantes exclusivos sem EA em todas as categorias que o genótipo RB855113. Para F36-819 e SP80-3280, foram obtidos dados de colmo, que têm uma tendência de apresentar mais SNVs exclusivos e, portanto, mais SNVs sem EA. Pode-se notar que entre os SNVs exclusivos desse par de genótipos houve apenas 4 (1,6%) SNVs com expressão alelo-específica em todos os casos, comparado com 16 (8,8%) SNVs com EA em todos os casos em RB855113.

Foi observado apenas um SNV presente em todos os 13 genótipos simultaneamente. Os grupos de genótipos com maior quantidade de sítios variantes e de transcritos com expressão alelo-específica são formados por genótipos presentes no experimento de colmos. Em maior quantidade, como já foi discutido, F36-819 (baixo teor de sólidos solúveis) e SP80-3280 (teor de sólidos solúveis muito alto), seguido de R570 (alto) e F36-819 (baixo), R570 (baixo) e SP80-3280 (muito alto) e então os grupos envolvendo IN84-58 (muito baixo). Isto reforça a ideia de que os genótipos híbridos são mais homogêneos entre si do que se comparados a genótipos selvagens.

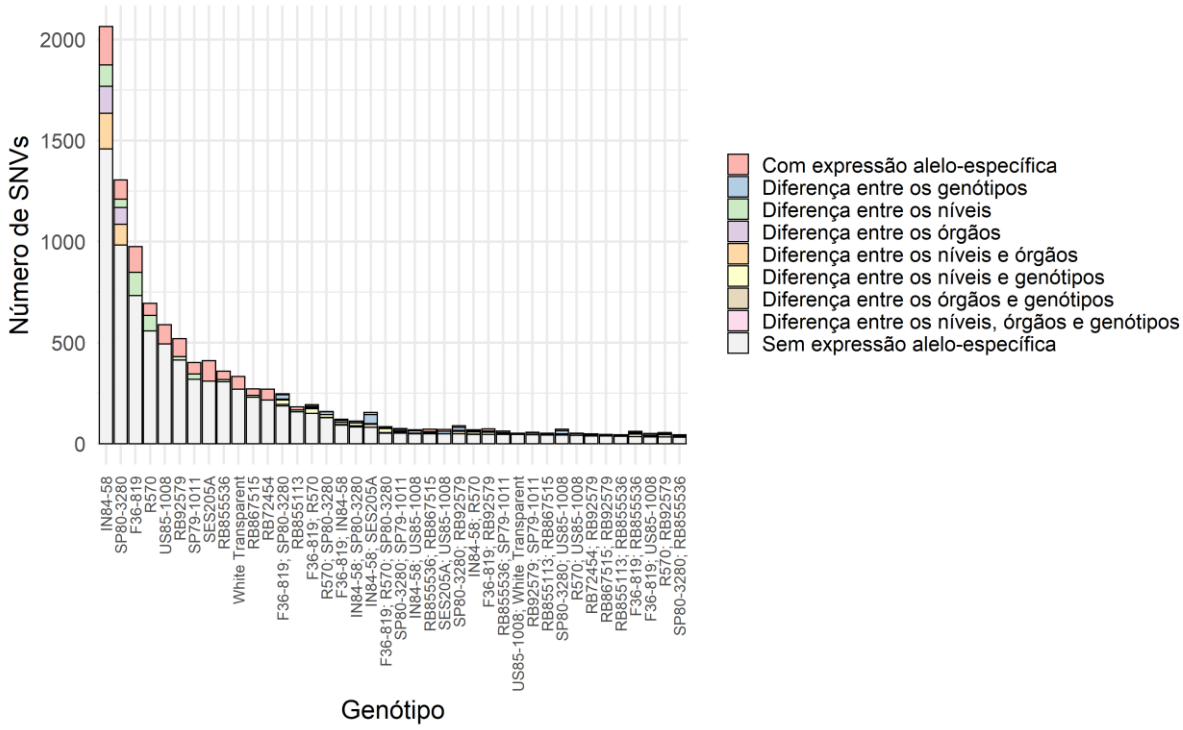

Figura 12. Número de sítios variantes exclusivos de cada genótipo ou grupo de genótipos com e sem expressão alelo-específica (40 grupos de genótipo com maior número de sítios variantes). O preenchimento das barras indica o número de SNVs de determinado genótipo ou grupo de genótipos com ou sem expressão alelo-específica em todos os casos, ou com diferença na expressão alelo-específica em casos específicos (entre níveis, órgãos ou genótipos do mesmo grupo).

Observando-se os transcritos exclusivos por grupos de genótipos (Figura 13), observa-se algo semelhante à Figura 12, mas apenas os genótipos IN84-58 (colmos, folhas e raízes), SP80-3280 (colmos, folhas e raízes), F36-819 (colmos), R570 (colmos), US85-1008 (folhas) e RB92579 (raízes) aparecem antes dos primeiros grupos de genótipos.

Inclusive, um grupo com nove genótipos (F36-819; IN84-58; R570; SP80-3280; RB855113; RB855536; RB867515; RB92579; SP79-1011) está entre os 40 com mais transcritos exclusivos sem EA.

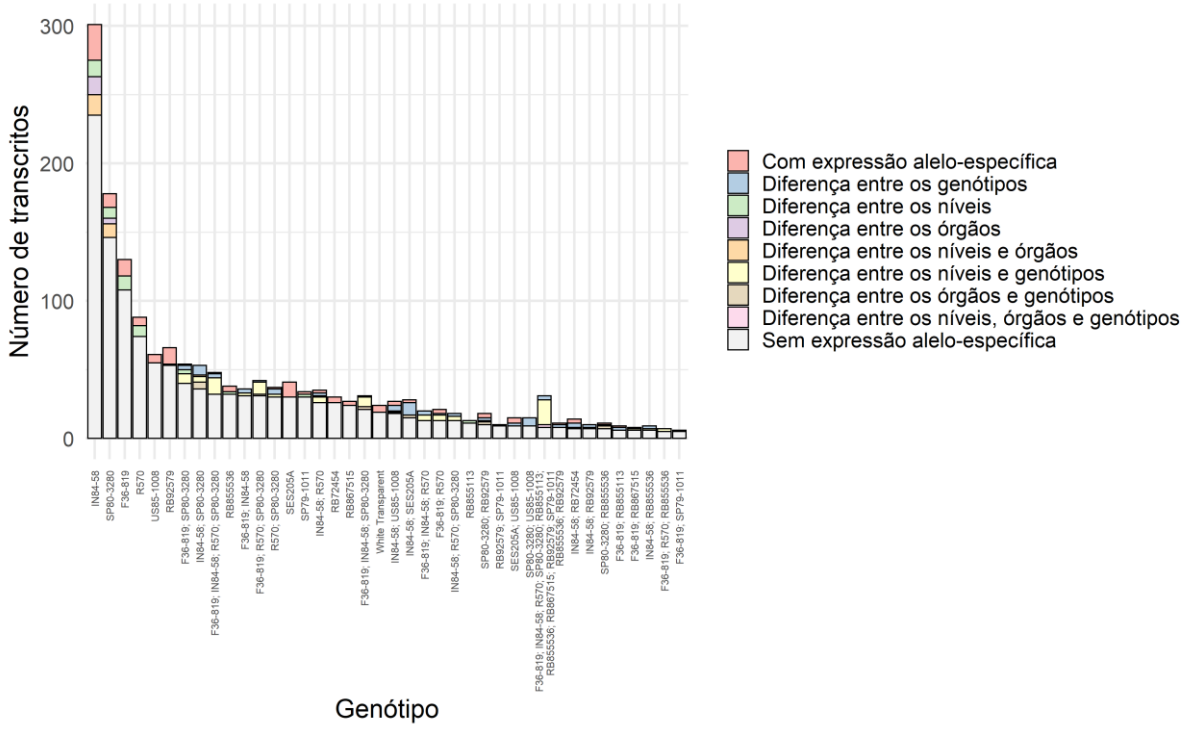

Figura 13. Número de transcritos exclusivos de cada genótipo ou grupo de genótipos com e sem expressão alelo-específica (40 grupos de genótipo com maior número de transcritos). O preenchimento das barras indica o número de transcritos de determinado genótipo ou grupo de genótipos com ou sem expressão alelo-específica em todos os casos, ou com diferença na expressão alelo-específica em casos específicos (entre níveis, órgãos ou genótipos do mesmo grupo).

Nas Figuras 14 e 15, observa-se o número de SNVs e de transcritos exclusivos de grupos de categorias dos fatores experimentais. Como foram observadas 122 diferentes combinações de níveis com SNVs exclusivos e 110 combinações com transcritos exclusivos, são mostrados nas figuras os 40 grupos com mais SNVs ou transcritos sem EA. Observando-se os SNVs exclusivos (Figura 14), em primeiro lugar, existem mais PEA exclusivos de folha (381 PEA, 17,9%). Em segundo lugar, estão os PEA exclusivos de raízes, mas compartilhados nas duas condições (329, 16,6%). Em terceiro lugar, estão os PEA compartilhados por todos os níveis dos três órgãos (298, 8,6%). Em quarto lugar, aparecem os SNVs exclusivos de colmo, mas compartilhados nos quatro tempos (209, 12,9%). A seguir, estão as demais combinações de níveis. Observa-se, na Figura 14, como o número de SNVs com diferenças entre genótipos sobressai com relação ao número de sítios variantes com diferença na EA entre níveis ou entre órgãos.

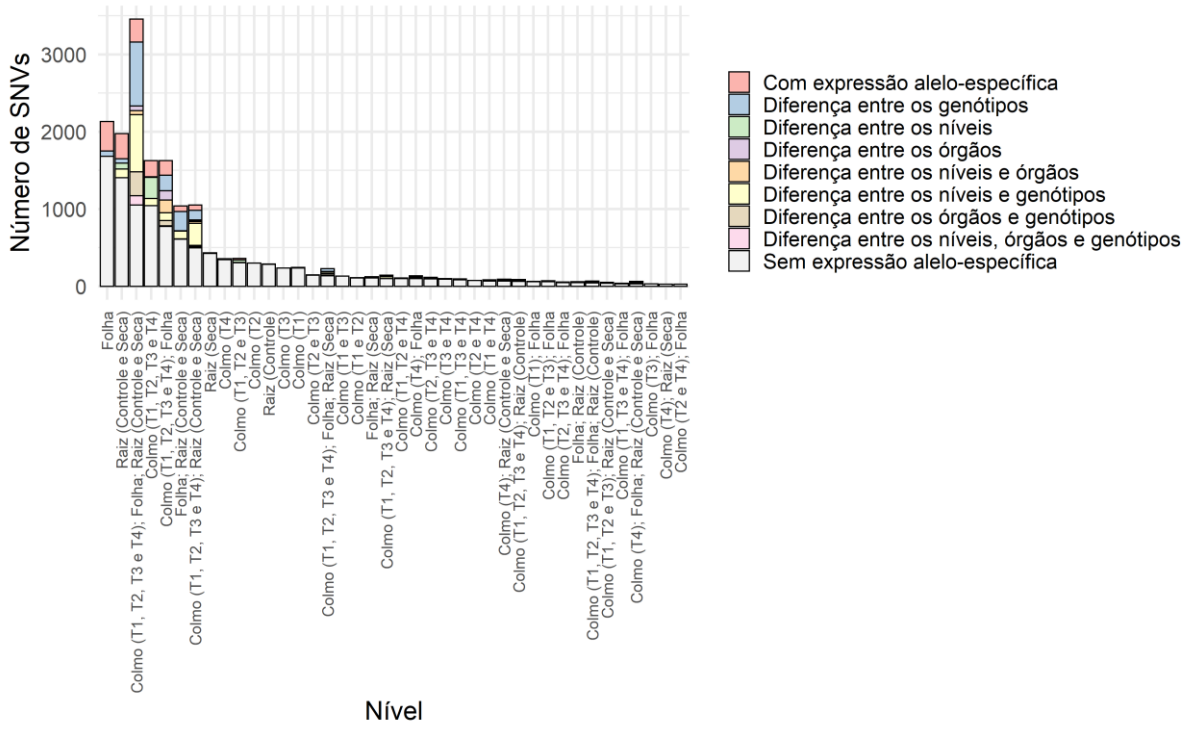

Figura 14. Número de sítios variantes exclusivos de cada nível ou grupo de níveis com e sem expressão alelo-específica (40 grupos de níveis com maior número de sítios variantes). O preenchimento das barras indica o número de SNVs de determinado nível ou grupo de níveis com ou sem expressão alelo-específica em todos os casos, ou com diferença na expressão alelo-específica em casos específicos (entre genótipos, órgãos ou níveis do mesmo grupo).

Como se pode observar na Figura 15, a grande maioria dos transcritos apareceu em todos os níveis estudados. Observa-se também que o tempo T4 apresentou mais SNVs (Figura 14) e transcritos exclusivos (Figura 15) que os demais tempos em colmo. Além disso, nota-se como a diferença entre genótipos predomina com relação à diferença entre órgãos e entre níveis.

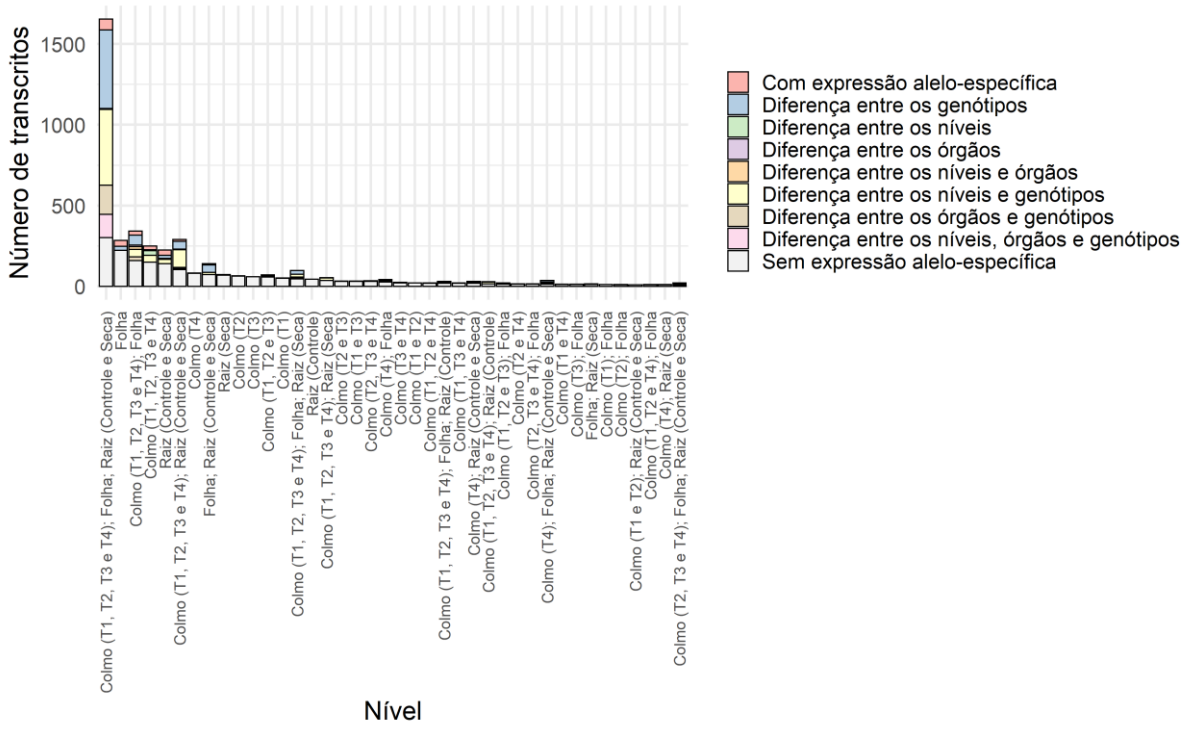

Figura 15. Número de transcritos exclusivos de cada nível ou grupo de níveis com e sem expressão alelo-específica (40 grupos de níveis com maior número de transcritos). O preenchimento das barras indica o número de transcritos de determinado nível ou grupo de níveis com ou sem expressão alelo-específica em todos os casos, ou com diferença na expressão alelo-específica em casos específicos (entre genótipos, órgãos ou níveis do mesmo grupo).

2.3.4 Análises de enriquecimento funcional

Os únicos termos GO enriquecidos foram GO:0005737 (citoplasma) e GO:0005634 (núcleo). Em parte, esse resultado é provavelmente devido às limitações da técnica de GBS: como a detecção dos sítios variantes depende da presença do sítio de restrição dentro das regiões transcritas, os termos GO enriquecidos tendem a ser aqueles com maior número de genes, os quais se referem a estruturas mais gerais da célula e não a processos específicos. Na Figura 16, observam-se os termos GO com mais TEA associados, em função da amostra. Dentre os termos associados a mais transcritos com SNVs diferencialmente expressos, também estavam GO:0008422 (atividade beta-glicosidase), GO:0043531 (ligação ao ADP), GO:0046872 (ligação a íon metálico) e GO:0005524 (ligação ao ATP). Também havia termos relacionados à atividade endopeptidase (GO:0004190), ligação ao DNA (GO:0003677), ubiquitinação de proteínas (GO:0016567) e atividade oxirredutase (GO:0016491).

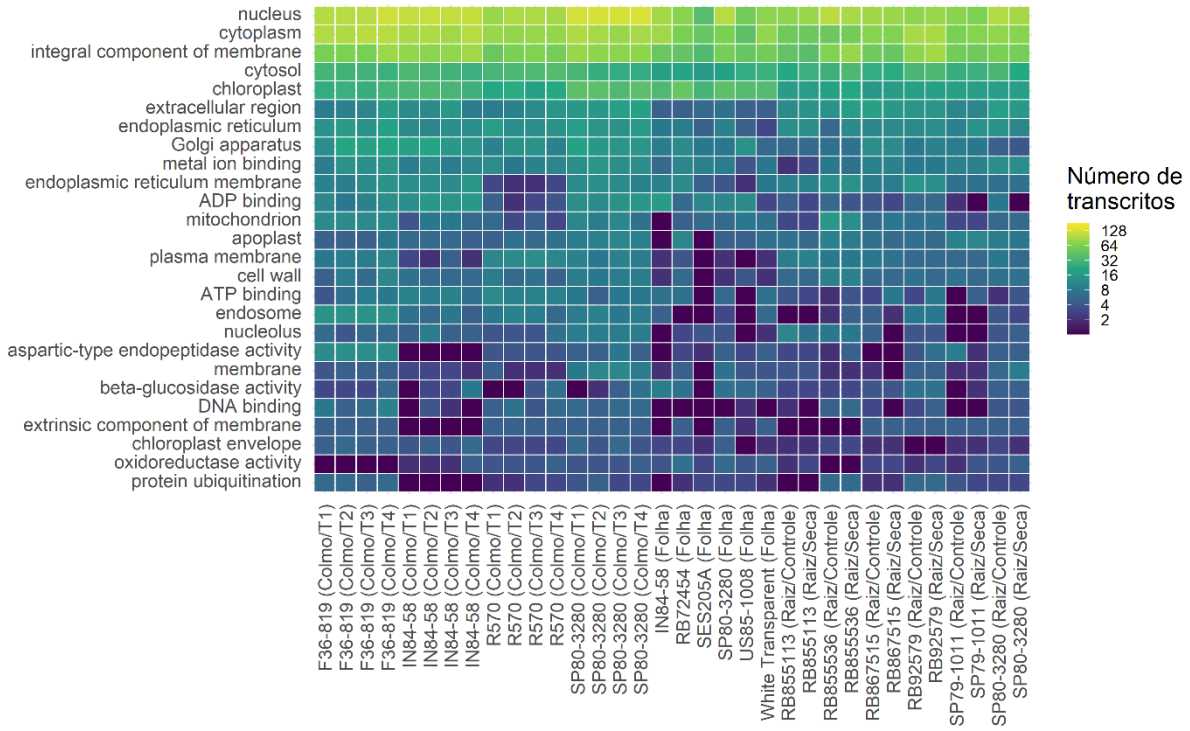

Figura 16. Termos GO associados a genes com expressão alelo-específica no conjunto de dados. A escala de cores representa o número de transcritos associados a dado termo GO para cada amostra.

Pelo número de TEA associados aos termos GO (Figura 17), é possível perceber um forte agrupamento por genótipos e, dentro desses grupos, um agrupamento por órgãos. Observam-se oito grupos: colmos de F36-819; colmos, raízes e folhas de SP80-3280; colmos e folhas de IN84-58; folhas de RB72454, White Transparent, SES205A e US85-1008; colmos de R570; raízes de RB92579 e RB867515; raízes de SP79-1011; e raízes de RB855536 e RB855113.

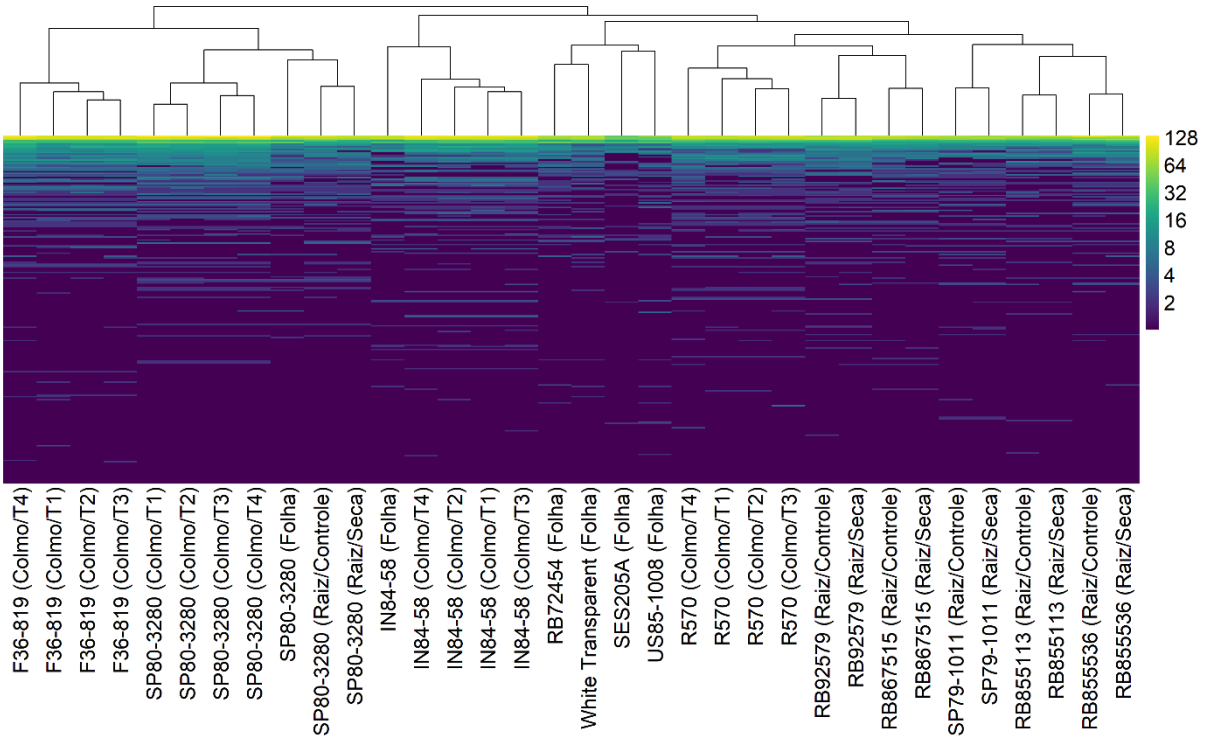

Figura 17. Agrupamento hierárquico a partir do número de transcritos com expressão diferencial de alelos associados aos termos GO. A escala de cores representa o número de transcritos associados a dado termo GO para cada amostra.

Além disso, buscando pelas rotas metabólicas do banco de dados KEGG (Tabela 2) observam-se entre os sítios variantes com expressão alelo-específica, muitos relacionados à biossíntese de metabólitos secundários, à transdução do sinal de fito-hormônios, ao metabolismo de carbono, ao metabolismo de amido e sacarose, à rota de sinalização MAPK, à proteólise mediada por ubiquitina e à interação planta-patógeno. No entanto, esses valores foram proporcionalmente bastante semelhantes aos encontrados em sítios variantes sem expressão alelo-específica. Isso indica que não houve rotas metabólicas enriquecidas para transcritos com expressão alelo-específica.

Tabela 2. Número de sítios variantes com e sem expressão alelo-específica em transcritos presentes em cada uma das rotas metabólicas encontradas no banco de dados KEGG (continua).

Rota metabólica	Sítios variantes		Transcritos	
	Sem expressão alelo-específica	Com expressão alelo-específica	Sem expressão alelo-específica	Com expressão alelo-específica
Metabolic pathways	857 (18,0%)	420 (17,5%)	101 (20,4%)	142 (16,7%)
Biosynthesis of secondary metabolites	514 (10,8%)	249 (10,4%)	57 (11,5%)	87 (10,2%)
Plant hormone signal transduction	166 (3,5%)	77 (3,2%)	12 (2,4%)	28 (3,3%)
Plant-pathogen interaction	136 (2,9%)	44 (1,8%)	20 (4,0%)	16 (1,9%)
MAPK signaling pathway - plant	123 (2,6%)	56 (2,3%)	9 (1,8%)	20 (2,4%)
Carbon metabolism	121 (2,5%)	74 (3,1%)	9 (1,8%)	22 (2,6%)
Ubiquitin mediated proteolysis	108 (2,3%)	55 (2,3%)	5 (1,0%)	19 (2,2%)
Starch and sucrose metabolism	106 (2,2%)	59 (2,5%)	12 (2,4%)	19 (2,2%)
Biosynthesis of amino acids	97 (2,0%)	42 (1,7%)	7 (1,4%)	17 (2,0%)
Protein processing in endoplasmic reticulum	76 (1,6%)	41 (1,7%)	9 (1,8%)	15 (1,8%)
Biosynthesis of various plant secondary metabolites	73 (1,5%)	43 (1,8%)	3 (0,6%)	10 (1,2%)
Endocytosis	71 (1,5%)	30 (1,2%)	4 (0,8%)	12 (1,4%)
Nucleocytoplasmic transport	71 (1,5%)	47 (2,0%)	3 (0,6%)	17 (2,0%)
Cyanoamino acid metabolism	62 (1,3%)	37 (1,5%)	5 (1,0%)	9 (1,1%)
Amino sugar and nucleotide sugar metabolism	59 (1,2%)	34 (1,4%)	3 (0,6%)	10 (1,2%)
Phenylpropanoid biosynthesis	56 (1,2%)	28 (1,2%)	9 (1,8%)	8 (0,9%)
Biosynthesis of cofactors	53 (1,1%)	17 (0,7%)	9 (1,8%)	9 (1,1%)
DNA replication	52 (1,1%)	25 (1,0%)	7 (1,4%)	10 (1,2%)
Homologous recombination	51 (1,1%)	25 (1,0%)	6 (1,2%)	12 (1,4%)
Diterpenoid biosynthesis	50 (1,0%)	17 (0,7%)	4 (0,8%)	5 (0,6%)
Mismatch repair	48 (1,0%)	21 (0,9%)	6 (1,2%)	9 (1,1%)
Glycerophospholipid metabolism	46 (1,0%)	17 (0,7%)	8 (1,6%)	7 (0,8%)
Glutathione metabolism	45 (0,9%)	26 (1,1%)	7 (1,4%)	8 (0,9%)
Nucleotide excision repair	44 (0,9%)	21 (0,9%)	5 (1,0%)	10 (1,2%)
Ribosome	44 (0,9%)	19 (0,8%)	4 (0,8%)	7 (0,8%)
Gluconeogenesis	41 (0,9%)	30 (1,2%)	1 (0,2%)	8 (0,9%)
Glycolysis	41 (0,9%)	30 (1,2%)	1 (0,2%)	8 (0,9%)
Fatty acid metabolism	38 (0,8%)	12 (0,5%)	4 (0,8%)	6 (0,7%)
Oxidative phosphorylation	38 (0,8%)	20 (0,8%)	4 (0,8%)	6 (0,7%)
Biosynthesis of unsaturated fatty acids	36 (0,8%)	11 (0,5%)	4 (0,8%)	5 (0,6%)
Carbon fixation in photosynthetic organisms	35 (0,7%)	23 (1,0%)	3 (0,6%)	5 (0,6%)

Tabela 2. Número de sítios variantes com e sem expressão alelo-específica em transcritos presentes em cada uma das rotas metabólicas encontradas no banco de dados KEGG (continuação).

Rota metabólica	Sítios variantes		Transcritos	
	Sem expressão alelo-específica	Com expressão alelo-específica	Sem expressão alelo-específica	Com expressão alelo-específica
Galactose metabolism	34 (0,7%)	17 (0,7%)	4 (0,8%)	5 (0,6%)
Phenylalanine, tyrosine and tryptophan biosynthesis	34 (0,7%)	10 (0,4%)	1 (0,2%)	4 (0,5%)
Pyruvate metabolism	34 (0,7%)	29 (1,2%)	2 (0,4%)	6 (0,7%)
Autophagy - other	32 (0,7%)	14 (0,6%)	1 (0,2%)	6 (0,7%)
Tyrosine metabolism	31 (0,6%)	15 (0,6%)	1 (0,2%)	5 (0,6%)
Pentose phosphate pathway	30 (0,6%)	11 (0,5%)	2 (0,4%)	3 (0,4%)
Ribosome biogenesis in eukaryotes	30 (0,6%)	20 (0,8%)	5 (1,0%)	6 (0,7%)
Glycerolipid metabolism	29 (0,6%)	12 (0,5%)	7 (1,4%)	4 (0,5%)
Spliceosome	29 (0,6%)	15 (0,6%)	3 (0,6%)	6 (0,7%)
alpha-Linolenic acid metabolism	28 (0,6%)	22 (0,9%)	1 (0,2%)	8 (0,9%)
Terpenoid backbone biosynthesis	27 (0,6%)	14 (0,6%)	4 (0,8%)	4 (0,5%)
2-Oxocarboxylic acid metabolism	26 (0,5%)	15 (0,6%)	2 (0,4%)	5 (0,6%)
Biosynthesis of nucleotide sugars	26 (0,5%)	17 (0,7%)	1 (0,2%)	5 (0,6%)
Peroxisome	26 (0,5%)	18 (0,7%)	3 (0,6%)	9 (1,1%)
Aminoacyl-tRNA biosynthesis	25 (0,5%)	11 (0,5%)	2 (0,4%)	4 (0,5%)
Valine, leucine and isoleucine degradation	25 (0,5%)	19 (0,8%)	3 (0,6%)	6 (0,7%)
Citrate cycle (TCA cycle)	24 (0,5%)	14 (0,6%)	1 (0,2%)	3 (0,4%)
Glycine, serine and threonine metabolism	24 (0,5%)	22 (0,9%)	1 (0,2%)	7 (0,8%)
Arginine and proline metabolism	23 (0,5%)	18 (0,7%)	1 (0,2%)	5 (0,6%)
beta-Alanine metabolism	23 (0,5%)	22 (0,9%)	2 (0,4%)	8 (0,9%)
Carotenoid biosynthesis	22 (0,5%)	13 (0,5%)	2 (0,4%)	4 (0,5%)
mRNA surveillance pathway	21 (0,4%)	11 (0,5%)	0 (0,0%)	5 (0,6%)
N-Glycan biosynthesis	21 (0,4%)	12 (0,5%)	1 (0,2%)	3 (0,4%)
Fatty acid elongation	20 (0,4%)	3 (0,1%)	1 (0,2%)	1 (0,1%)
Purine metabolism	20 (0,4%)	10 (0,4%)	1 (0,2%)	3 (0,4%)
Cysteine and methionine metabolism	18 (0,4%)	6 (0,2%)	2 (0,4%)	4 (0,5%)
RNA degradation	18 (0,4%)	5 (0,2%)	3 (0,6%)	3 (0,4%)
RNA polymerase	18 (0,4%)	3 (0,1%)	4 (0,8%)	1 (0,1%)
Arginine biosynthesis	17 (0,4%)	4 (0,2%)	1 (0,2%)	2 (0,2%)
Fatty acid degradation	17 (0,4%)	16 (0,7%)	1 (0,2%)	7 (0,8%)
Pantothenate and CoA biosynthesis	17 (0,4%)	13 (0,5%)	1 (0,2%)	4 (0,5%)
Pyrimidine metabolism	17 (0,4%)	5 (0,2%)	1 (0,2%)	2 (0,2%)
Sphingolipid metabolism	17 (0,4%)	10 (0,4%)	2 (0,4%)	5 (0,6%)
Various types of N-glycan biosynthesis	17 (0,4%)	8 (0,3%)	0 (0,0%)	2 (0,2%)
Viral life cycle - HIV-1	17 (0,4%)	12 (0,5%)	2 (0,4%)	4 (0,5%)
Vitamin B6 metabolism	17 (0,4%)	1 (0,0%)	4 (0,8%)	1 (0,1%)

Tabela 2. Número de sítios variantes com e sem expressão alelo-específica em transcritos presentes em cada uma das rotas metabólicas encontradas no banco de dados KEGG (continuação).

Rota metabólica	Sítios variantes		Transcritos	
	Sem expressão alelo-específica	Com expressão alelo-específica	Sem expressão alelo-específica	Com expressão alelo-específica
Nucleotide metabolism	16 (0,3%)	6 (0,2%)	2 (0,4%)	2 (0,2%)
Linoleic acid metabolism	15 (0,3%)	7 (0,3%)	0 (0,0%)	3 (0,4%)
Monoterpenoid biosynthesis	15 (0,3%)	5 (0,2%)	1 (0,2%)	2 (0,2%)
Alanine, aspartate and glutamate metabolism	14 (0,3%)	7 (0,3%)	0 (0,0%)	4 (0,5%)
Circadian rhythm - plant	14 (0,3%)	13 (0,5%)	0 (0,0%)	3 (0,4%)
Phagosome	14 (0,3%)	3 (0,1%)	4 (0,8%)	1 (0,1%)
Phenylalanine metabolism	14 (0,3%)	6 (0,2%)	0 (0,0%)	3 (0,4%)
Propanoate metabolism	14 (0,3%)	12 (0,5%)	1 (0,2%)	6 (0,7%)
Valine, leucine and isoleucine biosynthesis	14 (0,3%)	11 (0,5%)	0 (0,0%)	4 (0,5%)
Fructose and mannose metabolism	13 (0,3%)	4 (0,2%)	0 (0,0%)	2 (0,2%)
Nicotinate and nicotinamide metabolism	13 (0,3%)	4 (0,2%)	1 (0,2%)	2 (0,2%)
Pentose and glucuronate interconversions	13 (0,3%)	0 (0,0%)	5 (1,0%)	0 (0,0%)
Proteoglycans in cancer	13 (0,3%)	7 (0,3%)	1 (0,2%)	2 (0,2%)
SNARE interactions in vesicular transport	13 (0,3%)	2 (0,1%)	5 (1,0%)	1 (0,1%)
Glyoxylate and dicarboxylate metabolism	12 (0,3%)	4 (0,2%)	3 (0,6%)	2 (0,2%)
Tropane, piperidine and pyridine alkaloid biosynthesis	12 (0,3%)	4 (0,2%)	0 (0,0%)	2 (0,2%)
Fanconi anemia pathway	11 (0,2%)	2 (0,1%)	2 (0,4%)	2 (0,2%)
Isoquinoline alkaloid biosynthesis	11 (0,2%)	5 (0,2%)	0 (0,0%)	2 (0,2%)
Selenocompound metabolism	11 (0,2%)	10 (0,4%)	0 (0,0%)	2 (0,2%)
Ubiquinone and other terpenoid-quinone biosynthesis	11 (0,2%)	3 (0,1%)	2 (0,4%)	2 (0,2%)
Arachidonic acid metabolism	10 (0,2%)	5 (0,2%)	2 (0,4%)	2 (0,2%)
Butanoate metabolism	10 (0,2%)	8 (0,3%)	0 (0,0%)	3 (0,4%)
Inositol phosphate metabolism	10 (0,2%)	1 (0,0%)	3 (0,6%)	1 (0,1%)
Phosphatidylinositol signaling system	10 (0,2%)	2 (0,1%)	3 (0,6%)	2 (0,2%)
ABC transporters	9 (0,2%)	6 (0,2%)	2 (0,4%)	3 (0,4%)
Ether lipid metabolism	9 (0,2%)	6 (0,2%)	0 (0,0%)	2 (0,2%)
Lysine degradation	9 (0,2%)	7 (0,3%)	1 (0,2%)	2 (0,2%)
Steroid biosynthesis	9 (0,2%)	4 (0,2%)	1 (0,2%)	1 (0,1%)
Tryptophan metabolism	9 (0,2%)	7 (0,3%)	1 (0,2%)	3 (0,4%)
Fatty acid biosynthesis	8 (0,2%)	1 (0,0%)	2 (0,4%)	1 (0,1%)
Lipoic acid metabolism	8 (0,2%)	3 (0,1%)	1 (0,2%)	1 (0,1%)
Phosphonate and phosphinate metabolism	8 (0,2%)	2 (0,1%)	1 (0,2%)	1 (0,1%)
Porphyrin metabolism	8 (0,2%)	6 (0,2%)	0 (0,0%)	3 (0,4%)
Proteasome	8 (0,2%)	2 (0,1%)	1 (0,2%)	2 (0,2%)
Basal transcription factors	7 (0,1%)	1 (0,0%)	1 (0,2%)	1 (0,1%)

Tabela 2. Número de sítios variantes com e sem expressão alelo-específica em transcritos presentes em cada uma das rotas metabólicas encontradas no banco de dados KEGG (continuação).

Rota metabólica	Sítios variantes		Transcritos	
	Sem expressão alelo-específica	Com expressão alelo-específica	Sem expressão alelo-específica	Com expressão alelo-específica
Chemical carcinogenesis - reactive oxygen species	7 (0,1%)	2 (0,1%)	3 (0,6%)	1 (0,1%)
Chemical carcinogenesis - receptor activation	7 (0,1%)	2 (0,1%)	3 (0,6%)	1 (0,1%)
Ascorbate and aldarate metabolism	5 (0,1%)	6 (0,2%)	1 (0,2%)	3 (0,4%)
Base excision repair	6 (0,1%)	4 (0,2%)	1 (0,2%)	2 (0,2%)
Brassinosteroid biosynthesis	6 (0,1%)	0 (0,0%)	2 (0,4%)	0 (0,0%)
Flavonoid biosynthesis	6 (0,1%)	3 (0,1%)	0 (0,0%)	2 (0,2%)
Folate biosynthesis	6 (0,1%)	6 (0,2%)	0 (0,0%)	2 (0,2%)
Glycosaminoglycan degradation	6 (0,1%)	3 (0,1%)	1 (0,2%)	1 (0,1%)
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	6 (0,1%)	0 (0,0%)	2 (0,4%)	0 (0,0%)
Other glycan degradation	6 (0,1%)	3 (0,1%)	2 (0,4%)	1 (0,1%)
Sulfur metabolism	6 (0,1%)	3 (0,1%)	1 (0,2%)	2 (0,2%)
Caffeine metabolism	5 (0,1%)	4 (0,2%)	0 (0,0%)	1 (0,1%)
Histidine metabolism	5 (0,1%)	5 (0,2%)	0 (0,0%)	2 (0,2%)
Methane metabolism	5 (0,1%)	3 (0,1%)	0 (0,0%)	1 (0,1%)
Microbial metabolism in diverse environments	5 (0,1%)	3 (0,1%)	0 (0,0%)	1 (0,1%)
NOD-like receptor signaling pathway	3 (0,1%)	5 (0,2%)	0 (0,0%)	2 (0,2%)
Non-homologous end-joining	5 (0,1%)	1 (0,0%)	0 (0,0%)	1 (0,1%)
Protein export	3 (0,1%)	5 (0,2%)	1 (0,2%)	1 (0,1%)
Aldosterone synthesis and secretion	4 (0,1%)	4 (0,2%)	0 (0,0%)	1 (0,1%)
Amyotrophic lateral sclerosis	4 (0,1%)	1 (0,0%)	0 (0,0%)	1 (0,1%)
Biotin metabolism	4 (0,1%)	0 (0,0%)	2 (0,4%)	0 (0,0%)
Cutin, suberine and wax biosynthesis	4 (0,1%)	3 (0,1%)	1 (0,2%)	1 (0,1%)
Nitrogen metabolism	4 (0,1%)	4 (0,2%)	0 (0,0%)	2 (0,2%)
Other types of O-glycan biosynthesis	4 (0,1%)	2 (0,1%)	0 (0,0%)	1 (0,1%)
Pathways of neurodegeneration - multiple diseases	4 (0,1%)	1 (0,0%)	0 (0,0%)	1 (0,1%)
Photosynthesis	4 (0,1%)	2 (0,1%)	0 (0,0%)	1 (0,1%)
Retrograde endocannabinoid signaling	4 (0,1%)	4 (0,2%)	0 (0,0%)	1 (0,1%)
Thiamine metabolism	4 (0,1%)	4 (0,2%)	1 (0,2%)	1 (0,1%)
Glucosinolate biosynthesis	3 (0,1%)	2 (0,1%)	1 (0,2%)	2 (0,2%)
Glycosphingolipid biosynthesis - ganglio series	3 (0,1%)	3 (0,1%)	0 (0,0%)	1 (0,1%)
Betalain biosynthesis	2 (0,0%)	2 (0,1%)	0 (0,0%)	1 (0,1%)
Chemical carcinogenesis - DNA adducts	1 (0,0%)	2 (0,1%)	1 (0,2%)	1 (0,1%)
Drug metabolism - cytochrome P450	1 (0,0%)	2 (0,1%)	1 (0,2%)	1 (0,1%)
Drug metabolism - other enzymes	1 (0,0%)	2 (0,1%)	1 (0,2%)	1 (0,1%)

Tabela 2. Número de sítios variantes com e sem expressão alelo-específica em transcritos presentes em cada uma das rotas metabólicas encontradas no banco de dados KEGG (conclusão).

Rota metabólica	Sítios variantes		Transcritos	
	Sem expressão alelo-específica	Com expressão alelo-específica	Sem expressão alelo-específica	Com expressão alelo-específica
Fluid shear stress and atherosclerosis	1 (0,0%)	2 (0,1%)	1 (0,2%)	1 (0,1%)
Hepatocellular carcinoma	1 (0,0%)	2 (0,1%)	1 (0,2%)	1 (0,1%)
Lysosome	2 (0,0%)	1 (0,0%)	1 (0,2%)	1 (0,1%)
Metabolism of xenobiotics by cytochrome P450	1 (0,0%)	2 (0,1%)	1 (0,2%)	1 (0,1%)
Pathways in cancer	1 (0,0%)	2 (0,1%)	1 (0,2%)	1 (0,1%)
Platinum drug resistance	1 (0,0%)	2 (0,1%)	1 (0,2%)	1 (0,1%)
Riboflavin metabolism	2 (0,0%)	1 (0,0%)	0 (0,0%)	1 (0,1%)
Zeatin biosynthesis	2 (0,0%)	0 (0,0%)	1 (0,2%)	0 (0,0%)
Benzoxazinoid biosynthesis	0 (0,0%)	1 (0,0%)	0 (0,0%)	1 (0,1%)
D-Amino acid metabolism	1 (0,0%)	0 (0,0%)	1 (0,2%)	0 (0,0%)
Protein digestion and absorption	1 (0,0%)	0 (0,0%)	1 (0,2%)	0 (0,0%)
Renin-angiotensin system	1 (0,0%)	0 (0,0%)	1 (0,2%)	0 (0,0%)
Taurine and hypotaurine metabolism	0 (0,0%)	1 (0,0%)	0 (0,0%)	1 (0,1%)

Foram selecionados alguns transcritos para uma investigação mais detalhada da expressão diferencial alelo-específica nessas amostras. Observou-se que muitos deles apresentaram expressão exclusiva de apenas um dos alelos, apesar de ter ambas as cópias no genoma. Esse fenômeno pode estar associado com sítios variantes em regiões regulatórias em desequilíbrio de ligação com os sítios variantes observados em transcritos. Esses sítios que afetam a regulação das porções expressas do genoma, então, fariam com que um dos alelos não fosse transcrito ou com que a molécula transcrita fosse degradada rapidamente.

Na Figura 18, observam-se os valores de proporção relativa no DNA e de expressão relativa no RNA do alelo de referência em quatro SNVs presentes em um gene associado à formação do complexo trimérico da quinase relacionada a SNF1 (SnRK1). Este complexo está ligado à resposta ao estresse e ao metabolismo de carboidrato (WANG et al., 2017). Nota-se que nos tempos T1 a T3 de colmos de SP80-3280 houve uma tendência para a expressão do alelo alternativo do sítio variante encontrado na posição 1986, acentuada também em folhas deste mesmo genótipo. Não houve evidências de expressão alelo-específica nas raízes desse genótipo. Isso sugere que órgãos diferentes podem apresentar padrões de expressão alelo-específica diferentes. Nesse caso, tanto o alelo alternativo quanto o alelo de referência foram expressos, mas a expressão do alelo alternativo foi maior do que a explicada pelo seu número de cópias no genoma, sugerindo que há um mecanismo que favorece a expressão do alelo alternativo nesses órgãos de SP80-3280, sem interromper por completo a expressão do outro alelo.

Figura 18. Proporção no DNA e expressão no RNA do alelo de referência de SNVs em um gene ligado à formação do complexo trimérico da SnRK1 para cada genótipo, órgão e nível do fator estudado. A altura das barras vermelhas corresponde à proporção do alelo de referência nos dados de GBS, enquanto a altura das barras azuis, à proporção do alelo de referência nos dados de RNA-Seq. A intensidade das cores das barras representa o número de leituras de RNA-Seq para dado sítio em dada amostra. Barras com contorno preto representam sítios variantes em que foi rejeitada a hipótese nula de que a probabilidade de expressão do alelo de referência corresponde à proporção do alelo de referência no genoma.

A expressão de outro gene é analisada na Figura 19. Este transcrito está associado à atividade UDP-glicosiltransferase, relacionada a muitas funções em plantas, como resposta ao estresse (REHMAN et al., 2018). Ainda, essa família de proteínas também está associada à formação, alteração e degradação de paredes celulares (LIN et al., 2016). Outros estudos também identificaram a ocorrência de expressão alelo-específica em transcritos associados à resposta de defesa e à parede celular (MARGARIDO et al., 2022). Em colmos de F36-819 e SP80-3280 nos tempos T1 a T3 e no tempo T3 de IN84-58, há evidência de expressão alelo-específica estimulando a produção do alelo alternativo do sítio variante presente na posição 229. Nas folhas de SP80-3280 e de SES205A, no entanto, nota-se o padrão de expressão oposto: a maior parte dos transcritos vem do alelo de referência. Esses resultados mostram que também há diferença na expressão alelo-específica entre órgãos, não apenas entre genótipos.

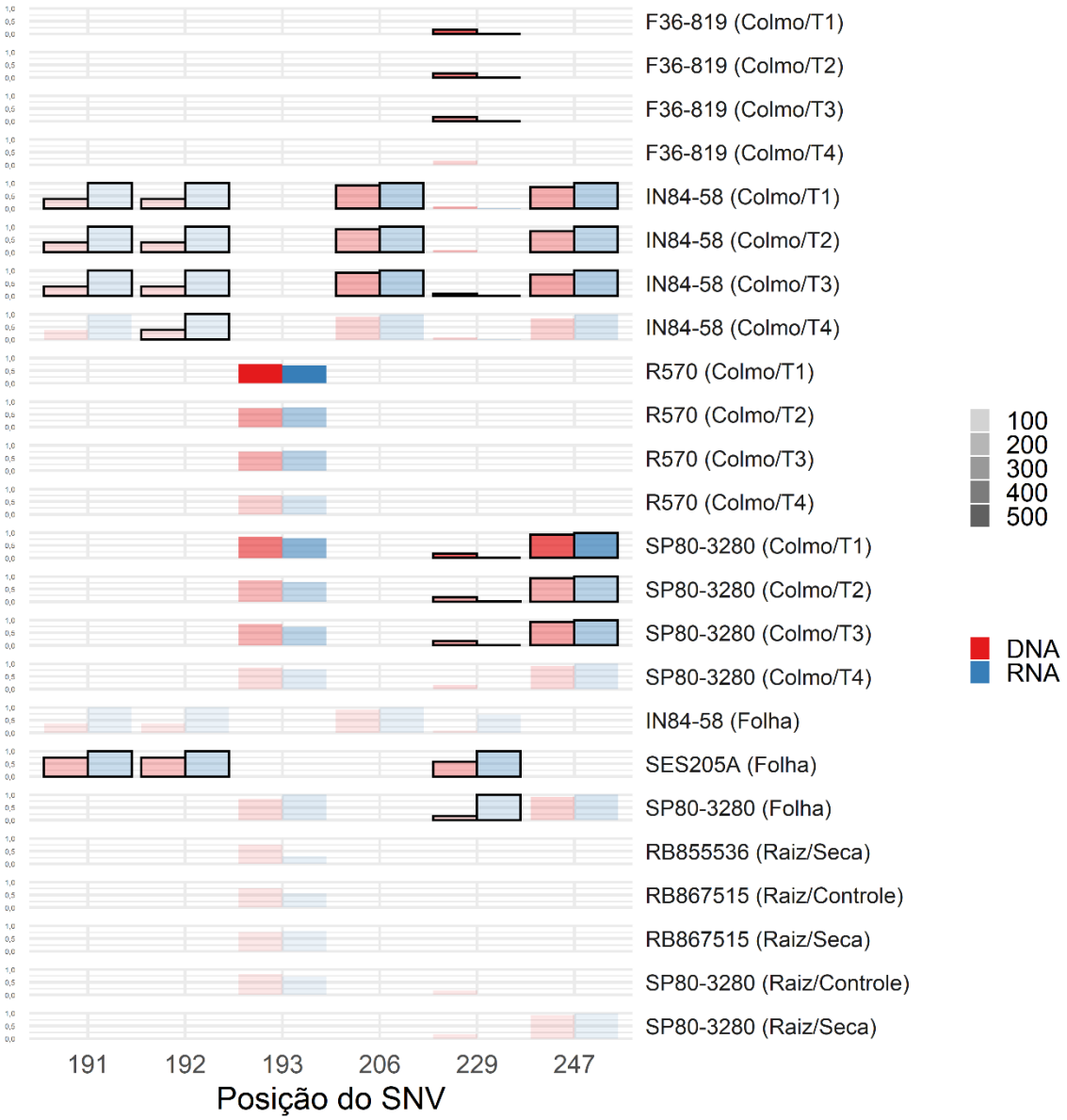

Figura 19. Proporção no DNA e expressão no RNA do alelo de referência de SNVs em um transcrito relacionado à atividade UDP-glicosiltransferase para cada genótipo, órgão e nível do fator estudado. A altura das barras vermelhas corresponde à proporção do alelo de referência nos dados de GBS, enquanto a altura das barras azuis, à proporção do alelo de referência nos dados de RNA-Seq. A intensidade das cores das barras representa o número de leituras de RNA-Seq para dado sítio em dada amostra. Barras com contorno preto representam sítios variantes em que foi rejeitada a hipótese nula de que a probabilidade de expressão do alelo de referência corresponde à proporção do alelo de referência no genoma.

Na Figura 20, observam-se as proporções de SNVs em um transcrito relacionado à proteína DETOXIFICATION 40. Nota-se a diferença na magnitude da expressão alelo-específica em diferentes tempos do desenvolvimento no SNV da posição 238 de SP80-3280. Além disso, na posição 379 em colmos de IN84-58, nota-se que houve uma maior expressão do alelo alternativo no tempo T4, diferentemente do que aconteceu nos tempos T1 a T3.

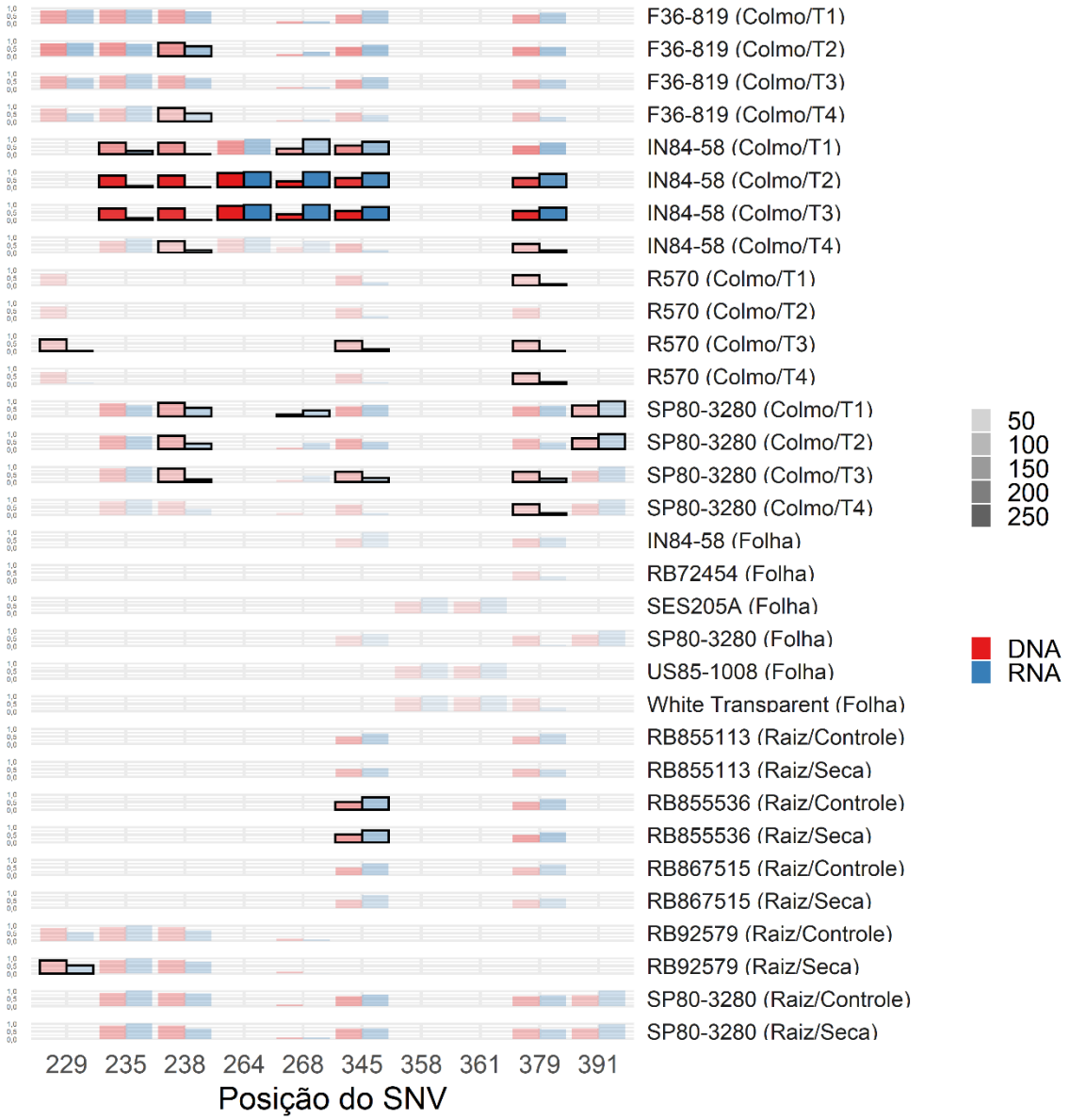

Figura 20. Proporção no DNA e expressão no RNA do alelo de referência de SNVs em um transcrito relacionado à proteína DETOXIFICATION 40 para cada genótipo, órgão e nível do fator estudado. A altura das barras vermelhas corresponde à proporção do alelo de referência nos dados de GBS, enquanto a altura das barras azuis, à proporção do alelo de referência nos dados de RNA-Seq. A intensidade das cores das barras representa o número de leituras de RNA-Seq para dado sítio em dada amostra. Barras com contorno preto representam sítios variantes em que foi rejeitada a hipótese nula de que a probabilidade de expressão do alelo de referência corresponde à proporção do alelo de referência no genoma.

Na Figura 21 observam-se os dados de um transcrito relacionado a uma proteína que faz o transporte transmembrana de íons de Fe(2+) e zinco. Em raízes de RB855536, RB92579 e SP79-1011, nota-se que houve uma diminuição da expressão do alelo alternativo do SNV presente na posição 1104 em condição de seca. Isso mostra que além de associada às diferenças fenotípicas entre genótipos, ao desenvolvimento e à diferenciação de tecidos, a expressão alelo-específica também pode estar presente na resposta a estímulos e ao estresse.

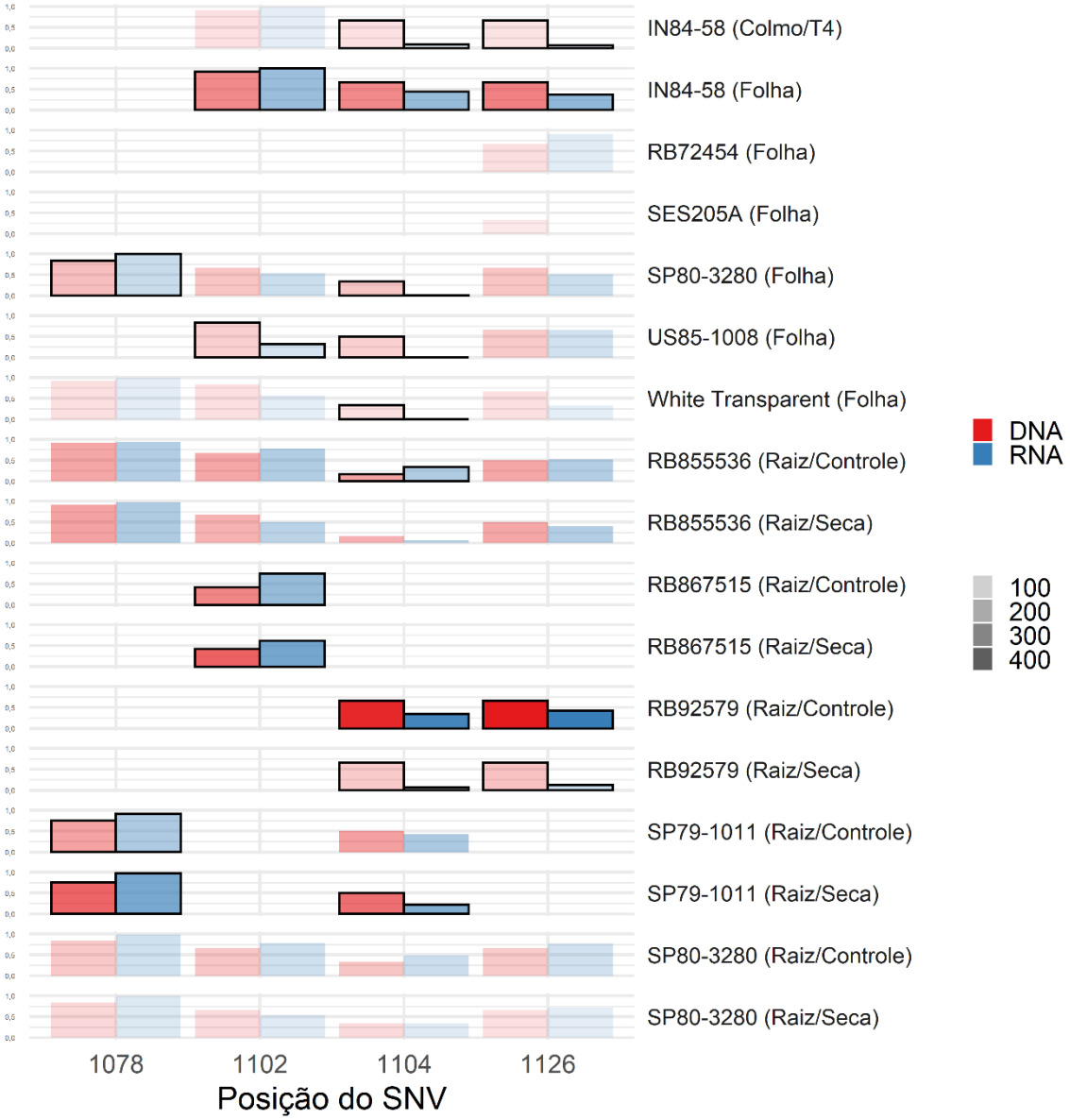

Figura 21. Proporção no DNA e expressão no RNA do alelo de referência de SNVs em um transcrito relacionado ao transporte transmembrana de íons de Fe(2+) e zinco para cada genótipo, órgão e nível do fator estudado. A altura das barras vermelhas corresponde à proporção do alelo de referência nos dados de GBS, enquanto a altura das barras azuis, à proporção do alelo de referência nos dados de RNA-Seq. A intensidade das cores das barras representa o número de leituras de RNA-Seq para dado sítio em dada amostra. Barras com contorno preto representam sítios variantes em que foi rejeitada a hipótese nula de que a probabilidade de expressão do alelo de referência corresponde à proporção do alelo de referência no genoma.

Na Figura 22 observa-se a expressão comparada à proporção no genoma de alelos em um gene relacionado à proteína serina/treonina quinase NeK5. Observa-se que órgãos dos genótipos SP80-3280 e F36-819 tendem a expressar o alelo de referência do sítio variante presente na posição 1502, mesmo tendo menos cópias deste no genoma. Por outro lado, órgãos do genótipo IN84-58 expressam mais transcritos contendo o alelo alternativo, mesmo tendo um número maior de cópias do alelo de referência no genoma. Uma maior expressão de um alelo distinto nos transcritos de IN84-58 quando comparado aos outros genótipos também pode ser observada na posição 297 de um transcrito relacionado ao metabolismo de xiloglucano (Figura 23). Além disso, observa-se uma diferença na magnitude da expressão alelo-específica na posição 297 em colmos de IN84-58 ao longo do tempo.

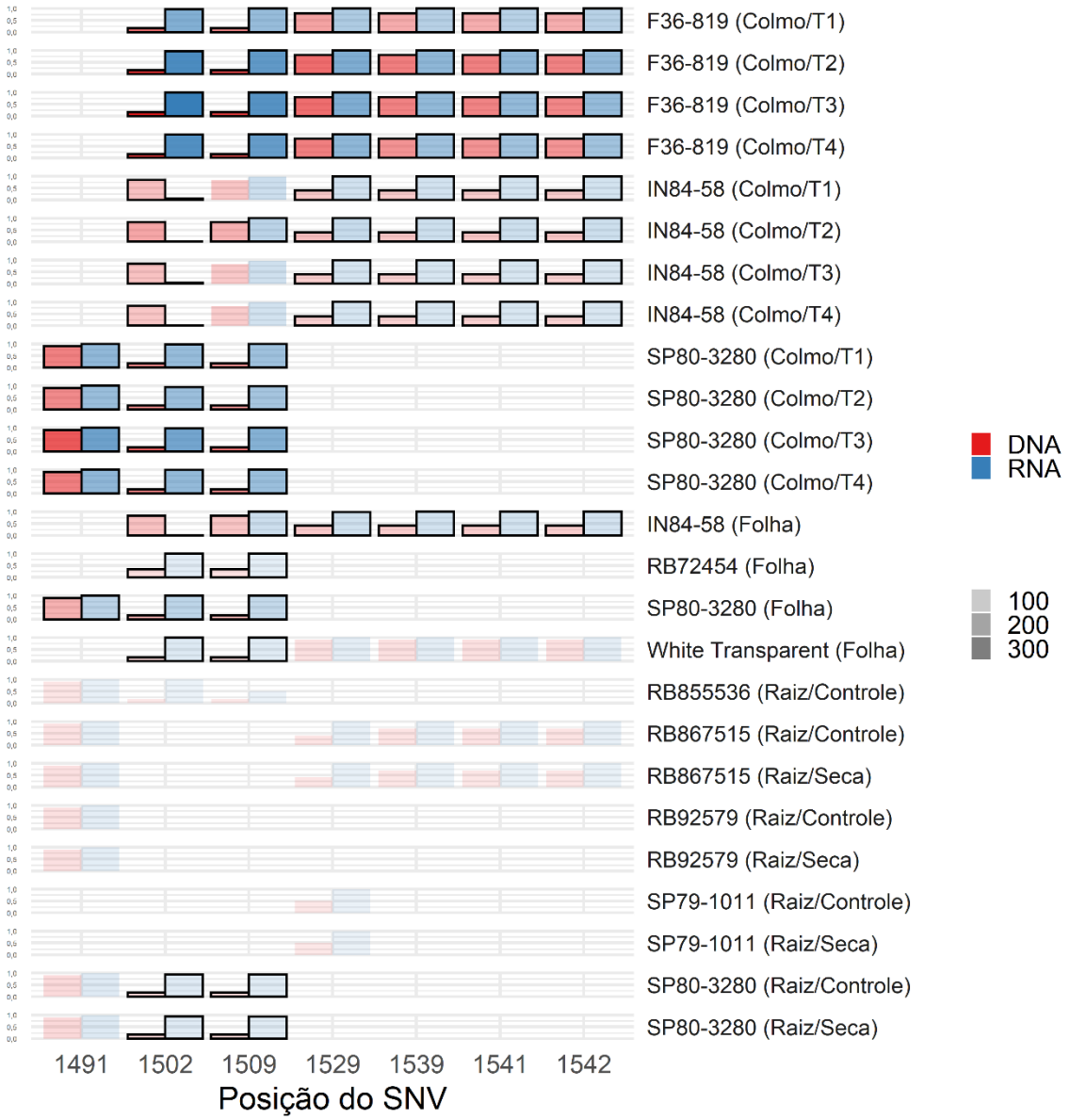

Figura 22. Proporção no DNA e expressão no RNA do alelo de referência de SNVs em um transcrito relacionado à proteína serina/treonina quinase NeK5 para cada genótipo, órgão e nível do fator estudado. A altura das barras vermelhas corresponde à proporção do alelo de referência nos dados de GBS, enquanto a altura das barras azuis, à proporção do alelo de referência nos dados de RNA-Seq. A intensidade das cores das barras representa o número de leituras de RNA-Seq para dado sítio em dada amostra. Barras com contorno preto representam sítios variantes em que foi rejeitada a hipótese nula de que a probabilidade de expressão do alelo de referência corresponde à proporção do alelo de referência no genoma.

Figura 23. Proporção no DNA e expressão no RNA do alelo de referência de SNVs em um transcrito relacionado ao metabolismo de xiloglucano para cada genótipo, órgão e nível do fator estudado. A altura das barras vermelhas corresponde à proporção do alelo de referência nos dados de GBS, enquanto a altura das barras azuis, à proporção do alelo de referência nos dados de RNA-Seq. A intensidade das cores das barras representa o número de leituras de RNA-Seq para dado sítio em dada amostra. Barras com contorno preto representam sítios variantes em que foi rejeitada a hipótese nula de que a probabilidade de expressão do alelo de referência corresponde à proporção do alelo de referência no genoma.

2.4 Conclusão

Observou-se baixa ocorrência de expressão alelo-específica nesse conjunto de dados de cana-de-açúcar. Ainda, nos sítios variantes em que foi observado esse fenômeno, houve pouca consistência em diferentes genótipos, órgãos, tempos e condições de estresse hídrico. Os padrões de expressão alelo-específica para determinado sítio diferem principalmente entre genótipos e menos entre órgãos e entre tempos. As menores diferenças nos padrões de expressão alelo-específica foram encontradas comparando-se as raízes de um mesmo genótipo nas condições de seca e controle. Observou-se um maior número de genes exclusivos com expressão alelo-específica nos genótipos em IN84-

58 e SP80-3280, mas houve mais diferenças dos genótipos representantes de *S. spontaneum* (SES205A e IN84-58) em relação aos demais do que entre os outros genótipos. Grande parte dos transcritos de IN84-58 e SP80-3280 apresentaram expressão alelo-específica apenas em um subconjunto de condição, tempo ou órgão, mas não em outros casos, reforçando a baixa consistência desse fenômeno nesse conjunto de dados. Foram encontrados genes com expressão diferencial alelo-específica associados a diferenças entre genótipos, órgãos, tempos durante o desenvolvimento das plantas e entre as condições de seca e controle. Devido a limitações da técnica de GBS, foram enriquecidos termos GO bastante gerais e não foi encontrada qualquer rota metabólica KEGG enriquecida para os genes com expressão alelo-específica. Ainda, foi encontrado um viés favorecendo sítios variantes que expressam mais cópias do alelo de referência. Este foi o primeiro estudo em que foi feita uma análise genômica da expressão alelo-específica em cana-de-açúcar, estudando-se múltiplos fatores, como genótipos, órgãos, tempos e condições de estrese hídrico. Apesar das limitações, os resultados apresentados iluminam o caminho para uma maior compreensão deste fenômeno nesse grupo de organismos poliploides complexos.

Referências

ANDREWS, S. et al. **FastQC: a quality control tool for high throughput sequence data**. , 2010.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 57, n. 1, p. 289–300, jan. 1995.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 1 ago. 2014.

BRYANT, D. M. et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. **Cell Reports**, v. 18, n. 3, p. 762–776, jan. 2017.

CAI, M. et al. Allele specific expression of Dof genes responding to hormones and abiotic stresses in sugarcane. **PLOS ONE**, v. 15, n. 1, p. e0227716, 16 jan. 2020.

CORRER, F. H. et al. Time-series expression profiling of sugarcane leaves infected with Puccinia kuehnii reveals an ineffective defense system leading to susceptibility. **Plant Cell Reports**, v. 39, n. 7, p. 873–889, 20 jul. 2020a.

CORRER, F. H. et al. Differential expression in leaves of Saccharum genotypes contrasting in biomass production provides evidence of genes involved in carbon partitioning. **BMC Genomics**, v. 21, n. 1, p. 673, 29 dez. 2020b.

CORRER, F. H. et al. Allele expression biases in mixed-ploid sugarcane accessions. **Scientific Reports**, 2022.

FAOSTAT. **Crops and livestock products**. Disponível em: <https://www.fao.org/faostat/en/#data/QCL>. Acesso em: 11 maio. 2022.

GARSMEUR, O. et al. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nature Communications, v. 9, n. 1, p. 2638, 6 dez. 2018.

GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, v. 29, n. 7, p. 644–652, 15 jul. 2011.

HOSAKA, G. K. et al. Temporal Gene Expression in Apical Culms Shows Early Changes in Cell Wall Biosynthesis Genes in Sugarcane. Frontiers in Plant Science, v. 12, 13 dez. 2021.

KIM, C. et al. Comparative Analysis of Miscanthus and Saccharum Reveals a Shared Whole-Genome Duplication but Different Evolutionary Fates. The Plant Cell, v. 26, n. 6, p. 2420–2429, jun. 2014.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. Nature Methods, v. 9, n. 4, p. 357–359, 4 abr. 2012.

LIN, J.-S. et al. UDP-glycosyltransferase 72B1 catalyzes the glucose conjugation of monolignols and is essential for the normal cell wall lignification in *Arabidopsis thaliana*. The Plant Journal, v. 88, n. 1, p. 26–42, out. 2016.

MARGARIDO, G. R. A. et al. Limited allele-specific gene expression in highly polyploid sugarcane. Genome Research, v. 32, n. 2, p. 297–308, fev. 2022.

PEREIRA, G. S.; GARCIA, A. A. F.; MARGARIDO, G. R. A. A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. BMC Bioinformatics, v. 19, n. 1, p. 398, 1 dez. 2018.

PIPERIDIS, N.; D'HONT, A. Sugarcane genome architecture decrypted with chromosome-specific oligo probes. The Plant Journal, v. 103, n. 6, p. 2039–2051, 12 set. 2020.

R CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna, AustriaR Foundation for Statistical Computing, , 10 mar. 2022.

REHMAN, H. M. et al. Comparative genomic and transcriptomic analyses of Family-1 UDP glycosyltransferase in three Brassica species and Arabidopsis indicates stress-responsive regulation. Scientific Reports, v. 8, n. 1, p. 1875, 30 dez. 2018.

SERANG, O.; MOLLINARI, M.; GARCIA, A. A. F. Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. PLoS ONE, v. 7, n. 2, p. e30906, 17 fev. 2012.

SFORÇA, D. A. et al. Gene Duplication in the Sugarcane Genome: A Case Study of Allele Interactions and Evolutionary Patterns in Two Genic Regions. Frontiers in Plant Science, v. 10, 7 maio 2019.

SOUZA, G. M. et al. Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. GigaScience, v. 8, n. 12, 1 dez. 2019.

VAN DER AUWERA, G. A.; O'CONNOR, B. D. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. [s.l.] O'Reilly Media, Inc., 2020.

VILELA, M. DE M. et al. Analysis of three sugarcane homo/homeologous regions suggests independent polyploidization events of *Saccharum officinarum* and *Saccharum spontaneum*. Genome Biology and Evolution, p. evw293, 12 jan. 2017.

WANG, F. et al. A sucrose non-fermenting-1-related protein kinase 1 gene from potato, StSnRK1, regulates carbohydrate metabolism in transgenic tobacco. Physiology and Molecular Biology of Plants, v. 23, n. 4, p. 933–943, 16 out. 2017.

YOUNG, M. D. et al. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biology, v. 11, n. 2, p. R14, 2010.

ZHANG, J. et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nature Genetics, v. 50, n. 11, p. 1565–1573, 8 nov. 2018.

ZHAO, C. et al. Quantification of allelic differential expression using a simple Fluorescence primer PCR-RFLP-based method. Scientific Reports, v. 9, n. 1, p. 6334, 19 dez. 2019.