EULÁLIO GUTEMBERG DIAS DOS SANTOS

ANÁLISE DOS MECANIMOS DE *SPLICING* ALTERNATIVO EM RESPOSTA AO DÉFICIT HÍDRICO EM SOJA (*Glycine max*)

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Aplicada, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2019

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T

S237a
2019

Santos, Eulálio Gutemberg Dias dos, 1991-

Análise dos mecanismos de *splicing* alternativo em resposta ao déficit hídrico em soja (*Glycine max*) / Eulálio Gutemberg Dias dos Santos. – Viçosa, MG, 2019.

xv, 80 f. : il. (algumas color.) ; 29 cm.

Inclui apêndice.

Orientador: Humberto Josué de Oliveira Ramos.

Dissertação (mestrado) - Universidade Federal de Viçosa.

Referências bibliográficas: f. 50-58.

1. Soja - Melhoramento genético. 2. Sequênciamento de nucleotideo. 3. Déficit hídrico. 4. Engenharia genética. I. Universidade Federal de Viçosa. Departamento de Bioquímica e Biologia Molecular. Programa de Pós-Graduação em Bioquímica Aplicada. II. Título.

CDD 22. ed. 633.342

EULÁLIO GUTEMBERG DIAS DOS SANTOS

ANÁLISE DOS MECANIMOS DE *SPLICING* ALTERNATIVO EM RESPOSTA AO DÉFICIT HÍDRICO EM SOJA (*Glycine max*)

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Aplicada, para obtenção do título de *Magister Scientiae*.

APROVADA: 29 de julho de 2019.

Pedro Marcus Pereira Vidigal

Otávio José Bernardes Brustolini

Camilo Elber Vital

Humberto Josué de Oliveira Ramos
(Orientador)

A todos aqueles que, de alguma forma, estiveram e estão próximos de mim, me dando forças e torceram pela minha recuperação, assim fazendo esta vida valer cada vez mais a pena.

AGRADECIMENTOS

A Deus por ter me dado mais uma chance de estar vivo, saúde e força para superar os obstáculos que encontrei em dias. Agradeço a minha mãe Danisete, heroína que me deu apoio, incentivo nas horas difíceis, de desânimo e cansaço. Ao meu pai, que apesar de todas as dificuldades me fortaleceu e que para mim foi muito importante. Ao meu irmão Breno pela força em todos os momentos bons e ruins, e a A minha namorada Taina por ser companheira em todos os momentos, nesses dois anos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Sendo assim, agradeço a CAPES pela oportunidade concedida, a esta universidade, seu corpo docente, direção e administração que oportunizaram a janela que hoje vislumbro um horizonte superior, eivado pela acendrada confiança no mérito e ética aqui presentes, enfatizando ao LBMP e ao Núcleo de Biomoléculas. Em especial ao meu orientador Humberto Ramos, pelo emprenho dedicado à elaboração deste trabalho e disposto a ajudar sempre que possível. A alguns colegas de trabalho que foram de fundamental importância na elaboração dessa dissertação como, Pedro, Flaviana, Otávio e Cleisinho que sempre foram solícitos em ajudar.

Meus agradecimentos aos diversos amigos e companheiros pela amizade e que fizeram parte da desta formação e que vão continuar presentes em minha vida com certeza. A todos que direta ou indiretamente fizeram parte da minha formação como pessoa e futuro profissional, o meu muito obrigado.

RESUMO

SANTOS, Eulalio Gutemberg Dias dos., M.Sc., Universidade Federal de Viçosa, julho de 2019. **Análise dos mecanismos de *splicing* alternativo em resposta ao déficit hídrico em soja (*Glycine max*)**. Orientador: Humberto Josué de Oliveira Ramos.

A soja [Glycine max (L.) Merr.] hoje em dia é uma das culturas de maior importância para o agronegócio mundial. O Brasil se posiciona como o segundo maior produtor mundial desta cultura, com mais de 30% da produção, atrás apenas dos Estados Unidos. Porém, nos últimos anos, a soja vem sofrendo grandes perdas em função das adversidades climáticas, apresentando enorme sensibilidade a alterações de frio, calor, chuvas e principalmente escassez hídrica, responsável pelas maiores perdas, impedindo uma maior produtividade e expansão agrícola para outras regiões. A solução para essas perdas seria o desenvolvimento de técnicas mais eficazes de plantio e produção, bem como o desenvolvimento de variedades genotípicas mais tolerantes à escassez hídrica. Dessa forma, nosso principal objetivo é a busca pela melhor elucidação de mecanismos e vias relacionadas ao estresse hídrico por predição em bioinformática analisando o transcriptoma após estresse. Portanto, estudos transcriptômicos oferecem uma visão dos mecanismos de resposta ao estresse das plantas. Aqui, apresentamos os resultados do perfil global de expressão gênica de folhas de dois cultivares de soja, BR16 e EMBRAPA48, com capacidade contrastante para lidar com o déficit hídrico, utilizando metodologia de sequenciamento de RNA. Para as análises iniciais do transcriptoma, a qualidade das sequências foi avaliada com o programa FastQC. Posteriormente, nós realizamos o alinhamento contra o genoma de referência da soja usando o alinhador Bowtie/TopHat, em seguida avaliado o perfil de expressão gênica pelo pacote Cufflinks/Cuffdiff, que mostraram uma gama de genes que foram alterados em função do estresse, com uma perturbação mais branda para a variedade mais tolerante EMBRAPA48. Nossa análise sugere que a tolerância à seca resulta de um certo nível de transcrição de genes que predispõem a planta mesmo antes do início da seca. No *splicing* alternativo diferentes éxons e íntrons de um mesmo pré- RNA mensageiro podem ser excluídos ou retidos para produzir diferentes RNAs maduros, expandindo o repertório do transcriptoma. Análises de expressão diferencial em nível de transcritos podem detectar alterações na expressão de isoformas do transcrito, a partir de um mesmo gene. Portanto, no presente estudo, nós avaliamos a ocorrência de

splicing alternativo diferencial. Inicialmente realizamos o alinhamento contra o genoma de referência de Soja usando o alinhador *splice-aware* STAR. Após o alinhamento, o programa rMATS foi usado para detectar os principais tipos de splicing alternativo. Em nossos resultados mostramos que o Alternativo 3' e Skipped exon foram os eventos de maior ocorrência para ambas variedades. Nossos dados sugerem que eventos de splicing alternativo são diferencialmente regulados em consequência da submissão ao estresse hídrico e revelam um mecanismo potencial na regulação da expressão de genes com importantes funções na tolerância ao estresse. Diante das análises de Ontologia revelou que os genes DAS estavam envolvidos principalmente em processos biológicos, incluindo resposta celular ao estresse como a biossíntese de polissacarídeos de membrana (dentre as quais se encontram a biossíntese de glicanos, betaglicanos e UDP-raminose), reparo da região 3' de DNA, transporte de auxina, regulação do desenvolvimento floral e resposta ao estímulo abiótico como manutenção do potencial hídrico das folhas. Dessa forma a avaliação da expressão de genes e do *splicing* alternativo do transcriptoma de soja para as variedades contrastantes nos fornecem entendimento de como ocorre a regulação da expressão diferencial bem como a descrição de mecanismos e vias para tolerância ao estresse hídrico.

ABSTRACT

SANTOS, Eulalio Gutemberg Dias dos., M.Sc., Universidade Federal de Viçosa, July 2019. **Mechanisms analysis of alternative *splicing* in response to drougth stress in soybean (*Glycine max*)** Orientador: Humberto Josué de Oliveira Ramos.

Soybean [Glycine max (L.) Merr.] is one of the most important cultures for agribusiness worldwide. Brazil is the second largest grower of this culture, with over 30% of its production, surpassed only by the United States. However, in the past few years, soybean has suffered great losses due to climate adversities, presenting enormous sensibility due to changes in heat, cold, rain and mostly hydric shortage, responsible for the largest losses. This context prevents more productivity and breaks agricultural expansion to other regions. The solution for this losses is the development of more effective techniques for growth and production of soybean, as well as the development of genotypic varieties with more tolerance to hydric shortage. Thus, our main goal is the elucidation of mechanisms and pathways related to hydric stress by bioinformatics prediction, analysing the transcriptome after stress. Therefore, transcriptomic studies offer a view of response mechanisms to plant stress. This study presents the global profile results of gene expression of two soybean types, BR16 and EMBRAPA48, with contrastant capacities to deal with hydric deficit, using RNA sequence methodology. For transcriptome initial analysis, the quality of the sequences was evaluated with FastQC software. Afterwards, the alignment against the soybean reference genome was made using the Bowtie/TopHat aligner, followed by the evaluation of the gene expression profile using Cufflinks/Cuffdiff package. The results showed an amount of genes that were altered due to stress, with a milder disturbance in the more tolerant variety, EMBRAPA48. The analysis suggests that the drought tolerance occurs due to a certain level of gene transcription that predisposes the plant even before the beginning of the drought. In alternative splicing different exons and introns of one same messenger pre-RNA may be excluded or retained for the production of different mature RNAs, expanding the repertoire of the transcriptome. Differential expression analysis in transcripts level can detect alterations in transcripts isoforms from one specific gene. Therefore, in the present study, we evaluate the occurrence of differential alternative splicing. Initially, the alignment against the soybean reference genome was made using the *splice-aware* STAR aligner. After the alignment, rMATS software was used to detect the main types of alternative splicing. The results show that Alternative 3' and

Skipped exon were the events of major occurrence for both varieties. The data suggests that the alternative splicing events are differentially regulated due to hydric stress submission and reveal a potential mechanism for gene expression regulation of important functions in stress tolerance. Ontology analysis revealed that the DAS genes were involved mainly in biological processes, including cellular response to stress, such as the biosynthesis of membrane polysaccharides (glycan, betaglycan, UDP-raminose, among others), DNA repair in the 3’ region, auxin transportation, flower development regulation and response to abiotic stimulation as maintenance of the leaves hydric potential. Thus, the evaluation of gene expression and alternative splicing of the soybean transcriptome for contrastant varieties provides the understanding of the differential expression and the description of mechanisms and pathways for hydric stress tolerance.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ABA– Ácido Abscísico

AS– *Alternative Splicing*

A3SS – *Alternative 3' splicing site*

A5SS – *Alternative 5' splicing site*

BP – Processo Biológico

CC – Componente Celular

DAS – *Diferencial alternative splicing*

DEGs – *Diferencial expression genes*

DNA – Ácido desoxirribonucleico

cDNA – DNA complementar

ESTs – Sequenciamento de sequências expressas

FC – *Fold change*

FDR – *False Discovery Rate*

FPKM – *Fragments Per Kilobase Million*

GPB – Giga pares de bases

IR – Irrigado

Log – Logaritmo

rMATS – *Multivariate Analysis of Transcript Splicing*

MF – Função Molecular

MXE – *Mutualy exclusive exons*

NGS – *Next Generation Sequence*

NI – Não irrigado

NMD – *Nonsense Mediated Decay*

PB – *Pair base*

PCR – *Polimerase Chain reaction*

qPRC – *Quantitative Polimerase Chain reaction*

Q – Qualidade

RI – *Retained intron*

RNA – Ácido ribonucleico

mRNA – Ácido ribonucleico mensageiro

RNA-seq – RNA *sequencing*

SE – *Skipped exon*

SNP – Polimorfismo de nucleotídeo único

SSH – Supressão de hibridização subtrativa

STAR – *Splicingd Transcripts Alignment to a Reference*

TFs – Fatores de transição

TSS – *Transcriptional Start Sites*

UDP – *Uridine diphosphate*

UTR – Região não traduzida

SUMÁRIO

1. INTRODUÇÃO

1.1 Soja inserida na economia

A soja [*Glycine max (L.) Merrill*] é originária de clima temperado com ampla adaptação nos climas subtropicais e tropicais. Nas primeiras três décadas da produção de soja no século XX, todo o cultivo era realizado no Oriente, sendo a China, Indonésia, Japão e Coréia os principais produtores. No entanto, no final da década de 1940 e início dos anos 50, os EUA ultrapassaram a China e eventualmente todo o Oriente na produção de soja. Em 1968, aproximadamente 28 milhões de hectares de soja foram semeados em quase 25 países. Hoje a cultura da soja é considerada a mais importante do agronegócio mundial, movimentando em 2018 cerca de 31,7 bilhoes de dólares com uma produção de 362,075 milhões de toneladas em uma área de 125,691 milhoes de hectares na safra 2018/2019 (AGROSTAT, 2019; USDA, 2019).

O Brasil é, atualmente, o segundo maior produtor mundial de soja e esta cultura é também o principal produto agrícola dentro do agronegócio brasileiro. A safra 2018/2019 produziu aproximadamente 114,843 milhões de toneladas de grãos, em ma área plantada de 35,822 milhões de hectares, com produtividade de 3,206 kg/ha (CONAB, 2019), responsável por cerca de 25% da produção e exportação deste grão (ABIOVE 2015, FAOSTAT 2015). Em função da boa aceitação das novas tecnologias por parte dos produtores, associada ao relevante esforço dos programas de melhoramento de soja, a produtividade da cultura tem aumentado consideravelmente. Entretanto, esse aumento é condicionado e influenciado por oscilações climáticas (TSUKAHARA et al., 2016).

Um dos grandes problemas que os agricultores encontram é a confirmação de genótipos tolerantes à escassez de água, que limita a produtividade em certas regiões. Eventos que levam à seca têm aumentado consideravelmente em função das mudanças climáticas no mundo (STOKSTAD, 2004; SCHIERMEIER, 2006). Nos estados do sul do Brasil, que são responsáveis por 40% da produção interna, as perdas foram de até 25% do total da produção nas safras recentes, (Vidal et al., 2012 e Rodrigues et al., 2012. Na safra 2015/16), os estresses abióticos foram os principais responsáveis pela quebra de produção, especialmente pelas altas temperaturas e pelos períodos de estresse hídrico.

O efeito do defict hídrico na produção depende da época de ocorrência e de sua severidade. O desenvolvimento de cultivares mais tolerantes a períodos de déficit hídrico, por meio de tecnologias relacionadas a genética e a biologia das plantas, auxilia a cultura a suportar períodos prolongados de estiagem, sendo essenciais na manutenção da produção agrícola.

A tolerância à seca é uma característica complexa desenvolvida por meio de mecanismos que funcionam em conjunto ou isoladamente para tolerar períodos de déficit hídrico (CASAGRANDE et al., 2001) e possui uma base molecular genética para todas as alterações fisiológicas, morfológicas e de desenvolvimento sobre as plantas. Dessa forma, a caracterização de genótipos tolerantes ou sensíveis à seca se tornou um pré-requisito para seleção e manipulação genética (TURNER, 1997; CASAGRANDE et al., 2001). Como não há uma classificação precisa, bem como regras quanto a tolerância à seca, foi elaborada uma classificação de cunho prático com o objetivo de estabelecer uma relação de tolerância (PITOL e BROCH, 2008).

A tabela abaixo relata alguns cultivares de soja e sua classificação quanto a tolerância à seca. Portanto, genótipos que diferem em tolerância ao déficit hídrico devem apresentar diferenças qualitativas e quantitativas em expressão gênica. Logo, compreender como esses eventos são ativados/desativados e como interagem entre si torna-se essencial no desenvolvimento de novas variedades de soja mais tolerantes a períodos de seca (CASAGRANDE et al., 2001).

Tabela 1 - Classificação das cultivares de soja quanto a tolerância ao déficit hídrico do solo.

TOLERANTE	MODERADAMENTE TOLERANTE	SUSCETÍVEL	ALTAMENTE SUSCETÍVEL
BRS 239 EMBRAPA 48	BR 16 BRS 241 BRS 268 BRS 282 CD 202 Fundacap 59 RR FTS Campo Mourão RR JB 101 M-Soy 8001 Vmax NK 7059 RR	BRS 133 BRS 181 BRS 232 BRS 245 RR BRS 246 RR BRS Charrua RR BRS Favorita RR BRS MG 68 – Vencedora MG/BR 46 (Conquista) CD 214 RR CD 219 CD 225 RR CD 226 RR 3MX Titan RR M-Soy 7908 RR Don Mario 7 Oi RR BMX Potência RR	BRS 244 RR BRS 247 RR CD 201 CD 205 CD 208 CD 213 RR

Fonte: FUNDAÇÃO MS, 2008

1.2 Escassez Hídrica

O estresse é uma condição fisiológica alterada causada por fatores que tendem a romper o equilíbrio. A tensão é qualquer mudança física e química produzida por um estresse (GASPAR et al., 2002). A flexibilidade do metabolismo permite a iniciação da resposta às mudanças ambientais, que flutuam regularmente e são previsíveis em ciclos diários e sazonais. O estresse é uma restrição ou flutuações altamente imprevisíveis impostas a alterações metabólicas regulares, padrões que causam lesão, doença ou fisiologia aberrante. Enquanto crescem na natureza, as plantas são frequentemente expostas a muitos estresses, como a seca, baixa temperatura, salinidade, inundação, calor, estresse oxidativo e toxicidade de metais pesados.

A seca é um risco natural relacionado a uma falta prolongada de chuvas que leva a uma diminuição temporária ou déficit na disponibilidade de água natural (VOGT E SOMMA, 2000; SPINONI et al. 2017), o que leva a condições atmosféricas alteradas e perda de água em função da evapotranspiração (JALEEL et al., 2007). A seca prolongada impede o desenvolvimento das plantas, alterando consistentemente a fisiologia e metabolismo do vegetal (JALEEL et al., 2007; NAKAYAMA et al., 2007). No entanto, as plantas evoluíram a fim de criar estratégias para lidar com a seca, incluindo um curto ciclo de vida que é uma forma de escape aos períodos de seca, maior absorção de água e redução da perda de água que é uma estratégia afim de evitar a seca, bem como ajuste osmótico, capacidade antioxidante o que confere tolerância à dessecação (FANG E XIONG, 2015). Dessa forma, a obtenção de genótipos mais tolerantes é fundamental para contornar a escassez hídrica e manutenção da produtividade agrícola e principalmente, evitar perdas na produção do agronegócio (CATTIVELLI et al., 2008). Apesar de mecanismos naturais terem favorecido a adaptação e a sobrevivência de alguns tipos de plantas, estudos de fisiologia molecular têm fornecido significativo ganho no entendimento de respostas fisiológicas e moleculares de plantas ao déficit hídrico.

A complexidade dos mecanismos de tolerância à seca explica o vagaroso processo do melhoramento genético visando a produtividade em ambientes propícios ao recorrer este tipo condição adversa (TUBEROSA e SALVI, 2006; RODRIGUES et al. 2012, BROWN E HUDSON, 2017). Entre esses mecanismos, estão aqueles que codificam proteínas quinases dependentes de cálcio, calmodulina e proteínas de cálcio relacionadas com calmodulina e fosfatases de proteína classe 2C (MOLINA et al. 2008; GUO et al. 2009; RANJAN E SAWANT, 2015), juntamente com um número de fatores de transcrição (TFs) (SAHOO et al. 2013, JANIAK et al. 2016). Em condições de estresse severo, as alterações fisiológicas podem resultar na parada da fotossíntese, perturbação do metabolismo e, finalmente, a morte da planta (JALEEL et al., 2008). Alguns processos fisiológicos são ativados por variações no conteúdo de água dos tecidos, enquanto outros são acionados por hormônios das plantas que sinalizam variações hídricas (CHAVES et al., 2003).

O ajuste osmótico é um mecanismo que possibilita às plantas manterem a absorção de água e a pressão de turgor, contribuindo para sustentar alta taxa fotossintética e expansão do crescimento. Uma análise comparativa de muitos

estudos dedicados ao ajuste osmótico tem sugerido que o ajuste não pode ser considerado igualmente útil em todas as culturas em condições de seca, mas que uma associação positiva entre produtividade e ajuste osmótico pode ser encontrada sob estresse severo onde a produtividade tende a ser baixa (SERRAJ e SINCLAIR, 2002). Alterações na estrutura da membrana celular também promovem mudanças em canais de transporte ativados por pressão, modifica a conformação ou a justaposição de proteínas sensoriais embebidas nas membranas celulares e altera a continuidade entre a parede e a membrana celular (SHINOZAKI e YAMAGUCHISHINOZAKI, 1999; SHINOZAKI e YAMAGUCHI-SHINOZAKI, 2000). A mudança no potencial osmótico pode ser uma resposta ao estresse hídrico em nível molecular (BRAY, 1993). A percepção do déficit hídrico celular precisa ser traduzida em compostos bioquímicos e metabólitos, gerando uma consequente resposta fisiológica ao estresse (INGRAM e BARTELS, 1996).

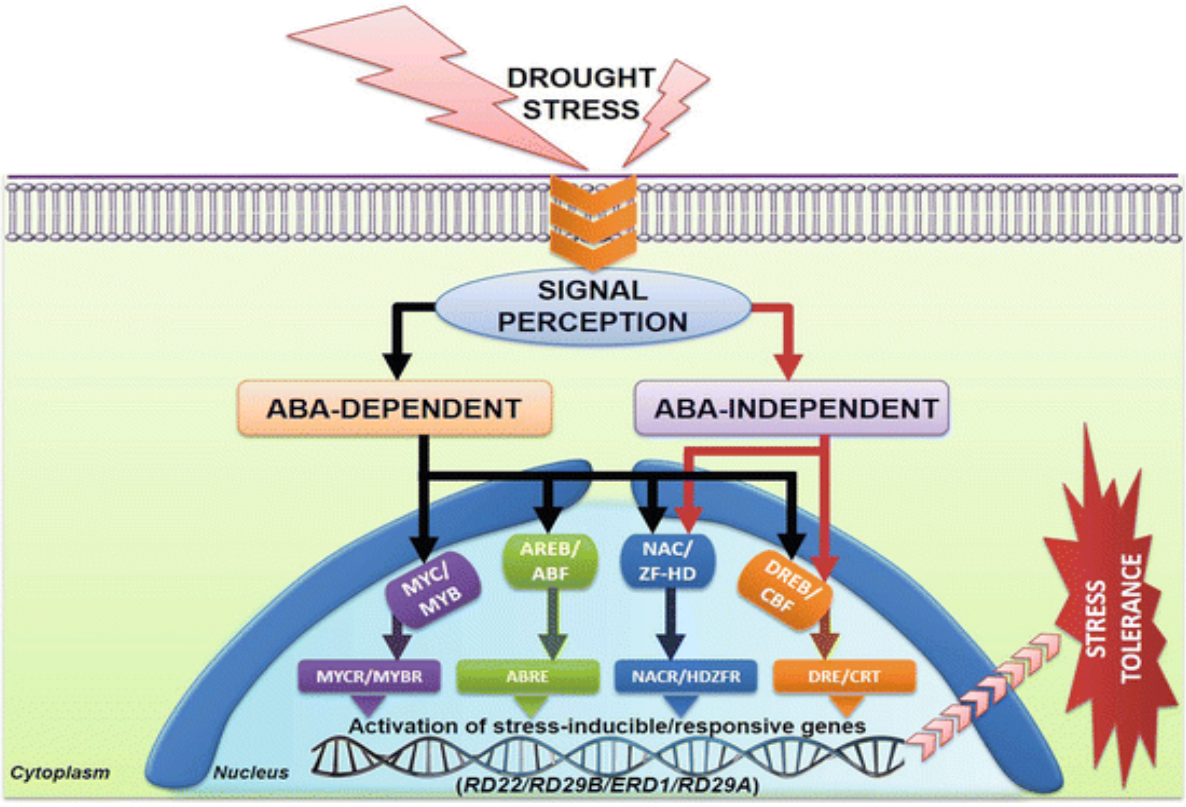
O déficit hídrico em plantas inicia-se a partir de uma complexa via de respostas, começando com a percepção do estresse, o qual desencadeia uma cascata de eventos moleculares, sendo finalizada em vários níveis de respostas fisiológicas, metabólicas e de desenvolvimento (BRAY, 1993), que auxiliam os vegetais a se adaptarem a condições adversas (ARORA, 2002; SEKI, 2003). O conhecimento dessas respostas é essencial para elucidar os mecanismos de resistência nas plantas (REDDY et al., 2004; JALEEL et al., 2006; SHAO et al.; 2008). Deste modo, uma melhor compreensão dos aspectos morfoanatômicos e base fisiológica de mudanças na resistência ao estresse hídrico poderia ser usada para selecionar ou criar novas variedades (NAM et al., 2001; MARTINEZ et al., 2007), tornando assim as culturas mais tolerantes ao estresse (REDDY et al., 2004; SHAO et al., 2008), até mesmo com a severidade e duração da escassez hídrica (FOYER, 2001; SHAO et al., 2006).

Os produtos dos genes induzidos por estresse podem ser classificados em dois grupos: aqueles que protegem diretamente contra estresses e aqueles que regulam a expressão de genes e a transdução de sinais em resposta ao estresse. O primeiro grupo inclui proteínas como chaperones e proteínas detoxificadoras, que provavelmente funcionam como protetores celulares contra desidratação. O segundo grupo é constituído de proteínas regulatórias, como fatores transcricionais que regulam a expressão de genes responsivos aos estresses, além de proteínas quinases e fosfatases que regulam transdução de sinais (HASSEGAWA et al., 2000;

SHINOZAKI e YAMAGUCHI-SHINOZAKI, 2000). Um padrão de expressão gênica pode ser estabelecido como resultado de condições específicas de estresse. Esse padrão pode sofrer variações nas etapas iniciais, alterando a taxa de transcrição de um gene específico ou, subsequentemente, controlando especificamente os níveis de mRNA e a tradução (BRAY et al., 2002).

A análise transcricional em folhas de soja submetidas ao estresse osmótico e ao estresse do retículo endoplasmático apresentou alterações na expressão gênica, predominantemente positivas, e identificou genes co-regulados por esses estresses (IRSIGLER et al., 2007). Análises genômicas e moleculares têm facilitado a descoberta de genes e capacitado a engenharia genética a usar vários genes regulatórios ou funcionais para investigar vias relacionadas a tolerância à seca em plantas (UMEZAWA et al., 2006), mesmo que a complexidade da resposta molecular à seca tenha sido só recentemente revelada por análises de transcriptoma (KOLLIPARA et al., 2002; BUCHANAN et al., 2005; HAZEN et al., 2005).

Figura 1: Efeitos do estresse hídrico sobre o desenvolvimento da planta e resposta ao mesmo.



Fonte: Lata et al, 2015.

1.3 Plataformas NGS

O desenvolvimento de tecnologias de sequenciamento de nova geração (NGS) expandiu as abordagens baseadas em sequenciamento no perfil de expressão gênica, bem como ampliou a abrangência da identificação de transcritos por meio de técnicas de sequenciamento (RNA-seq). O alto nível de sensibilidade e natureza de alta produtividade tornam as tecnologias NGS o método de escolha para a análise de expressão gênica (LI et al., 2008; MORTAZAVI et al., 2008; SULTAN et al., 2008)

Em geral, um ensaio de análise da expressão gênica por RNA-seq inclui a conversão de um conjunto de mRNAs numa biblioteca de fragmentos de cDNA aleatoriamente quebrados com adaptadores ligados a uma ou ambas as extremidades. Isto é seguido por sequenciamento com ou sem amplificação por PCR de uma extremidade (extremidade única) ou ambas as extremidades (extremidade emparelhada). O comprimento de leitura varia de 30 a 400 pb., dependendo da tecnologia de sequenciamento usada. Quatro tecnologias principais de NGS estão disponíveis para RNA-seq: RocheGS FLX 454 sequencer (Roche Diagnostics Corp., Branford, CT, USA), Illumina genome analyzer (Illumina Inc., San Diego, CA, USA), ABI SOLiD System (Life Technologies Corp., Carlsbad, CA, USA) e Ion Personal Genome Machine (Life Technologies, South San Francisco, CA, USA). Um segundo grupo inclui a HeliScope (Helicos BioScience Corp.,Cambridge, MA, USA) e PacBio RS single-molecule realtime (SMRT) system (Pacific Biosciences, Menlo Park, CA, USA) (MARDIZ, 2008; SIMON et al., 2009; VARSHNEY et al., 2009) que são tecnologias baseadas no sequenciamento de uma única molécula, portanto não requer o passo de amplificação prévio ao sequenciamento. Entre estas seis plataformas disponíveis, a Illumina/Solexa Genome Analyzer, a Roche 454 GS FLX sequencer, a Applied Biosystems SOLiD Analyzer e a HeliScope (que pertence às tecnologias de sequenciamento de segunda geração) dominam o mercado, enquanto que a Pacific Biosciences PacBio RS SMRT system e a Ion Personal Genome Machine da Life Technologies (terceira geração), têm sido introduzidas recentemente, portanto ainda não são de amplo uso (JAIN et al., 2005). No caso de Illumina, os custos são menores, no entanto, as leituras que se produzem são de menores

fragmentos em relação as leituras geradas por Roche/454 (OZSOLAK e MILOS, 2011; BARBA et al., 2014).

A Tabela 2 indica um resumo comparativo entre as principais plataformas de sequenciamento baseado em (BARBA et al., 2014), o qual facilita a escolha da plataforma mais apropriada. Em geral, a Roche 454 gera as leituras de maior comprimento, a Illumina tem a maior capacidade de sequenciamento e os menores custos, e a SOLiD 5500 xls gera a maior acurácia (LIU et al., 2012).

Tabela 2: Resumo comparativo entre as principais plataformas de sequenciamento.

Plataforma	Método de amplificação	Química do sequenciamento	Comprimento das leituras (pb)	Máxima produção per corrida	Acurácia (%)
454 (Roche)	PCR de emulsão	Pirosequenciamento	400-700	700 Mpb	99.9
Illumina (ILLUMINA)	Amplificação em ponte (*Bridge PCR*)	Terminadores reversíveis	100-300	600 Gpb	99.9
SOLiD (Life Technologies)	PCR de emulsão	Ligação	75-85	80-360 Gpb	99.99
PacBio (Pacific Biosciences)	Sequenciamento de molécula única em tempo real	Nucleotídeos fluorescentemente marcados	4000-5000	200 Mb-1 Gb	95
Helicos (Helicos Biosciences)	Sequenciamento de molécula única	Terminadores reversíveis	25-55	35 Gpb	97
Ion Torrent (Life Technlogies)	PCR de emulsão	Detecção da liberação do H^+	100-400	100 Mb-64Gpb	99
Nanopore (Oxford Technologies	Sequenciamento de molécula única	-	Leituras muito extensas até de 50 kpb	Dezenas de Gpb	96

Fonte: Adaptada de: Barba, 2014.

Cada plataforma de sequenciamento possui suas vantagens e desvantagens. Porém, para projetos de RNA *sequencing*, é preciso ter uma alta cobertura, uma média elevada de leituras que estejam sobrepondo num determinado nucleotídeo na sequência reconstruída. A Illumina é uma das opções mais recomendáveis para este tipo de projeto, a qual oferece o menor custo, uma boa precisão e o maior rendimento (RADFORD et al., 2012) e que durante os últimos 5 anos tem sido usada com maior frequência em diferentes projetos que envolvem sequenciamento de plantas (BARBA et al., 2014)

A Illumina tem desenvolvido a série de plataformas que incluem a HiSeq 2500, a HiSeq 2000, a HiSeq 1500 e a HiSeq 1000, as quais têm sido vantajosas em relação a outras plataformas, devido principalmente a quantidade de nucleotídeos que são capazes de sequenciar numa mesma corrida, bem como o tempo que leva o sequenciamento, a longitude das leituras geradas, a precisão no sequenciamento e os baixos custos. A Hiseq 2500 tem a capacidade de sequenciar um genoma em 24 horas, 20 exomes num dia ou 30 amostras para RNA *sequencing* em aproximadamente 5 horas (BARBA et al., 2014). O verdadeiro desafio dentro de um projeto de RNA-seq consiste nas análises de compreensão e interpretação da grande quantidade de dados gerados, cujo objetivo é reconstruir o transcriptoma a partir das milhões de leituras, e encontrar padrões que respondam uma pergunta biológica. O processo de montagem, contagem, normalização e análises estatísticas requeridas para o processamento de giga pares de bases (Gpb) de informação produzidas no sequenciamento é realizado por meio de abordagens computacionais de bioinformática. Atualmente estão disponíveis muitas ferramentas para estes tipos de análises, tanto de uso livre (*open source*) como de uso comercial. Consequentemente, uma compreensão completa é necessária para escolha da combinação mais mais apropriada das ferramentas disponíveis.

1.4 RNA-Seq

RNA-seq é uma ferramenta voltada para o estudo de perfis de transcriptoma, associada à tecnologia das plataformas NGS (Next Generation Sequence), a qual fornece uma qualidade de informação bem mais precisa de dados quando comparada a outros métodos (WANG et al., 2009), também tem a vantagem de ser mais sensível e dinâmica quando comparada a outras tecnologias antes usadas como os microarrays (MORTAZAVI et al., 2008), apresentam algumas desvantagens, tais como a hibridização cruzada, a hibridização não específica e uma sensibilidade limitada.

O primeiro passo para a resposta ao estresse é o reconhecimento do sinal de estresse e subsequentes respostas moleculares, bioquímicas e fisiológicas ativadas por meio da transdução de sinal (KOMATSU et al., 2009 ; GE et al., 2010 ; LE et al., 2012), e mediado pela atividade transicional de ativação e repressão de genes.

Portanto, abordagens de RNAseq são apropriadas para visualização destas cascatas de respostas.

Uma gama de mecanismos de defesa é ativada, aumentando a tolerância da planta contra condições adversas, a fim de evitar danos impostos por estresses abióticos. O conjunto de todos os transcritos derivados de genes produzidos numa célula em uma determinada condição fisiológica é conhecido como o transcriptoma. A avaliação das flutuações no transcriptoma é fundamental para compreender a função, estrutura e as interações dos genes envolvidos num determinado processo, o que permite também avaliar muitos fenômenos biológicos, incluindo Single Nucleotide Polimorphism (SNP), eventos epigenéticos, *splicing* alternativo e o estudo de interações proteína-DNA (WANG et al., 2009; GONÇALVES, 2013). O conhecimento global desses mecanismos moleculares pode ser aplicado na modulação e alteração dos padrões de expressão numa determinada condição, visando melhorar e otimizar os processos biológicos envolvidos. Uma visão geral de uma *pipeline* típica de RNA-seq para análise de expressão diferencial é delineada na Figura 2. Primeiro, as leituras são mapeadas para o genoma ou transcriptoma, em seguida são mapeadas para cada amostra que são montadas em nível de gene, éxon ou nível de expressão de transcrição, dependendo dos objetivos do experimento. Em seguida, os dados sumarizados são normalizados em conjunto com teste estatístico, levando a uma lista de genes candidatos. Finalmente, a visão biológica dessas listas da *pipeline* de análise RNA-Seq para detectar a expressão diferencial. (TRAPNELL et al., 2012).

Muitas ferramentas que se concentram em testes de conjunto de genes, inferências de redes e bancos de dados de conhecimento têm sido projetadas para analisar listas de genes diferencialmente expressos de conjuntos de dados. As ferramentas de ontologia a qual é uma representação formal de um corpo de conhecimento dentro de um determinado domínio e que geralmente consistem em um conjunto de classes (ou termos ou conceitos). A Ontologia Gênica (GO) descreve nosso conhecimento do domínio biológico em relação a três aspectos:

Função Molecular, em termos descrevem atividades que ocorrem no nível molecular, como "catálise" ou "transporte". Os termos da função molecular GO representam atividades e não as entidades (moléculas ou complexos) que executam as ações e não especificam onde, quando ou em que contexto a ação ocorre. Funções moleculares geralmente correspondem a atividades que podem ser executadas por

produtos génicos individuais (isto é, uma proteína ou RNA). Exemplos de termos funcionais amplos são atividade catalítica e atividade transportadora. Para evitar confusão entre os nomes dos produtos gênicos e suas funções moleculares, as funções moleculares GO são frequentemente anexadas à palavra "atividade" (uma proteína quinase teria a atividade da proteína quinase da função molecular GO).

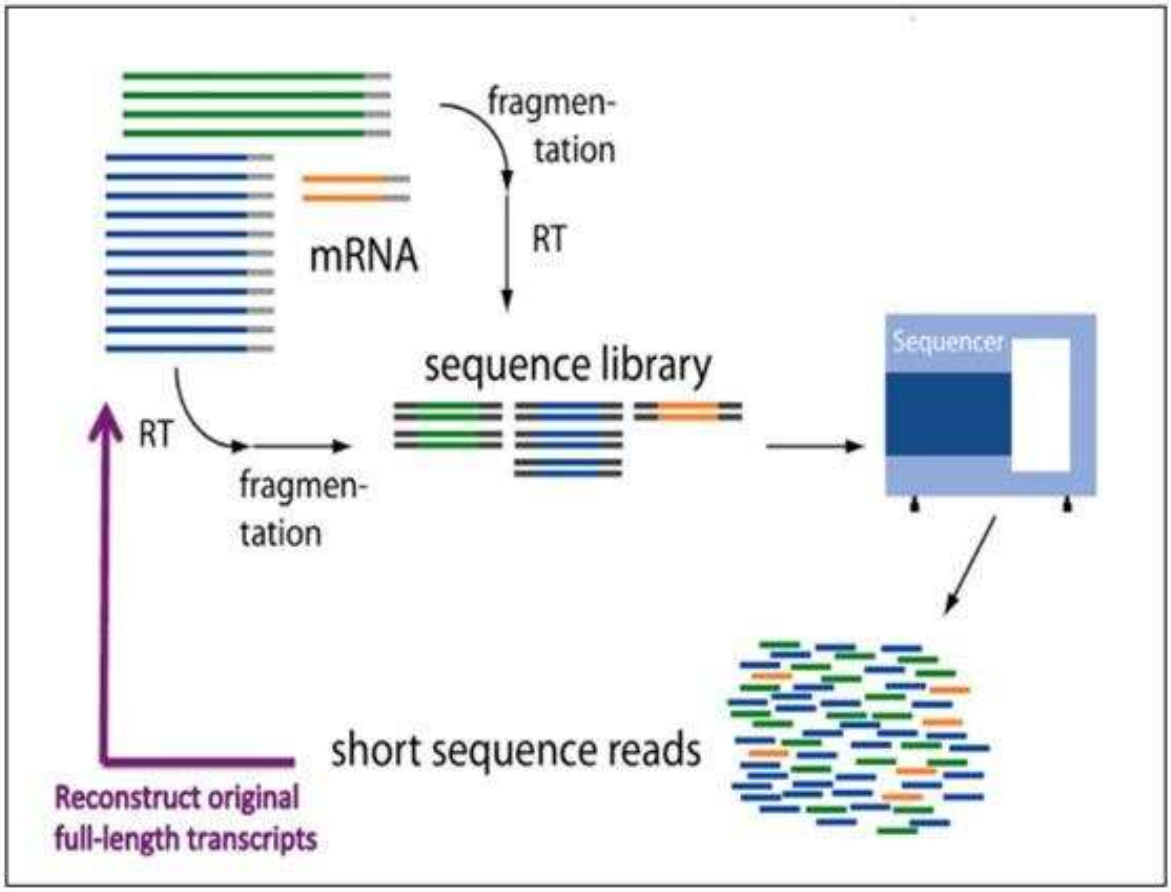
Componente Celular, são locais relativos às estruturas celulares nas quais um produto genético desempenha uma função, compartimentos celulares (por exemplo, mitocôndria) ou complexos macromoleculares estáveis dos quais fazem parte (por exemplo, o ribossomo). Diferentemente dos outros aspectos do GO, as classes de componentes celulares não se referem a processos, mas a uma anatomia celular.

Processo Biológico, ou 'programas biológicos', realizados por múltiplas atividades moleculares. Exemplos de termos gerais de processo biológico são reparo de DNA ou transdução de sinal. Exemplos de termos mais específicos são o processo biossintético da pirimidina nucleobase ou o transporte transmembranar da glicose. Observe que um processo biológico não é equivalente a um caminho. No momento, o GO não tenta representar a dinâmica ou as dependências necessárias para descrever completamente um caminho. O enriquecimento via ontologia é uma importante iniciativa de bioinformática para unificar a representação de atributos de genes e produtos genéticos em todas as espécies, mais especificamente, o projeto visa: 1) manter e desenvolver seu vocabulário controlado de atributos de genes e produtos genéticos; 2) anotar genes e produtos gênicos, assimilar e disseminar dados de anotação; e 3) fornecer ferramentas para facilitar o acesso a todos os aspectos dos dados fornecidos pelo projeto e permitir a interpretação funcional dos dados experimentais usando o GO, por exemplo, por meio de análise de enriquecimento (YOUNG *et al.* 2010). O sequenciamento de RNA tem conseguido superar as limitações de outras tecnologias de amplo uso, como os Microarrays, devido principalmente a necessidade de quantidades menores de RNA, a possibilidade de encontrar estrutura de éxons, introns e locais de *splicing* alternativo, assim como a identificação das extremidades 5’ e 3’ dos genes. Além disso, essa tecnologia possibilita a quantificação dos níveis de expressão de éxons e as variantes de *splicing* (MARGUERAT et al., 2008; SHENDURE, 2008; WANG et al., 2009).

O perfil do transcriptoma oferece uma oportunidade para investigar a regulação da resposta das plantas e identificar os genes envolvidos nos mecanismos de

tolerância ao estresse específico de um determinado genótipo. Anteriormente, abordagens que usam *Expressed Sequence Tass* (ESTs) juntamente com várias técnicas, como a supressão de hibridização subtrativa (SSH), foram amplamente utilizadas para o perfil transcriptoma de soja sob condições de estresse abiótico (CLEMENT et al., 2008). Além disso, informações de ESTs foram usadas para desenvolver microarranjos marcados (O'ROURKE et al., 2007). Essas técnicas são eficientes, mas não garantem a análise de genes inteiros no genoma da soja. Diversas técnicas de alto rendimento foram desenvolvidas para a análise do transcriptoma devido ao avanço da tecnologia de sequenciamento e a disponibilidade de toda a sequência do genoma da soja, (LIBAULT et al., 2010 ; SCHMUTZ et al., 2010 ; CHENG et al., 2013).

Figura 2: Metodologia geral usada na técnica de sequenciamento RNA-seq. Uma biblioteca de cDNA deve ser preparada após do isolamento e fragmentação do mRNA. Esta biblioteca será sequenciada usando uma plataforma de sequenciamento que gera milhões de leituras curtas.



Fonte: (MSKCC, 2014)

1.5 Transcritoma e Splicing Alternativo

Transcritoma é o conjunto completo de transcritos da célula em um estágio específico de desenvolvimento ou condição fisiológica. Portanto, a identificação dos transcritos expressos é essencial para o entendimento do genoma e do organismo como um todo (NOBUTA, VENU, et al., 2007). Além disso, compreender o transcriptoma é essencial para interpretação dos elementos funcionais do genoma, bem como o entendimento dos constituintes moleculares de células e tecidos ou da compreensão do desenvolvimento de doenças, por exemplo (WANG, GERSTEIN e SNYDER, 2009).

A regulação da transcrição envolve não apenas as proporções diferentes da transcrição nas diferentes partes do genoma, como também a escolha das regiões que devem ser transcritas, e a extensão desta transcrição. Desse modo, diferentes conjuntos de genes podem ser transcritos em diferentes células, ou na mesma célula, em momentos diferentes (WATSON, BAKER, et al., 2006). A transcrição passa por uma série de etapas bem definidas, sendo as principais: iniciação, alongamento e terminação. O nucleotídeo no DNA que codifica o início da cadeia de RNA é chamado de sítio de início da transcrição, (do inglês Transcriptional Start Sites – TSS), designado pela posição +1. As sequências situadas no sentido da transcrição são referidas como a jusante ao ponto de início (downstream). Da mesma forma, as sequências situadas na região anterior ao TSS são referidas como sequências à montante (upstream) (WATSON, BAKER, et al., 2006).

Nos eucariotos, os genes transcritos geram o mRNA que se organiza entre sequencias codificantes denominadas (éxons), separadas entre si por sequências não-codificantes, denominadas (íntrons). Desta forma, há um padrão de alternância entre éxons e íntrons, estes por sua vez são removidos do pré-mRNA por meio de um mecanismo denominado processamento de RNA (*splicing*). Este processo converte o pré-mRNA em RNA mensageiro maduro. O *splicing* é um mecanismo que envolve duas etapas. A primeira etapa envolve a clivagem no sítio intrônico 5’ (RUSKIN et al., 1984); em seguida ocorre a ligação do fosfato na extremidade 5’ a uma adenosina (KONARSKA et al., 1985). Frequentemente, pré-mRNAs podem ser processados de mais de um modo, originando mRNAs alternativos pela remoção de diferentes combinações de íntrons. Este processo é denominado *splicing* alternativo, dessa

maneira, um gene pode dar origem a mais de um produto polipeptídico (WATSON, BAKER, et al., 2006).

Os diferentes mRNAs formam variantes proteicas denominadas isoformas (Hanke et al., 1999). Os mecanismos moleculares da reação de *splicing* precisam distinguir o que é éxon e íntron para remover os íntrons e ligar os éxons com alta precisão. Com esta finalidade, as fronteiras entre íntrons e éxons são marcadas por sequências de nucleotídeos específicas nos pré-mRNAs. Essas sequências determinam onde ocorrerá o *splicing* e são denominadas sítio de processamento 5' e sítio de processamento 3'. Outro fator importante é a presença de uma Adenina (Resíduo A) no sítio de ramificação (WATSON, BAKER, et al., 2006).

Grande parte dos genomas podem possuir múltiplas isoformas transcritas efetivas (WANG et al., 2008; PAN et al., 2008), as quais podem ser geradas por *splicing* alternativo do transcrito primário de mRNA, transcrição de promotores alternativos e clivagem em locais alternativos de poliadenilação 3 (LICATALOSI, 2010). As variantes de transcrição podem levar a isoformas de proteínas com função distinta, UTRs modificadas com potencial regulatório alterado ou transcritos não funcionais, que estão sujeitos a decaimento mediado por nonsense (NMD). Além disso, a variação genética em uma população pode afetar a expressão de isoformas transcritas individuais, bem como o uso de diferentes isoformas pode contribuir para a diversidade fenotípica (PICKRELL et al., 2010; LAPPALAINEN et al., 2013).

RNA-seq é o método mais indicado para o estudo das isoformas transcritas existentes e das que vierem a existir. Os primeiros estudos de *splicing* baseados em dados de RNAseq debruçaram-se na análise de eventos de *splicing* (WANG et al., 2008), onde, tipicamente, consideram um conjunto definido de eventos de ocorrência comum. Entre eles, *Skipped* ou inclusão de um único éxon em pares de éxons que são incluídos nos transcritos de uma maneira mutuamente excludente, de extensão ou encurtamento de éxon devido ao *splicing* nos sítios 5′ ou 3′. Contudo, geralmente, os eventos de *splicing* de um gene podem ser descritos por seu gráfico de *splicing*, conforme Heber et al., 2002. Para tornar a tarefa matematicamente tratável, os métodos se baseiam em suposições como, por exemplo, abotoaduras que se baseiam em parcimônia, prevendo um conjunto mínimo de transcrições consistentes com os dados observados. Para isto, o sequenciamento de pequenos fragmentos precisa ser remontado para retornar às moléculas iniciais de RNA, causando esforço na tarefa de

remontagem (MARTIN e WANG, 2011). Seja na presença ou na ausência de um genoma de referência, mas com o objetivo geral de identificar e quantificar todas as moléculas de RNA inicialmente presentes na amostra. O principal desafio consiste no fato de que as leituras são curtas e podem ser atribuídas ambiguamente à várias transcrições.

Basicamente são dois elementos, cis e trans-acting, responsáveis pela regulação do *Splicing* alternativo. Os elementos cis-acting incluem sequências consenso e elementos auxiliares, enquanto o trans-acting compreendem um grupo de *Ser/Arg-Rich proteins* (SRs), *heteroneous nuclear ribonucleoproteins* (hnRNPs) e *small nuclear ribonucleoproteins* (snRNPs), que estão diretamente envolvidos no processo de *Splicing*. Todo esse processo é mediado via spliceossomo, que abrange os elementos cis e trans (LE et al, 2015). Basicamente, existem 5 principais tipos de *Splicing* alternativo, sendo-inclusão alternativa do exon (skipping exon), sítio de *Splicing* alternativo 5’ e 3’ (A3’SS e A5’SS), exon mutuamente exclusivos (MXE) e retenção de íntron (RI).

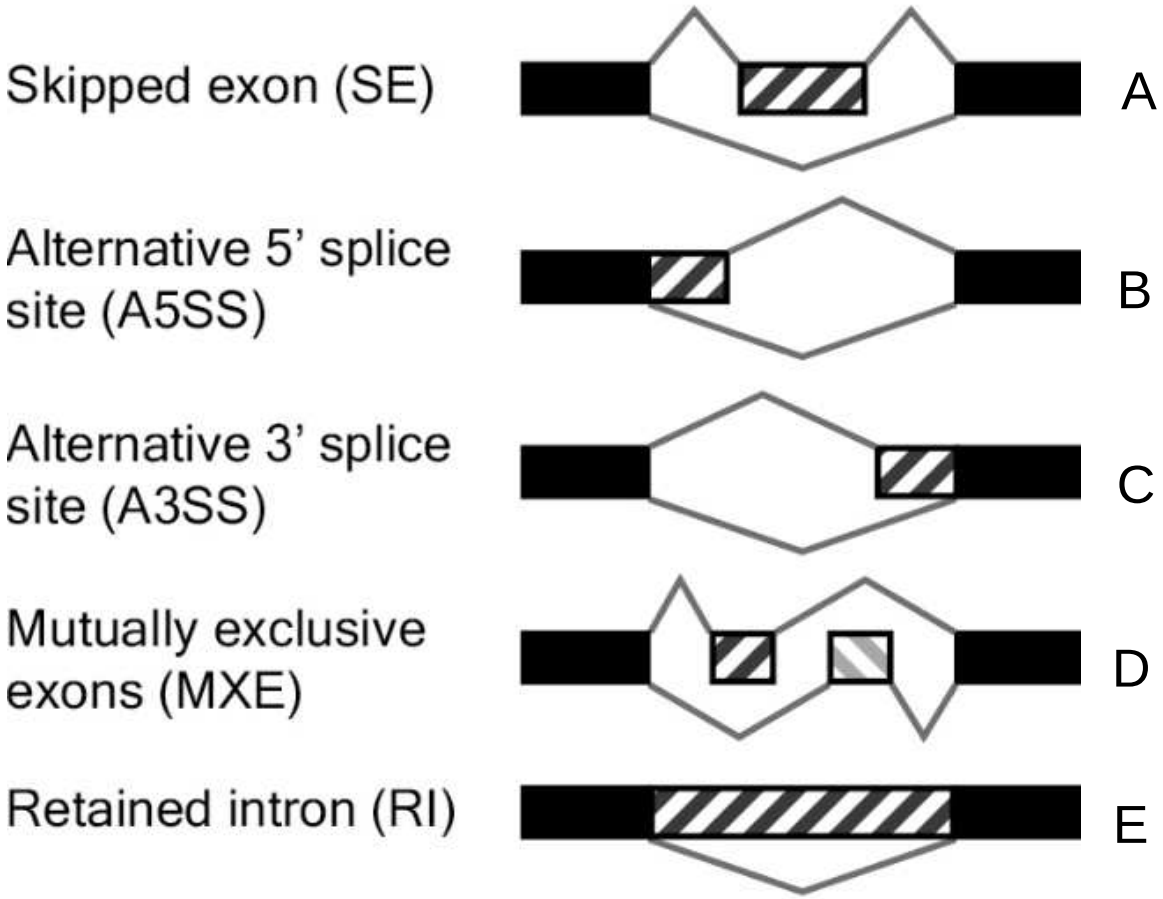
A inclusão alternativa do exon, também conhecida como Skipping exon, é o evento mais comum a ser encontrado entre os possíveis eventos de *Splicing* alternativo, no qual se concretiza através da inclusão ou exclusão de introns, regulando grande parte dos processos fisiológicos (MARLON et al., 2005). O sítio de *Splicing* alternativo das extremidades 5’ e 3‘ são eventos que ocorrem quando há o reconhecimento de dois ou mais sítios na extremidade de um exon. A ocorrência de *Splicing* nas extremidades correspondem à uma menor parcela entre os eventos acima mencionados (KOREN et al, 2007).

O exon mutuante exclusivo (MXE) é o evento de menor ocorrência entre os listados, caracterizado pela permanência ou flanqueamento de éxon na constituição das estruturas, alterando o tamanho final da proteína (POHL et al., 2013). Retenção de intron ocorre quando existe a permanência de um intron dentro do RNA maduro. É uma das formas de maior ocorrência em plantas. Dessa forma, há uma alteração no tamanho final do transcrito, estando ligado também a eventos de degradação, regulação da expressão gênica e ocorrência de novas isoformas. (WONG et al., 2016).

Em particular, no caso de *splicing* alternativo, as leituras provenientes de éxons constitutivos podem ser atribuídas a qualquer transcrição alternativa contendo este éxon. Contudo, encontrar a transcrição correta muitas vezes não é possível, pois,

como apontado na revisão de Martin e Wang (2011), os montadores de RNA-seq baseado em referência possuem suas próprias limitações. A integração e a interpretação dos dados fornecem informações úteis para uma compreensão abrangente, neste caso, do genoma de soja.

Figura 3: Uma forma comum de *splicing* alternativo é a presença ou não de éxon (mostrado em a), a de sítios 5' (mostrado em b) ou 3' (mostrado em c) que levam à inclusão ou não de um éxon. Alternativamente, éxons mutuamente excludentes podem permutar de lugar na forma madura no mRNA (mostrado em d). Íntrons internos podem ser incluídos ou não na sequência final do mRNA (mostrado em e)



Fonte: Shen et al., 2014.

2. OBJETIVOS

2.1 Objetivos Gerais

O objetivo deste projeto foi analisar perfis de transcrição de dois genótipos de soja contrastantes para tolerância à seca, BR16 e EMBRAPA48, para a identificação de genes relacionados a tolerância a seca, bem como analisar padrões de *splicing* alternativos relacionados aos mecanismos de regulação da expressão gênica em resposta ao estresse de seca.

2.2 Objetivos Específicos

i. Identificar transcritos diferencialmente expressos que possam estar envolvidos nas vias metabólicas responsivas ao estresse analisado;

ii. Identificar novas isoformas de *splicing* que estão diretamente ligadas ao processo de tolerância ao estresse via software especializado.

iii. Identificar genes candidados envolvido no mecanismo de tolerância da escassez hídrica

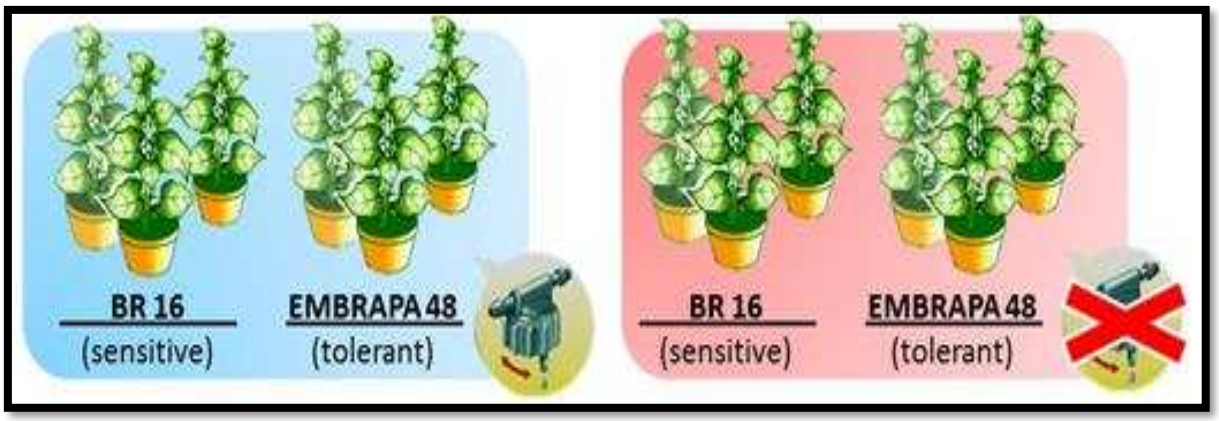
3. METODOLOGIA

3.1 Material vegetal, crescimento e estresse hídrico

Duas cultivares de soja que apresentam respostas contrastantes ao déficit hídrico, BR16 e EMBRAPA48, sensíveis e tolerantes à seca, respectivamente, foram utilizadas de experimentos anteriores (MESQUITA 2013; BALBI 2013), como descrito a seguir. As sementes das cultivares de soja BR16 e EMBRAPA48 foram germinadas em substrato orgânico e 3 plantas foram transferidas para vasos de 10L contendo uma mistura de solo, areia e terra (3: 1: 1) e cultivadas em casa de vegetação sob luz natural, umidade relativa (65 – 85%) e temperatura (15 – 35℃). As plantas foram mantidas irrigadas até atingirem o estágio de desenvolvimento V4, quando o suprimento de água foi interrompido. O delineamento experimental foi inteiramente casualizado, sendo o primeiro fator correspondente ao potencial hídrico das plantas (-1,0 MPa e controle irrigado) e o segundo fator foram duas cultivares (BR16 e EMBRAPA48). Os regimes hídricos foram atribuídos como irrigados (IR) e não irrigados (NI).

O potencial hídrico foliar (ψw) foi medido no terceiro trifólio emergente ao amanhecer, utilizando uma bomba de Scholander (SCHOLANDER et al., 1965) durante o período de estresse. As amostras foram coletadas em nitrogênio líquido e armazenadas a - 80℃.

Figura 4: Plantas de soja BR16 e EMBRAPA48 expostas a um regime de seca gradual para isolar o RNA para análise de transcriptômica. O potencial hídrico foi medido por Scholander.



3.2 Extração de RNA, construção da biblioteca e sequenciamento

A extração de RNA proveniente de folhas da soja foi realizada com reagente Trizol (Invitrogen). Foram usados 5 ug de RNA paired-end (extremidade emparelhada), gerando uma biblioteca com *reads* de 100pb com Kit RNA-Seq BIOO NEXT (Bioo-Scientific, Austin, TX). As qualidades das bibliotecas foram analisadas no BioAnalyzer (Agilent, Santa Clara, CA) e quantificadas por meio do equipamento Qubit (LifeTechnologies, Carls-bad, CA). As bibliotecas foram então agrupadas em proporções equimolares, quantificadas por qPCR com o kit Kapa Library Quant (Kapa, Cape Town, África do Sul). Três amostras biológicas de RNA em replicata para cada cultivar foram sequenciadas na Illumina Hi-Seq 2500 (Illumina, San Diego, Califórnia) do NuBioMol (Núcleo de Análise de Biomoléculas - UFV, Brasil). O processamento das *reads* foi realizado usando o *pipeline* de análise Illumina (no formato Fastq).

3.3 Qualidade de dados e trimagem

As leituras dos dados brutos foram subsequentemente submetidas a trimagem e filtragem usando o software Trimmometic (BOLGER et al. 2014), já a qualidade de dados foi averiguada pelo FastQC v0.10.0 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Parâmetros como a qualidade da sequência por nucleotídeo (Q20 e Q30), por conteúdo de GC, por conteúdo de nucleotídeos não determinados (N) e outros são também avaliados com este software. FastQ usa o phred quality score para determinar os Q20 e Q30 score. Este é definido pela seguinte expressão:

$$Q= - 10 \log10P$$

Em que P é a probabilidade de que uma base na leitura tinha sido atribuída de modo errado durante o sequenciamento.

Segue a linha de comando usada para gerar os arquivos de trimagem dos arquivos fastq.

```
trimmomatic PE -phred33 /path/R1_paired.fq.gz path/R2_paired.fq.gz
ILLUMINACLIP:contams_forward_rev.fa:2:30:10 LEADING:3 TRAILING:3
SLIDINGWINDOW:4:15 MINLEN:36
```

3.4 Indexação, alinhamento e mapeamento do transcriptoma com Bowtie2/Tophat

O alinhamento do transcriptoma primário de *Glycine max* cultivar Williams 82 (Wm82.a2.v1) (SCHMUTZ et al. 2010) foi realizado pelo Bowtie2 (LANGMEAD B., et al. 2012). O mapeamento das leituras foi realizado com o software TopHat (KIM et al. 2013) v1.3.3, disponível no site http://ccb.jhu.edu/software/tophat/index.shtml. TopHat é um mapeador *splicing junction* que usa o Bowtie aligner para alinhar as leituras ao genoma de referência e posteriormente identifica *splicing juction* entre os éxons.

Segue a linha de comando usada para gerar os arquivos de indexação do transcritoma pelo software Bowtie2.

```
bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r> | --interleaved <i> | --sra-acc <acc> | b <bam>} -S [<sam>]
```

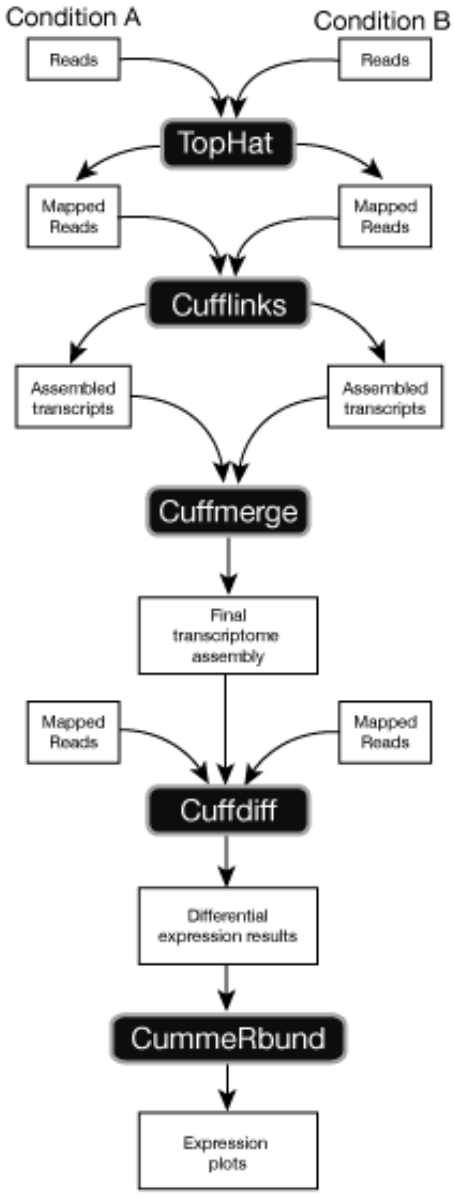
Segue a linha de comando usada para gerar os arquivos de mapeamento do transcritoma pelo software TopHat.

```
tophat [options]* <genome_index_base> <reads1_1[,...,readsN_1]> [reads1_2,...readsN_2]
```

3.5 Quantificação e expressão diferencial

Na montagem dos transcritos, determinação de abundâncias e análises de expressão diferencial, foi usado o software Cufflinks v2.2.1 (Trapnell et al 2012) disponível no site: http://cufflinks.cbcb.umd.edu/. As análises de expressão diferencial de genes (Differentially expression analysis - DE) foram realizadas pelo programa Cuffdiff, incluso no pacote de Cufflinks. Foram considerados como diferencialmente expressos apenas os genes cujos valores p foram ajustados a uma FDR <0,05 e que apresentassem log2 fold change valores maiores que 2, up regulados (>2) e menor que 0,5, down regulados (<0,5).

Figura 5. Sequência de análises pelo *tophat* e *cufflinks*



L

Fonte: TRAPNEL,2012.

Segue a linha de comando usada para gerar os arquivos de quantificação dos fragmentos mapeados pelo software cufflinks

```
cufflinks [options] <aligned_reads.(sam/bam)>
```

Segue a linha de comando usada para gerar os arquivos de Expressão diferencial pelo software cuffdiff.

```
cuffdiff [options]* <transcripts.gtf> \
<sample1_replicate1.sam[,…,sample1_replicateM.sam]> \
<sample2_replicate1.sam[,…,sample2_replicateM.sam]> … \
[sampleN.sam_replicate1.sam[,…,sample2_replicateM.sam]]
```

3.6 Indexação, alinhamento e mapeamento do transcriptoma com Generate/Star

O alinhamento do transcriptoma primário de *Glycine max* cultivar Williams 82 (Wm82.a2.v1) (SCHMUTZ et al. 2010) foi realizado pelo Generate. STAR é implementado como um código autônomo em C++, livre de código aberto distribuído sob licença GPLv3 e pode ser baixado em http://code.google.com/p/rna-star/ (A. Dobin et al, 2012) . STAR é um alinhador ultra-rápido para dados de RNA-Seq.

Segue a linha de comando usada para gerar os arquivos de Indexação do genoma pelo software Generate.

```
STAR --runThreadN --runMode genomeGenerate --genomeDir /path --
genomeFastaFiles /path --sjdbGTFfile /path --sjdbOverhang 1
```

Segue a linha de comando usada para gerar os arquivos de mapeamento do genoma pelo software STAR.

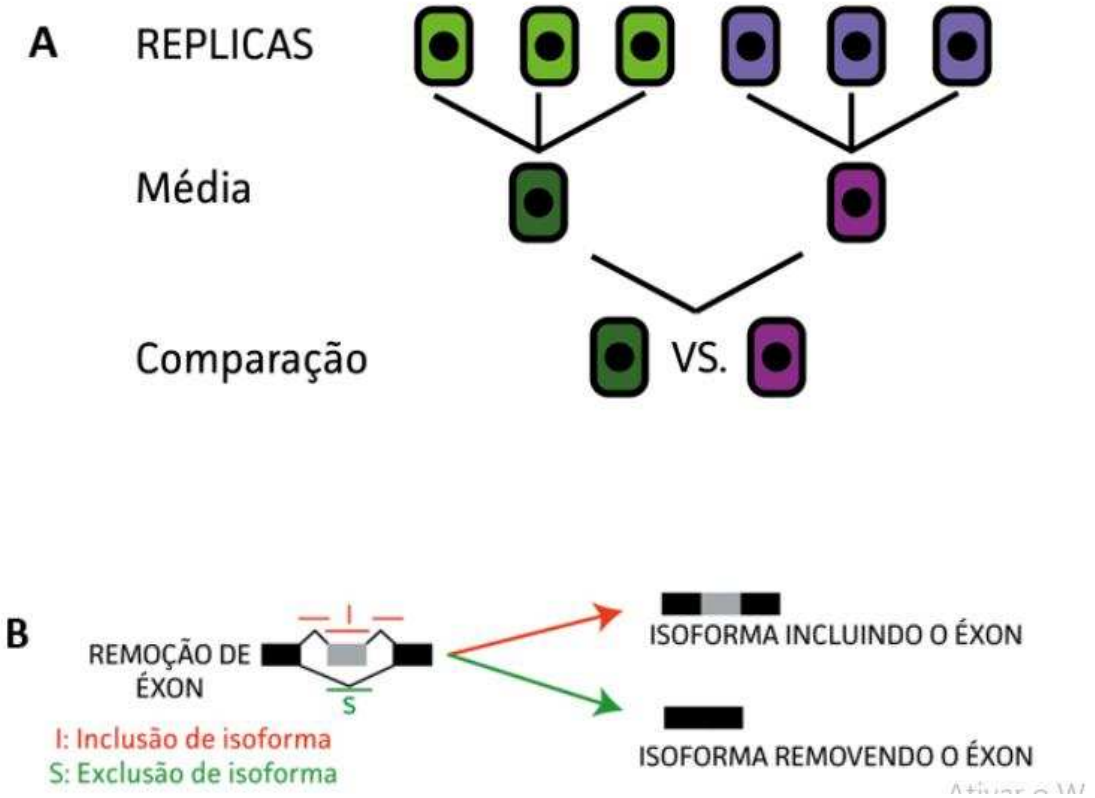
```
STAR --runMode alignReads --outSAMtype BAM Unsorted --
readFilesCommand zcat --genomeDir /path/generate --outFileNamePrefix
star --readFilesIn /path/right path/left
```

3.7 Identificação de *splicing* alternativo

rMATS v.3.2.5 encontrado em http://rnaseq-mats.sourceforge.net/ user_guide.htm (Shen et al, 2012; Park et al, 2013; Shen et al, 2014), foi usado para detectar eventos de *splicing* alternativos, o qual identifica todas as classes de *splicing* possíveis, nominalmente identificadas, Retained intron (RI), Mutually exclusive éxon (MXE), alternative 5′ *Splicing* Site (A5SS), alternative 3′ *Splicing* Site (A3SS), Skipped éxons (SE). O modelo estatístico do rMATS calcula a razão de verossimilhança. Os eventos com p-valor <0,05 foram identificados como eventos significativamente diferentes. Tendo como exemplo o evento de exon skipping, o nível de inclusão do éxon é estimado pela contagem de reads específicas para a inclusão de uma isoforma e a contagem de reads específica para a exclusão de uma isoforma. A inclusão de

uma isoforma é representada por reads que alinham para o éxon alternativo e aquelas que alinham para suas junções com éxons constitutivos. Já a exclusão de uma isoforma é caracterizada por reads que alinham para as junções entre éxons constitutivos. Então, a partir das análises individuais para cada replicata, é feita uma média do nível de inclusão para cada grupo e, em seguida, a variabilidade entre as amostras é calculada pela diferença de seus níveis de inclusão (Figura 6)

Figura 6: Estrutura hierárquica das análises de rMATS e diagrama esquemático para um evento de exon skipping. A. Inicialmente é feita uma estimativa do nível de inclusão para cada réplica. Em seguida é feita uma média para cada grupo seguida do cálculo da variabilidade entre os grupos pela diferença de seus níveis de inclusão. B. A inclusão de uma isoforma (I) é representada pelas reads que alinham no éxon alternativo e em suas junções com éxons constitutivos. A exclusão de uma isoforma (S) é representada por reads que alinham nas junções entre éxons constitutivos



Fonte: Adaptado de Shihao Shen,2014.

Segue a linha de comando usada para gerar os arquivos de Expressão diferencial do *Splicing* alternativo pelo software rMATS

```
python /rMATS-turbo-Linux-UCS2/rmats.py\
 --b1 /condition1/b1.txt\
 --b2 /condition2/b2.txt\
```

```
--gtf /path/.gtf --od /output \
-t paired --readLength 50 --cstat 0.0001 --libType fr-unstranded
```

3.8 Análise de dados e anotação funcional

As relações entre os conjuntos de genes expressos diferencialmente nas duas condições, separados também entre genes Up e Down regulados, foram analisadas por meio de diagramas de Venn, usando a ferramenta interativa online Draw Venn Diagram segundo endereço online http://bioinformatics.psb.ugent.be/webtools/Venn/. Por meio da mesma ferramenta foi analisada a expressão diferencial de cada genótipo em função de cada classe de *splicing*, RI, MXE, A3SS, A5SS e SE. Foi realizada a contagem dos eventos de *splicing* separados por classes de *splicing* para os genótipos BR16 e EMBRAPA48

Os genes comuns encontrados em cada evento do diagrama de Venn foram confrontados com a análise funcional da ferramenta phytoMine do grupo de analises do Phytozome https://phytozome.jgi.doe.gov/phytomine/begin.do (1997-2017 The Regents of the University of California). Foram categorizadas em Função Molecular (MF), Componente Celular (CC) e Processo Biológico (BP) em paralelo para cada genótipo. O teste exato de Fisher, que é baseado na distribuição hipergeométrica, será usado para calcular o p-valor.

Também foi usado para fins de enriquecimento o ClueGO a qual possui duas características principais: pode ser usado para a visualização de termos correspondentes a uma lista de genes ou para a comparação de anotações funcionais de dois grupos. Os conjuntos de identificadores de genes homólogos em arabidopsis foram carregados diretamente em formato de texto simples. Para permitir uma análise rápida, o ClueGO usa arquivos de anotação pré-compilados, incluindo GO, KEGG e BioCarta para uma ampla variedade de organismos. Um recurso de atualização com um clique baixa automaticamente as fontes mais recentes de ontologia e anotação e cria novos arquivos pré-compilados que são adicionados aos existentes. Isso garante uma análise funcional atualizada.

Para criar a rede de anotações com os dados inseridos, o ClueGO fornece configurações de análise funcional predefinidas que variam de gerais a muito específicas. Além disso, foi ajustado os parâmetros de análise para se concentrar em termos, por exemplo, em determinados intervalos de nível GO, com códigos de

evidência específicos ou com um certo número e porcentagem de genes associados. A relação entre os termos selecionados é definida com base em seus genes compartilhados de maneira semelhante à descrita por (HUANG et al. 2007) O ClueGO cria primeiro uma matriz binária de termos gênicos com os termos selecionados e seus genes associados. Com base nessa matriz, uma matriz de similaridade termo-termo é calculada usando estatísticas de Kappa corrigidas ao acaso para determinar a força da associação entre os termos. Como a matriz termo-termo é de origem categórica, a estatística kappa foi considerada o método mais adequado. Finalmente, a rede criada representa os termos como nós que são vinculados com base em um nível de pontuação kappa predefinido. O tamanho dos nós reflete a importância do enriquecimento dos termos. A rede é organizada automaticamente e usa o algoritmo de layout orgânico suportado pelo Cytoscape. Os grupos funcionais são criados por mesclagem iterativa de grupos definidos inicialmente com base no limite de pontuação kappa predefinido. Os grupos finais são fixos ou coloridos aleatoriamente e sobrepostos à rede. Os grupos funcionais representados pelo termo mais importante (principal) são visualizados na rede, que fornece uma visão perspicaz de suas inter-relações através do Cytoscape

4. RESULTADOS

4.1 Análise do transcritoma

A resposta da soja ao estresse de seca foi investigada ao nível transcricional por uma abordagem de RNA-Seq. As plantas foram submetidas ao estresse hídrico de forma moderada e caracterizadas em suas respostas morfológicas e fisiológicas nos estudos realizados por MESQUITA (2013), BALBI (2013) e LIMA (2016). No delineamento experimental, três repetições biológicas de duas cultivares de soja, BR16 e EMBRAPA48, sensíveis e tolerantes à seca, respectivamente, tiveram o RNA isolado do tecido foliar e sequenciado na Illumina Hi-Seq 2500 (Illumina, São Diego, CA). A amostra inicial coletada após o déficit hídrico foi designada como “NI” e as plantas de controle “IR” (Figura S1). Aproximadamente 45 a 50 milhões de leituras foram geradas de cada amostra, sendo 33,7 milhões de leituras com fragmentos de 35 a 110 pb. A soma das leituras entre as réplicas para cada condição oscilou entre os 50 e 53 milhões, uma porcentagem entre o 70 e 72% conseguiu ser mapeada no genoma de *G. max*. Os resultados do mapeamento conservaram homogeneidade, o qual é um bom indicativo de confiabilidade nos dados obtidos. As leituras brutas foram submetidas a um passo de pré-processamento/trimagem para remover sequências curtas ou de baixa qualidade além de adaptador/iniciador. O fluxo de trabalho de análise de RNA-Seq é mostrado na Figura S2 e foi utilizado para a análise de dados.

Em relação com a qualidade das leituras produzidas no sequenciamento, três parâmetros são mostrados: o Q20, o Q30 e o conteúdo de GC%. O Q20 nos indica a porcentagem das leituras com 99% ou mais de acurácia (do inglês *accuracy*), ou seja, há probabilidade de que 1 base seja atribuída de forma errada a cada 100 vezes nos picos do cromatograma durante o sequenciamento (do inglês *base calling*). O resultado de Q30 representa 99.9% ou mais de acurácia, ou seja, há a probabilidade de 1 base ser atribuída de forma incorreta a cada 1000 vezes. Quando um sequenciamento alcança um *phred quality* Q30 (ILLUMINA, 2011). Baseados nestas definições, observou-se que aproximadamente 97% das leituras geradas neste experimento, em cada condição, têm uma confiabilidade de 99% ou mais, bem como, aproximadamente 92% das leituras têm uma acurácia de 99.9%. Estes valores nos indicam confiabilidade nas sequencias geradas para as análises seguintes.

Por outro lado, tem sido reportado que para dados gerados em plataformas de sequenciamento que usam a plataforma *Illumina*, o viés (do inglês *bias*) no conteúdo de GC, sequencias ricas ou sequencias pobres no conteúdo destas bases, induzem alinhamentos irregulares ou não alinhamentos das leituras no genoma (CHEN *et al.*, 2013). A porcentagem média no conteúdo de CG obtida em nosso projeto oscila em torno de 42%. Para dados procedentes de uma biblioteca com distribuição normal, como é nosso caso, o esperado é um conteúdo de GC nas leituras, também ajustadas a uma distribuição normal, com medias entre 40 e 60% (BABRAHAM INSTITUTE, 2014), o que nos indica que nossos dados estão dentro destes limites.

Figura 7. Dispersão dos valores de qualidade por base ao longo da sequência da *read*. No eixo das ordenadas estão localizados os valores de qualidade Phred e no eixo das abscissas as posições das bases ao longo da sequência. As linhas vermelha e azul correspondem, respectivamente, aos valores da mediana e da média.

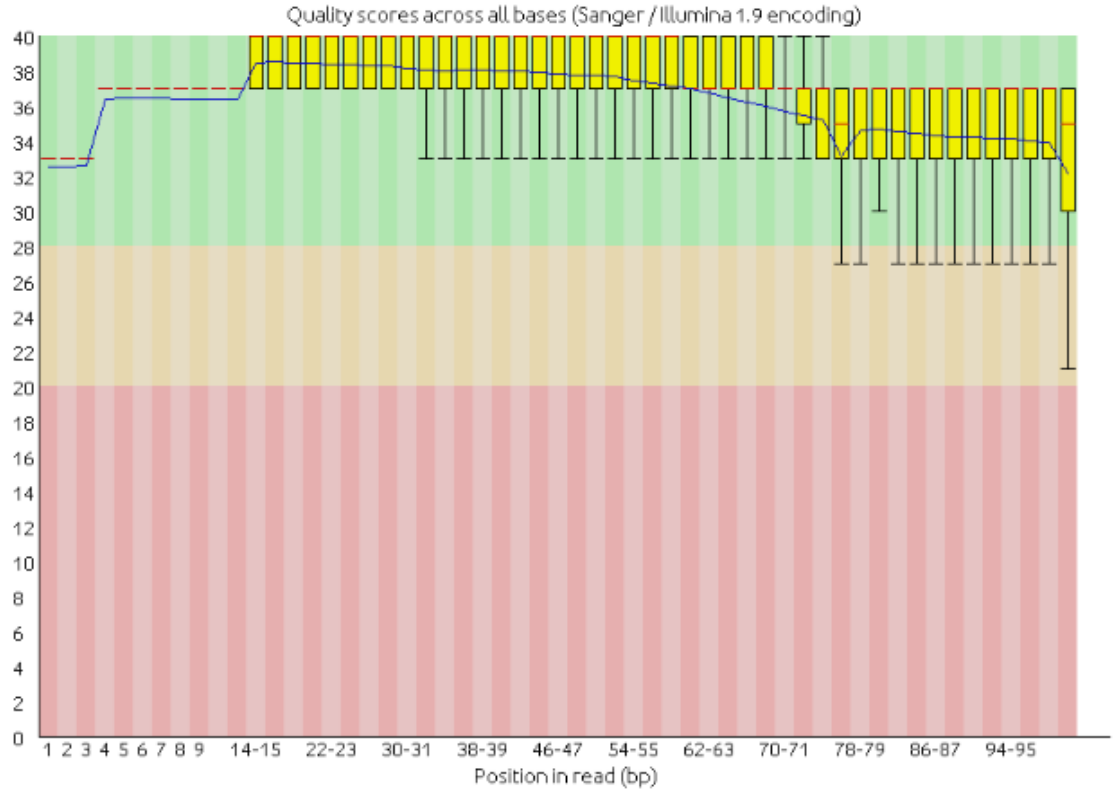


Figura 8. Distribuição da qualidade média por *read*. O gráfico mostra a distribuição da qualidade média por sequência, com média centrada em Phred 37.

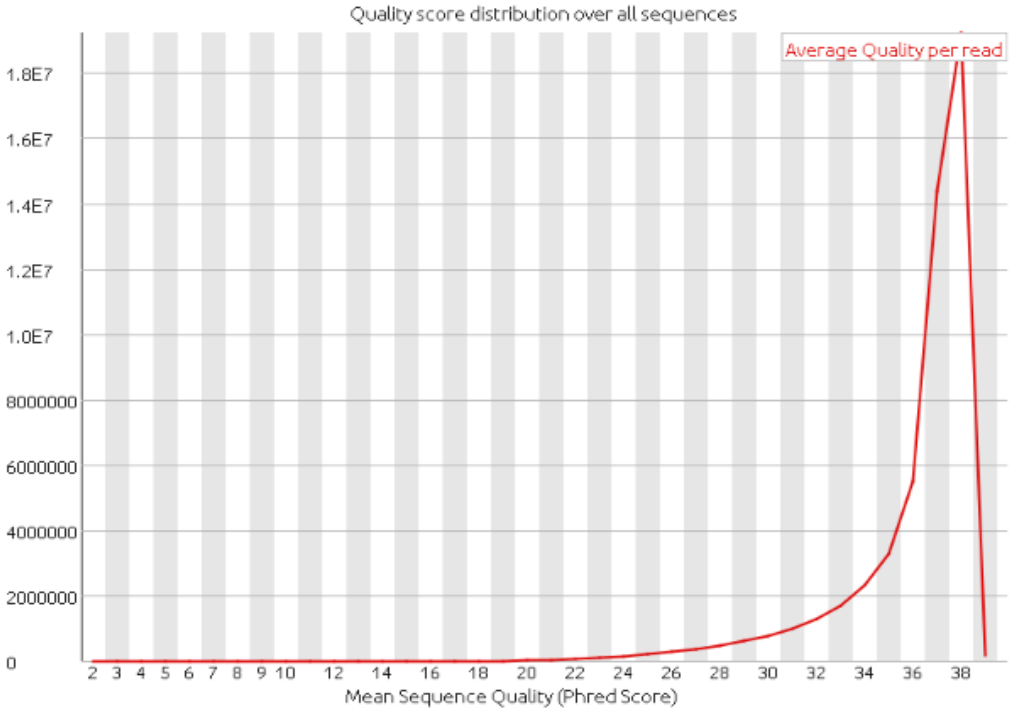


Figura 9. Conteúdo de bases por posição. As quatro linhas representam a proporção de cada uma das quatrobases possíveis (vermelho: timina, azul: citosina, verde: adenina e preto: guanina) ao longo de todas as *reads*.

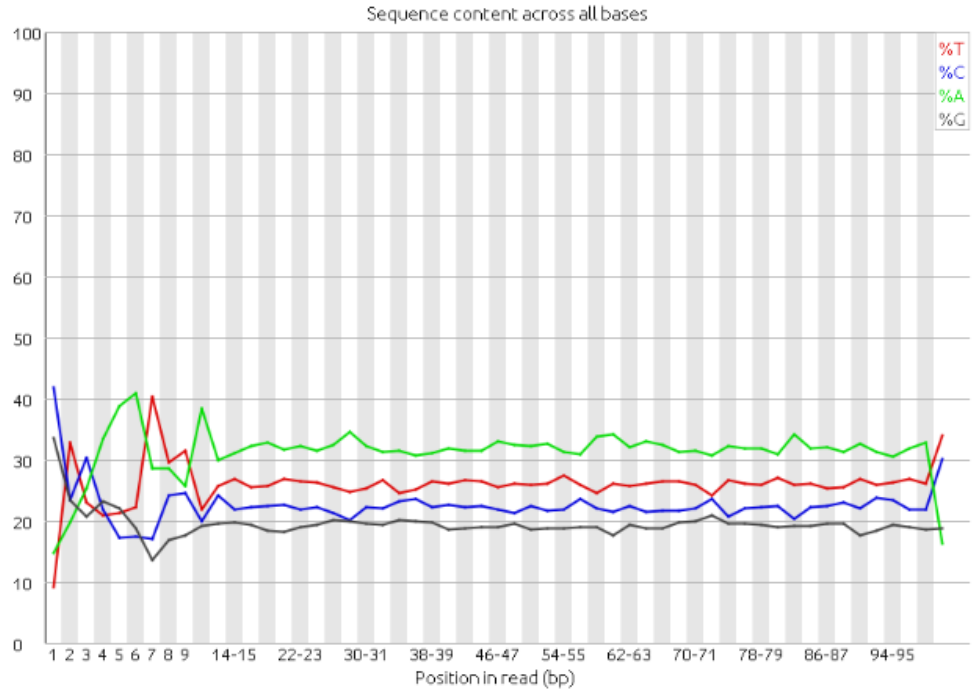
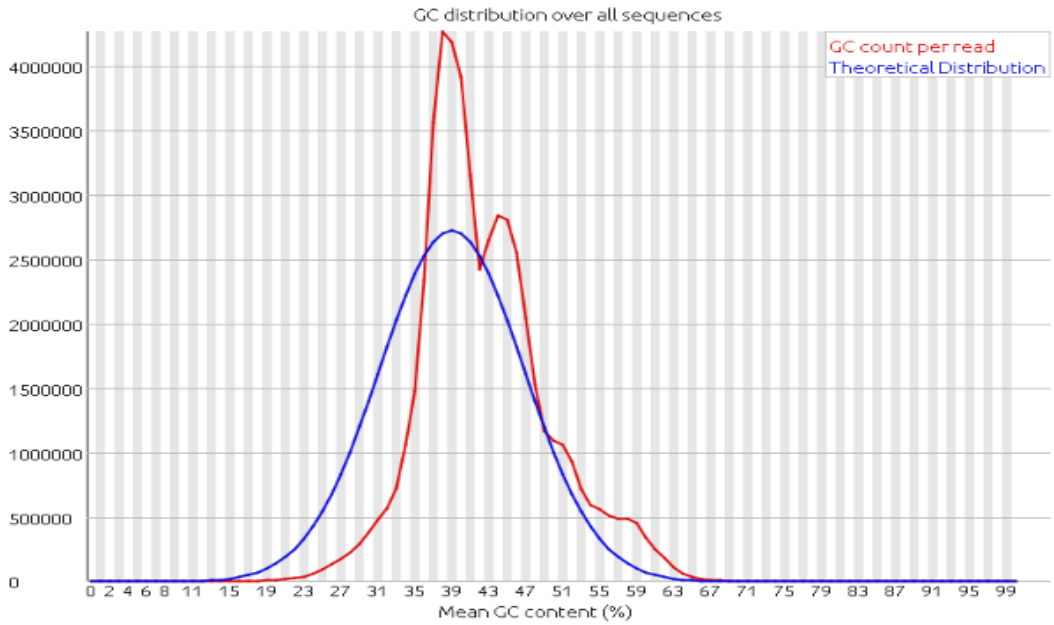


Figura 10. Distribuição normal do conteúdo GC. Em azul a curva normal teórica do conteúdo de GC e em vermelho a distribuição verdadeira encontrada nos dados de RNA-seq.

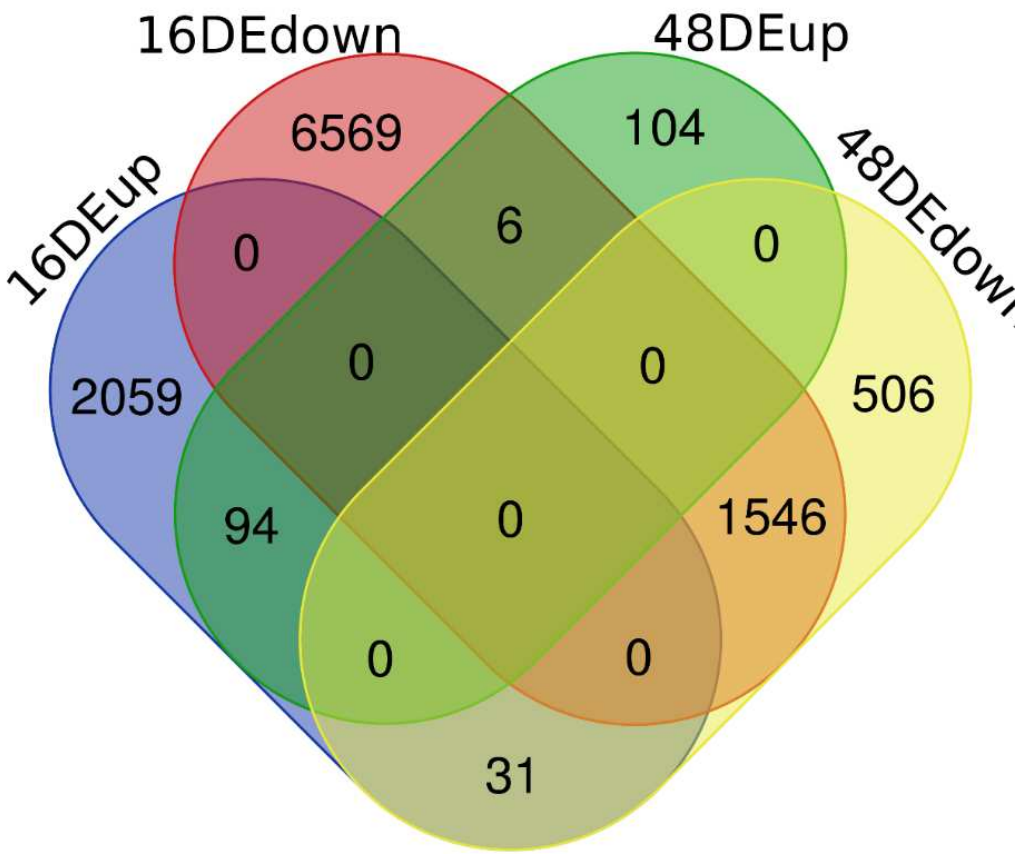


4.2 Análise de expressão gênica diferencial

Para quantificar e identificar os genes responsívos à escassez hídrica, foi usada uma *pipeline* para expressão diferencial sob estresse em soja, como descrito a seguir. A análise de sequenciamento de RNA de alto rendimento foi realizada usando o mapeador TopHat2, que utiliza o alinhador Bowtie2 por referência do transcriptoma de soja em seus processos iniciais. Em seguida, com o pacote Cufflinks foi realizada a quantificação dos transcritos por valores de FPKM e análise de expressão diferencial. Todo processamento foi efetuado utilizando arquivos de anotações para *Glycine max* cultivar Williams 82 (Wm82.a2.v1) disponível no Phytozome0 (https://phytozome.jgi.doe.gov/pz/portal.html).

Para análise dos genes diferencialmente expressos no cultivar sensível, BR16, foram identificados 10.035 genes DE nas condições de seca, quando comparados ao controle, sendo destes 2184 genes Up regulados e 8121 genes Down regulados. Enquanto o número de genes diferencialmente expressos entre a cultivar tolerante, EMBRAPA48, foram 2287 nas condições de seca, quando comparado ao controle, 204 *Up* regulados e 2082 genes *Down* regulados (Figura 11). Nestas análises foram utilizados critérios para *p* valor ajustados a uma FDR ≤ 0,05 e para os dados de níveis de expressão por *Log2 fold change* (FC), quando comparados com as plantas controle.

Figura 11: Número de genes diferentemente expressos sob a condição de escassez hídrica nas cultivares BR16 e EMBRAPA48. O diagrama de Venn mostra a comparação do número de genes diferencialmente expressos.



4.3 Expressão diferencial do *splicing* alternativo

Para identificação da expressão diferencial do *splicing* alternativo foi usado um algoritmo de alinhamento denominado *Spliced Transcripts Alignment to a Reference* (STAR). Foram realizados experimentos de validação de alto rendimento que corroboraram a precisão da STAR para detecção de novas junções. A alta velocidade e precisão de mapeamento da STAR foram cruciais para a análise do grande conjunto de dados do transcriptoma ENCODE (DJEBALI *et al.*, 2012) (> 80 bilhões de leituras da Illumina). Também foi demonstrado que o STAR tem o potencial de alinhar com precisão leituras longas, que estão surgindo das tecnologias de sequenciamento de terceira geração. A tabela 3 detalha os resultados obtidos no alinhamento das *reads* das doze amostras no genoma de referência. A porcentagem de mapeamento único variou entre 79,21% a 81,45%. A porcentagem de alinhamento único ficou, em média, 80,40%.

Tabela 3: Resultados obtidos do alinhamento das *reads* no genoma de referência utilizando o programa STAR.

ID	Grupos	Número/ Sequencias[1]	*Reads* unicamente mapeadas	(%)
1	BR16 Irrigada	52511315	42458492	80,86%
2	BR16 Irrigada	49374279	39233765	79,46%
3	BR16 Irrigada	44525275	35347472	79,39%
4	BR16 não irrigada	38327787	30965278	80,79%
5	BR16 não irrigada	40491356	32074945	79,21%
6	BR16 não irrigada	41325150	33657327	81,45%
7	EMB48 Irrigada	45957689	36907357	80,31%
8	EMB48 Irrigada	44314108	35625838	80,39%
9	EMB48 Irrigada	44223317	35314104	79,85%
10	EMB48 não irrigada	47584434	38870729	81,69%
11	EMB48 não irrigada	44614603	36031457	80,76%
12	EMB48 não irrigada	44998995	36292215	80,65%
Total		538248308	432778979	80,40%

Em todos os grupos prevalecem os sítios doadores e aceptores de *splicing* 5'GT e 3'AG em relação aos pares GC/AG e AT/AC. Os pares de sítios de *splicing* GT/AG, GC/AG e AT/AC são frequentemente encontrados em mamíferos, sendo que o sítio doador 5' GT e o sítio aceptor 3' AG são altamente conservados. Vários outros casos de sítios de *splicing* com GG/AG, GT/TG, GT/CG ou CT/AG são observados e estão envolvidos com a regulação de outros eventos de *splicing* alternativo. Entretanto, apenas os pares GT/AG, GC/AG e AT/AC podem recrutar a maquinaria de *splicing* efetivamente, e os demais sítios, considerados não-canônicos, podem funcionar em associação com esses pares canônicos.

Tabela 4: Número total de eventos de *splicing* alternativo detectados nas amostras dos grupos genotípico BR16 irrigado e BR16 não irrigado. Sítios de splicing GT/AG: sítio doador 5' GT e sítio aceptor 3' AG; sítios de splicing GC/AG: sítio doador 5' GC e sítio aceptor 3' AG; sítios de splicing AT/AC: sítio doador 5' AT e sítio aceptor 3' AC

	BR16IRP1	**BR16IRP2**	**BR16IRP3**	**BR16NIP1**	**BR16NIP2**	**BR16NIP3**
Número total de *splicing* alternativo	**7268582**	**6598570**	**5758878**	**5806639**	**5220368**	**6436662**
Número de *splicing* anotado	**6888554**	**6274541**	**5466392**	**5492099**	**4913180**	**6092221**
Número de sítios de *splicing* GT/AG	**7136213**	**6461429**	**5644023**	**5712482**	**5121947**	**6331266**
Número de sítios de *splicing* GC/AG	**89105**	**90154**	**74626**	**63623**	**58561**	**70399**
Número de sítios de *splicing* AT/AC	**2836**	**2179**	**2053**	**2286**	**2314**	**2520**
Número de *splicing* não canônico	**40428**	**44808**	**38176**	**28248**	**37546**	**32477**

Tabela 5: Número total de eventos de *splicing* alternativo detectados nas amostras dos grupos genotípico EMBRAPA48 irrigado e EMBRAPA48 não irrigado. Sítios de splicing GT/AG: sítio doador 5’ GT e sítio aceptor 3’ AG; sítios de splicing GC/AG: sítio doador 5’ GC e sítio aceptor 3’ AG; sítios de splicing AT/AC: sítio doador 5’ AT e sítio aceptor 3’ AC.

.

	EMB48IRP1	EMB48IRP2	EMB48IRP3	EMB48NIP1	EMB48NIP2	EMB48NIP3
Número total de *splicing* alternativo	**6725068**	**6849456**	**6949832**	**7978725**	**7527101**	**7012480**
Número de *splicing* anotado	**6394981**	**6523730**	**6616685**	**7594188**	**7168224**	**6655193**
Número de sítios de *splicing* GT/AG	**6593236**	**6713785**	**6809456**	**7836623**	**7388774**	**6884742**
Número de sítios de *splicing* GC/AG	**80036**	**92472**	**96249**	**98490**	**93928**	**84812**
Número de sítios de *splicing* AT/AC	**2301**	**2347**	**2416**	**2794**	**2605**	**2653**
Número de *splicing* não canônico	**43495**	**40582**	**44711**	**40818**	**41794**	**40273**

Foram analisados dados de RNA-Seq obtidos de folhas de 12 amostras, sendo dois genótipos contrastantes sob duas condições diferentes e três réplicas biológicas de cada condição usando rMATS. Este é um *pipeline* de bioinformática projetado para detectar eventos de *splicing* alternativos envolvendo duas isoformas de uma região alternativamente unida. Esses eventos são categorizados como uso alternativo de 3′ *splicing* site (A3SS), alternativo 5′ *splicing* site (A5SS), Skipped éxon (SE), éxons mutuamente exclusivos (MXE) ou retenção de íntrons (RI).

Neste estudo, identificamos um total de 29477 eventos de *splicing* alternativo*,* desses 508 foram diferenciais (FDR <0,05) no genótipo BR16, um total de 5025 eventos de Retenção Intrônica, desses 14 foram diferenciais; 6870 alternativos 3′SS, desses 140 diferenciais; 3773 alternativos 5′SS, desses 73 difereciais; 13204 Skipped

éxon, desses 278 diferenciais e 605 mutual éxon exclusivo, desses 3. Em contrapartida houve um total de 29101 eventos de *splicing* alternativo, onde 306 foram diferenciais (FDR <0,05) no genótipo EMBRAPA48, um total de 5031 eventos de Retenção Intrônica, desses 19 diferenciais; 6869 alternativos 3′SS, desses 87 diferenciais; 3763 alternativos 5′SS, desses 60 diferenciais; 1856 Skipped éxon, desses 132 diferenciais e 582 mutual éxon exclusivo, desses 8 diferenciais. Os resultados de contagem são apresentados na Figura 12A e 12B.

Figura 12A: Gráfico de contagem total dos eventos de *splicing* alternativo entre os genótipos BR16 e EMBRAPA48. Foram identificados os eventos Sítio Splicing Alternativo 3' (A3'SS), Sítio Splicing Alternativo 5' (A5'SS), Éxon Mutual Exclusivo (MXE), Retenção de Íntron (RI) e Skipped Éxon (SE).

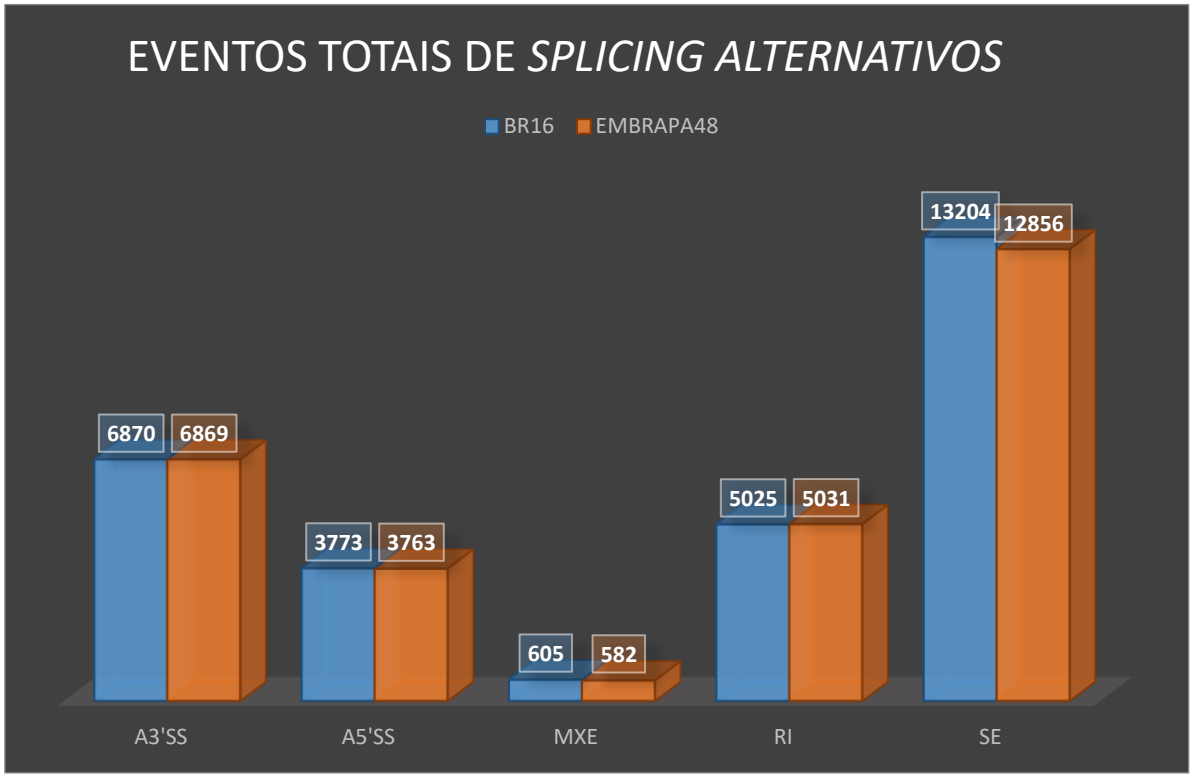


Figura 12B: Gráfico de contagem dos eventos da expressão diferencial do *splicing* alternativo entre os genótipos BR16 e EMBRAPA48 levando em conta os eventos significativos com FDR<0,05. Foram identificados os eventos Sítio Splicing Alternativo 3' (A3'SS), Sítio Splicing Alternativo 5' (A5'SS), Éxon Mutual Exclusivo (MXE), Retenção de Íntron (RI) e Skipped Éxon (SE).

4.4 Análise funcional de conjuntos de genes DE e DAS

A anotação dos eventos de *splicing* diferencialmente expressos foi conduzida com o uso de ferramentas de Gene Ontology (GO), disponíveis em The Arabidopsis Information Resource (http://www.arabidopsis.org). Os genes anotados foram agrupados em três principais categorias funcionais, sendo-as: componente celular, função molecular e processo biológico, nas quais posteriormente se subdivide em subcategorias. Nas Figuras 13A e 14A estão representadas as Ontologias anotadas representando os processos biológicos identificados pelo ClueGO.

As categorias funcionais dos genes responsivos para as duas cultivares apresentaram resposta semelhantes, entretanto, as análises de ontologia que sofreram splicing alternativo diferencial, realizadas pelo plugin ClueGO a Cytoscape (SHANNON et al., 2003) para facilitar a interpretação biológica e visualizar termos funcionalmente agrupados na forma de redes e gráficos, usa estatísticas kappa para vincular os termos na rede. Comparado com a abordagem de Ramos et al. (2008), cria uma análise *in silico* formando uma rede de anotações baseada em caminhos e dados de interação de proteínas e mapeia a lista de genes de interesse posteriormente. O ClueGO integra termos de GO, bem como caminhos de KEGG /

BioCarta, e cria uma rede de termos de GO / caminho funcionalmente organizada. Uma variedade de critérios de restrição flexíveis permite visualizações em diferentes níveis de especificidade. Além disso, o ClueGO pode comparar grupos de genes e visualizar suas diferenças funcionais. O ClueGO aproveita a estrutura versátil de visualização do, onde estão representados pelas figuras 12B e 13B.

Como pode ser visto na Figura 13, foram confrontados os resultados da expressão diferencial de genes (Figura 11) com os resultados da expressão diferencial do *Splicing* alternativo (Figura 12B). Dessa forma foram identificados 446 genes para expressão diferencial de *splicing* (DAS) relativo ao genótipo BR16, onde 23 além de DAS foram também Up regulados (UP) e 139 além de DAS foram Down regulados (DOWN). Para o genótipo EMBRAPA48, houve 290 genes para *splicing* diferencialmente expressos (DAS), onde 23 além de DAS foram Down regulados (DOWN), e nenhum DAS Up regulado (UP).

Além disso, a análise de enriquecimento funcional de EMBRAPA48, demonstrada na Figura 14A e 15B, revelou que os genes DAS estavam envolvidos principalmente em processos biológicos, incluindo resposta celular ao estresse como a biossíntese de polissacarídeos de membrana (dentre as quais se encontram a biossíntese de glicanos, betaglicanos e UDP-raminose), reparo da região 3’ de DNA, transporte de auxina, regulação do desenvolvimento floral e resposta ao estímulo abiótico como manutenção do potencial hídrico das folhas.

Os genes regulados da cultivar sensível BR16 compartilham aglomerados relacionados a processos biológicos como manutenção da resposta celular à citocinina, processo de biossíntese de celulose, processo de biossíntese de cisteína, fragmentacao do DNA dupla fita, metabolismo do RNA mitocondrial e processo de biossíntese de esfingolipídio. Notavelmente, esses resultados sugerem que existem diferenças consideráveis entre os processos fisiológicos de plantas sensíveis e tolerantes quando submetidas à deficiência hídrica.

Figura 13: Contagem dos eventos de Splicing Alternativo dividido em classes de splicing. SE (Skipped Exon), RI (Retenção de Intron), MXE (Exon Mutualmente Exclusivo), A5'SS (Sítio de Splicing Alternativo 5') A3'SS (Sítio de Splicing Alternativo 3') respectivamente. Foi avaliado também a expressão diferencial dos eventos de splicing, agrupados DAS (Diferencial Alternative Splicing), eventos de splicing diferencialmente Up regulados (UP) e eventos de splicing diferencialmente Down regulados (DOWN) levando em conta eventos significativos com FDR<0,05

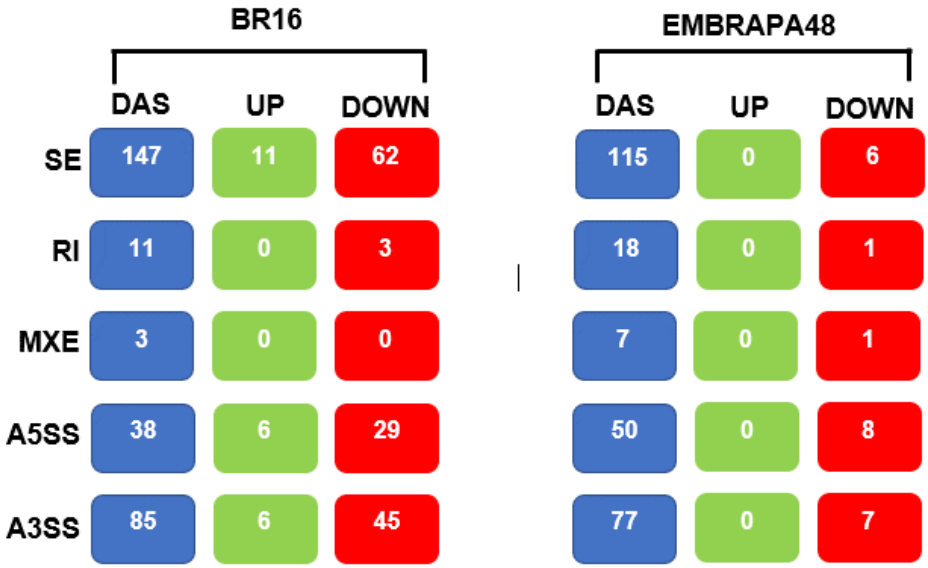


Figura 14A: Análise de categorização e agrupamento funcional pelo ClueGO. Distribuição dos genes processados por splicing em função do processo biológico para o genótipo tolerante EMBRPA48.

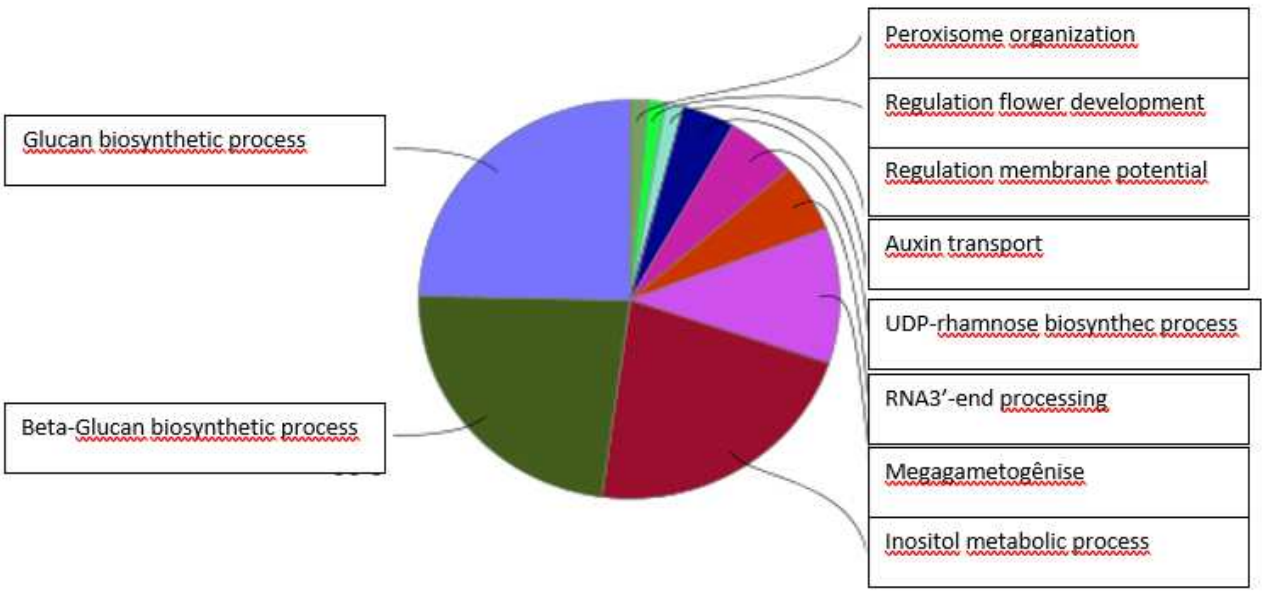


Figura 14B: Análise de categorização e agrupamento funcional pelo ClueGO. Distribuição dos genes processados por splicing para o genótipo tolerante EMBRAPA48.

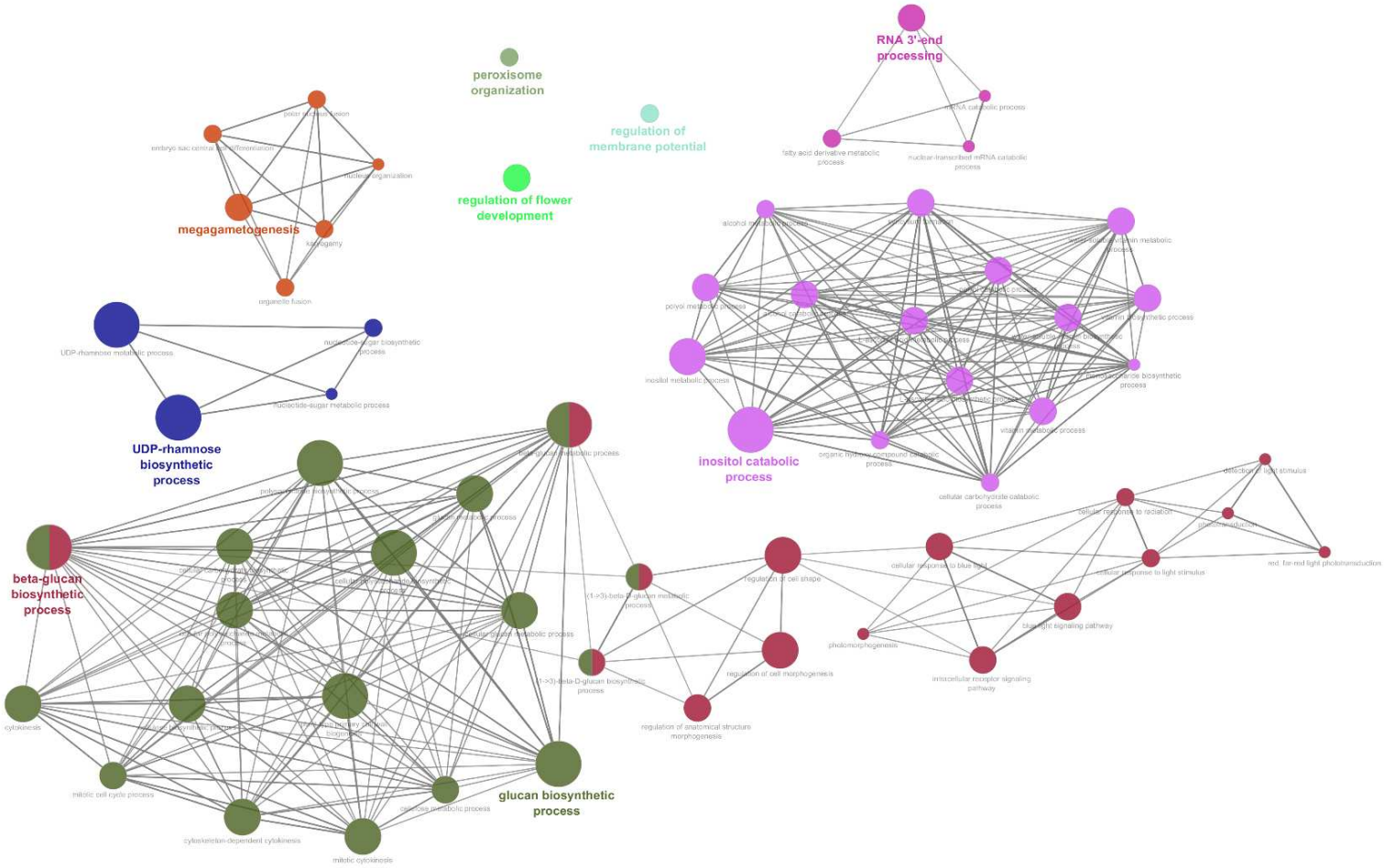


Figura 15A: Análise de categorização funcional pelo ClueGO. Distribuição dos genes processados por splicing em função da função biológica para o genótipo tolerante BR16.

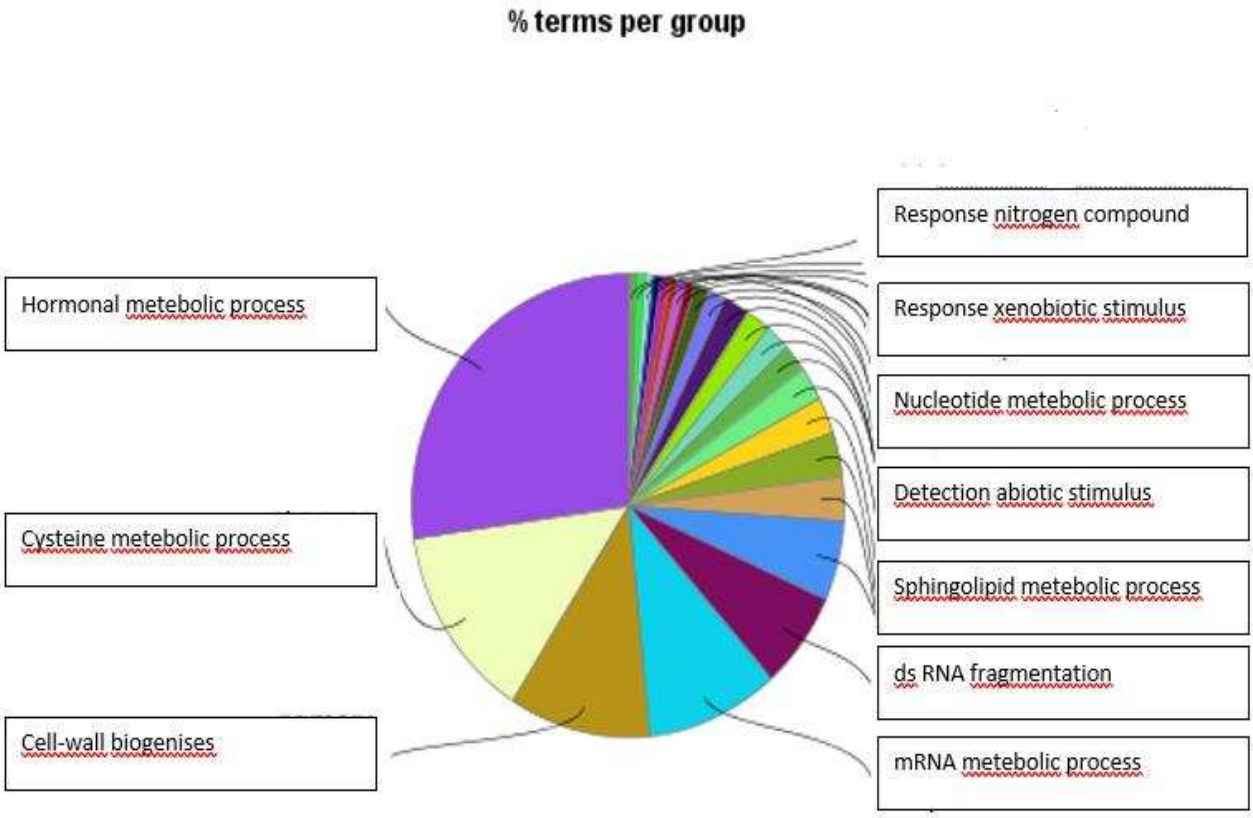


Figura 15B: Análise de categorização e agrupamento funcional pelo ClueGO. Distribuição dos genes processados por splicing para o genótipo tolerante BR16.

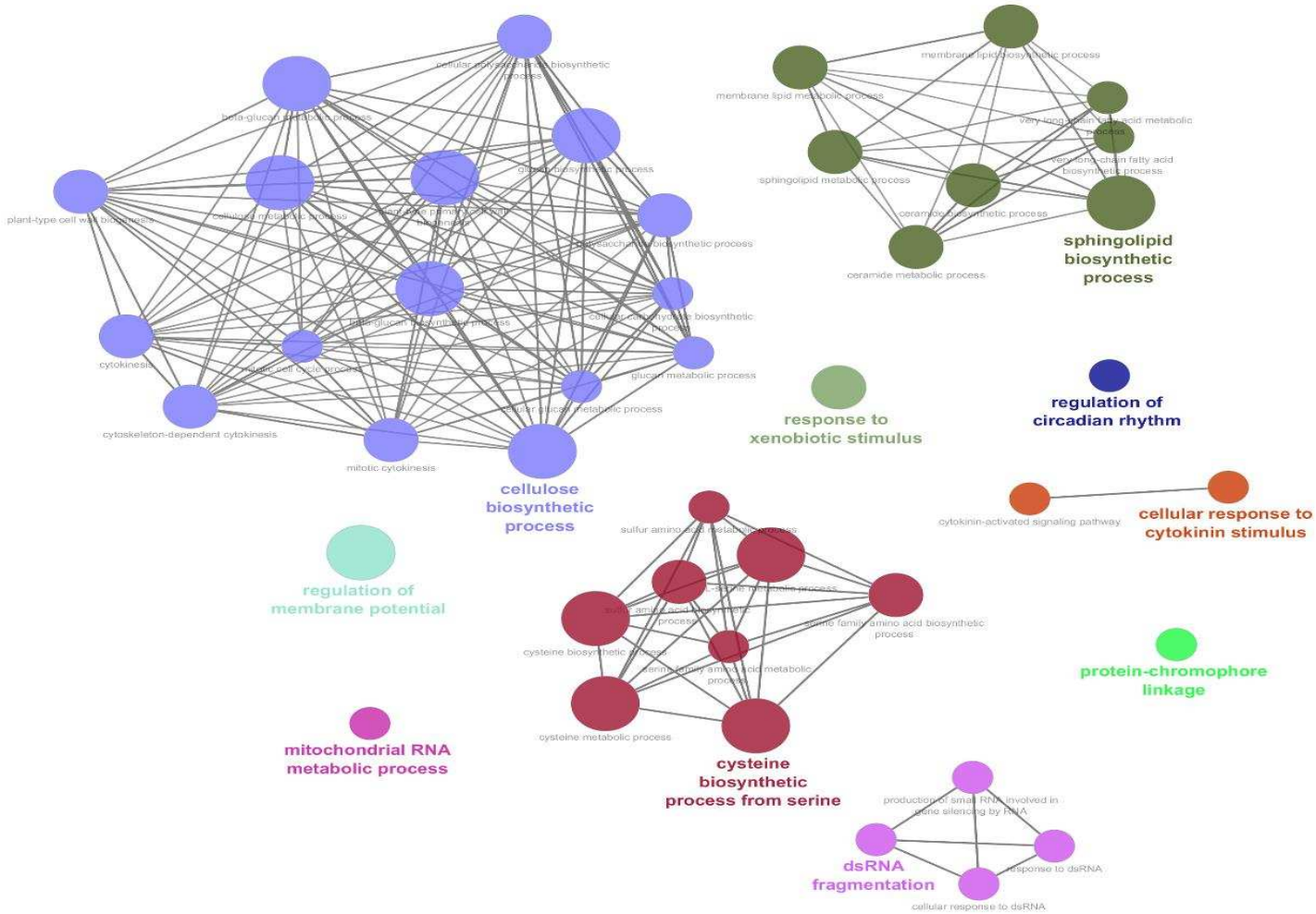


Tabela 6: Lista de genes correspondentes ao evento de *splicing* alternativo (DAS) Down regulado para o genótipo tolerante EMBRAPA48, seus respectivos ortólogos em *Arabidopsis* e a função biológica correspondente.

GENE ID	ORTÓLOGO	FUNÇÃO BIOLÓGICA/FONTE
GLYMA_13G299300	AT1G03590	Probable protein phosphatase 2C 1 [Source:UniProtKB/Swiss-Prot
GLYMA_12G109700	AT1G03620	ELMO/CED-12 family protein [Source:TAIR
GLYMA_20G034600	AT1G08750	Peptidase C13 family [Source:UniProtKB/TrEMBL
GLYMA_11G088600	AT1G20925	Protein PIN-LIKES 1 [Source:UniProtKB/Swiss-Prot
GLYMA_08G292700	AT1G24160	F3I6.9 protein [Source:UniProtKB/TrEMBL
GLYMA_13G178200	AT1G28260	Protein SMG7L [Source:UniProtKB/Swiss-Prot
GLYMA_13G272100	AT1G29020	Calcium-binding EF-hand family protein [Source:TAIR
GLYMA_13G272100	AT1G29025	Calcium-binding EF-hand family protein [Source:UniProtKB/TrEMBL
GLYMA_08G124200	AT1G30500	Nuclear transcription factor Y subunit A-7 [Source:UniProtKB/Swiss-Prot
GLYMA_13G353200	AT1G52080	Actin binding protein family [Source:UniProtKB/TrEMBL
GLYMA_20G237200	AT1G56220	Dormancy/auxin associated family protein [Source:TAIR
GLYMA_08G292700	AT1G70100	unknown protein
GLYMA_02G078500	AT1G71697	Probable choline kinase 1 [Source:UniProtKB/Swiss-Prot

GLYMA_11G088600	AT1G76520	Protein PIN-LIKES 3 [Source:UniProtKB/Swiss-Prot
GLYMA_11G088600	AT1G76530	Protein PIN-LIKES 4 [Source:UniProtKB/Swiss-Prot
GLYMA_13G272100	AT2G34020	Calcium-binding EF-hand family protein [Source:TAIR
GLYMA_13G272100	AT2G34030	Calcium-binding EF-hand family protein [Source:UniProtKB/TrEMBL
GLYMA_08G124200	AT2G34720	Nuclear transcription factor Y subunit A-4 [Source:UniProtKB/Swiss-Prot
GLYMA_02G253900	AT2G39920	Uncharacterized protein At2g39920 [Source:UniProtKB/Swiss-Prot
GLYMA_10G082800	AT3G05690	Nuclear factor Y, subunit A2 [Source:UniProtKB/TrEMBL
GLYMA_13G299300	AT4G03415	Probable protein phosphatase 2C 52 [Source:UniProtKB/Swiss-Prot
GLYMA_15G073000	AT4G04955	Allantoinase [Source:UniProtKB/Swiss-Prot
GLYMA_02G078500	AT4G09760	Probable choline kinase 3 [Source:UniProtKB/Swiss-Prot
GLYMA_10G130800	AT4G13550	triglyceride lipases
GLYMA_14G116000	AT4G32250	AT4g32250/F10M6_110 [Source:UniProtKB/TrEMBL
GLYMA_10G082800	AT5G06510	NF-YA10 [Source:UniProtKB/TrEMBL
GLYMA_17G146400	AT5G47180	Vesicle-associated protein 2-1 [Source:UniProtKB/Swiss-Prot

5. DISCUSSÃO

A abordagem de analise da expressão gênica por meio da tecnologia de sequenciamento RNAseq permitiu a geração de perfis diferenciais dos eventos de *splicing* alternativos para ambos os genótipos de soja contrastantes para a tolerância a seca, já caracterizado anteriormente (OYA et al. 2004; KU et al. 2013;MESQUITA 2013; FANG & XIONG, 2015). A principal característica da resposta do genótipo sensível BR16 ao estresse de seca é a indução da transcrição de um número relativamente grande de genes. Como o EMBRAPA 48 é tolerante à seca, pode ser mais um indicativo de que ainda, apesar das folhas apresentearem um potencial de 1.0 MP, ao nível celular mantenha uma condição homeostase em comparação ao BR16 e, portanto, ter uma menor reprogramação gênica.

De fato, os resultados indicam que as cultivares de soja BR16 e EMBRAPA48, apesar de compartilharem as mesmas categorias funcionais dos genes responsivos ao estresse, os efeitos ocorrem em menor extensão para a cultivar EMBRAPA 48. Esses dados corroboram com os obtidos por Coutinho et al. (2019), onde foi avaliado a expressão diferencial dos genes pelo alinhador Kallisto, um programa de quantificação de RNA-seq que é mais rápido do que outras abordagens e alcança exatidão similar. O método é um pseudo-alinhador que são lidos para uma referência, produzindo uma lista de transcrições compatíveis com cada leitura, evitando o alinhamento de bases individuais (BRAY et al., 2016). Essa resposta mais amenizada é provavelmente devida a genes regulados positivamente na planta EMBRAPA 48, (COUTINHO et al., 2019), os dados sugerem especificamente que a indução do estresse foi transcricionalmente mais acentuado na cultivar sensível (BR16), em função do maior número de genes diferencialmente expressos. Assim, essa hipótese pressupõe que a EMBRAPA48 esteja preparada para reagir a condições estressantes antes mesmo de perceber o estresse, sofrendo assim uma menor perturbação. É de entendimento que a capacidade de tolerar o déficit hídrico é uma característica complexa controlada por muitos genes como mostra os resultados de expressão diferencial gerados por RNA-seq (MOLINA et al. 2008, ERGEN & BUDAK, 2009).

O *splicing* alternativo é o um dos principais mecanismo envolvido na diversidade protéica e em diversas vias regulatórias em organismos eucarióticos multicelulares. A primeira evidência para a ocorrência do *splicing* alternativo no

desenvolvimento de plantas veio da expressão diferencial de proteína rica em Ser/Arg (SR) fatores de *splicing* em diferentes estruturas e durante o desenvolvimento (LOPATO et al., 1996, 1999b; KALYNA et al., 2003; PALUSA et al., 2007) indicando a regulação específica *splicing* em plantas. Estudos recentes mostraram como o *splicing* alternativo (AS) possui grande influência no desenvolvimento e vias de sinalização de muitas plantas, os quais foram abordados em estudo por Staiger e Brown (2013). Além disso, a triagem para mutantes em várias vias identificaram frequentemente fatores de *splicing* como moduladores de proteínas funcionais, indicando que algumas vias são reguladas por *splicing* diferencial (LEE et al., 2006; MONAGHAN et al., 2009; SUGLIANI et al., 2010; FOUQUET et al., 2011; KONCZ et al., 2012). Vários relatos demonstram que o *splicing* alternativo pode ser influenciado pelo estresse abiótico (MARRS E WALBOT, 1997; PALUSA et al. 2007; REDDY 2007; TANABE et al. 2007) e estresse biótico (IIDA et al. 2004; ATTALLAH et al. 2007; REDDY 2007). ZHANG e GASSMANN (2007) demonstraram que um gene de resistência a doença em Arabidopsis, RPS4, produz múltiplos transcritos via *splicing* alternativo. Dessa forma, avaliar a interferência do estresse à seca na regulação do *splicing* alternativo é de grande importância para a compreensão dos fatores envolvidos na resposta fisiológica da planta, e pode contribuir para a seleção de genes responsivos ao mecanismo de tolerância ao estresse hídrico.

No presente estudo, eventos de exon skipping, éxons mutuamente exclusivos, sítios de *splicing* 3’ e 5’ e retenção de íntrons foram detectados nos dados de transcriptoma dos dois genótipos de soja contrastantes. Os resultados obtidos demonstraram que os eventos de exon skipping e A3’SS parecem ser predominantemente alterados, tanto nos genótipos BR16 e EMBRAPA48 quando submetidos ao tratamento de escassez hídrica. Entretanto, esses dois eventos de splicing citados acima foram mais acentuados da variedade mais sensível BR16. Essa observação pode ser também em decorrencia de que um número maior de genes apresentou Splicing Alternativo diferencial para cultivar BR16. Nossos dados evidenciam que os genótipos se comportam em parte de maneira similar à regulação do *splicing* alternativo quando submetidos ao estresse, mas a cultivar tolerante apresenta menor perturbação e regula importante genes adaptativos, mostrando que o genótipo EMBRAPA48 apresenta um mecanismo adaptativo de resposta mais eficaz de tolerância ao estresse.

Desde que Walter Gilbert propôs que poderia ser possível para um gene produzir diferentes isoformas de mRNA através da recombinação de *splicing* de diferentes exons em 1978 (GILBERT, 1978), evidências crescentes mostraram que o AS ocorre comumente em eucariotos superiores. Em Drosophila melanogaster, o gene Dscam tem o potencial de produzir mais de 38.000 variantes de *splicing* alternativo Graveley, 2005). Em humanos, foi relatado que a proporção de genes AS é tão alta quanto 92 a 94% (WANG et al., 2008). Estudos previram que ~60% dos genes contendo introns são alternativamente processados em plantas (REDDY et al., 2013).

O rMATS foi usado para detectar a mudança dos padrões de AS em cada um dos genes AS (ou seja, genes DAS), que foi projetado para a detecção de emendas diferencialmente alternativas a partir de dados RNA-Seq replicados. Inicialmente, o rMATS utiliza um quadro hierárquico para modelar os níveis de inclusão do exon, denotados como percentual emendado (KATZ et al., 2010), que são estimados pelas contagens de leituras mapeadas para a inclusão do exon ou saltar isoforma. Em seguida, esta abordagem estima as mudanças de AS com base nesta informação. Especificamente, rMATS utiliza um teste de razão de verosimilhança para calcular o valor P dos eventos DAS e avaliar o seu significado. Apenas os eventos com valores de FDR<0,05 foram identificados como eventos significativamente diferenciais, o que, poderia eliminar eventos falsos positivos. Aqui, usando sequenciamento de RNA de alto rendimento e análises abrangentes, demonstramos que cerca de 44% dos genes são alternativamente processados em soja por mecanismos do tipo *skipped exon*. Tem sido sugerido que numerosos eventos de regulação do splicing são induzidos apenas por estresses abióticos e bióticos (STAIGER e BROWN, 2013).

A análise do GO indicou que estes genes AS constitutivos foram altamente enriquecidos em vias relacionadas ao *splicing* de RNA. Este resultado pode ser parcialmente apoiado pela observação de que, além de desempenhar papéis importantes na regulação do AS de outros genes, os genes relacionados ao *splicing* de RNA também são regulados por AS ou por outros fatores de *splicing* (SALTZMAN et al., 2011; THOMAS et al., 2012).

Nesse processo, para o genótipo tolerante EMBRAPA48, células vegetais perceberam estímulos de estresse através de sensores ou receptores localizados

principalmente na membrana celular, que levam resposta de AS em grupos de funções celulares, como, processamento, processamento final da região 3', processo de biossíntese da UDP-rhamnose, processos relacionados ao transporte de auxina, processos processos de biossíntese relacionados ao metabolismo de glicanos e beta-glicanos, grupos relacionados ao catabolismo de inositol, regulação do desenvolvimento floral e regulação do potencial de membrana. Foi observado proteínas quinases que se relacionam principalmente na regulação do ABA mediada por Ca^{+2}, uma vez que este está relacionado com o equilíbrio do potencial hídrico durante a condição do estresse hídrico (YANG, et al, 2011). Fatores de transcrição (TFs) MYB também foram caracterizados pelo papel do controle estomático, que também está relacionado ao balanço do potencial hídrico foliar homeostasia, e sinalização de Ca (Fasani, et al, 2019). Este grupo apresentou regulação reprimida (Down) e AS, que pode estar relacionado a condutância estomática da EMBRAPA48, quando comparada à sensível (MESQUITA, 2010).

Nos últimos anos, uma ampla gama de famílias de TF com relevância na resposta ao estresse hídrico foi identificada, como AREB, DREB, MYB, WRKY, NAC, ZFP e bzip (GOLLDACK *et al.*2011, JIN et al., 2014, ANBAZHAGAN et al., 2015). Em genes ortólogos de soja, fator de transcrição MYB61 foi down-regulado na EMBRAPA 48 sob condição de escassez hídrica (COUTINHO et al., 2019). Outros fatores de transcrição como MYB61 e MYB103 TFs foram descritos como reguladores positivos da biossíntese de parede celular secundária em Arabidopsis e gramíneas. Curiosamente, a expressão de um ortologo de soja em Arabidopsis AtPME foi up regulado na sensível BR16 e down regulado na tolerante Embrapa 48 (Coutinho et al., 2019). MYB55/61 foi capaz de modular o conteúdo de lignina em feixes vasculares através de activação de genes da biossíntese de lignina (Rao e Dixon 2018), e promovendo a parede secundária relacionada com síntese de celulose. Tais genes MyB de soja podem contribuir para coordenar a biossíntese de celulose e lignina na formação da parede celular secundária na Embrapa 48 tolerante.

Foi observado um grupo de genes que sofreram *splicing* alternativo diferencial no cultivar EMBRAPA48, relacionado a biossíntese de UDP-rhamnose (MUM4, RHM1, RHM3), que compõe a parede celular ramnogalacturonan-I, ramnogalacturonan-II e compostos naturais em plantas. Foi sugerido que esses genes estão envolvidos na conversão de UDP-D-glicose em UDP-L-ramnose com base no

seu efeito no desenvolvimento, dirigido por ramnogalacturonan-I. Os genes relacionados a RHM2 / MUM4, RHM1 e RHM3, podem ser encontrados no genoma de *A. thaliana*, onde essas proteínas possuem UDP-D-glucose 4,6-desidratase, UDP-4-ceto-6-desoxi-D-glucose 3,5-epimerase, e UDP-4-ceto-L- ramnose 4-ceto-redutase no citoplasma quando expresso na levedura *Saccharomyces cerevisiae*. A análise do domínio funcional revelou que a região N-terminal de RHM2 (RHM2-N; aminoácidos 1 a 370) tem a primeira atividade e a região C-terminal de RHM2 (RHM2-C; aminoácidos 371-667) tem os dois seguintes actividades. Isto sugere que RHM2 converte UDP-D-glucose em UDP-L-ramnose através de um intermediário UDP-4-ceto-6-desoxi-D-glucose (TAKUJI et al.,2006).

A dinâmica da parede celular sob estresse em tolerante de plantas parece ser importante para manter possibilidade de células e órgãos se expandirem. O principal grupo de polímeros nas paredes celulares dicotípicas primárias são pectinas, um grupo heterogêneo de ácido homogalacturônico, ramnogalacturonan I (RG-I) e ramnogalacturonan II (RG-II) (MOHNEN, 2008). A pectina é frequentemente modificada em plantas expostas ao estresse hídrico, proporcionando um aumento na elasticidade da parede celular, pode contribuir para a manutenção do turgor celular ou do volume simplificado (DE DIEGO et al. 2013, MARTÍNEZ et al. 2007). Essa elasticidade pode ser correlacionada com a tolerância à seca da planta, principalmente pelo aumento das cadeias laterais dos polímeros pécticos ramnogalacturonana RGI e RGII (COUTINHO et al., 2019)

6. CONCLUSÃO

Neste trabalho, nós avaliamos o perfil do transcriptoma de folhas de duas cultivares de soja com capacidade de contraste para lidar com o estresse hídrico, BR16 sensível e EMBRAPA48 tolerante, em função da expressão diferencial de genes o do *splicing* alternativo. Nosso estudo mostra que a indução do estresse é reponsável por alterar o perfil da expressão gênica das duas cultivares, com uma maior alteração do número de genes por parte da BR16 sensível e menor alteração do número de genes por parte da tolerante EMBRAPA48, isso se deve a expressão de muitos genes de resposta à seca, que está operando mesmo antes da ocorrência de estresse e torna a planta pronta para responder a condições ambientais adversas. Mostramos que os eventos *Skipped* exon e Alternativo 3'SS e tem maior nível de inclusão entre os eventos analisados. Propomos a hipótese de que a biossíntese de genes relacionados a rahmnose (MUM4, RHM1, RHM3) são regulados por *splicing* durante o estresse hídrico seja responsável pela dinâmica da parede celular no tolenrante da planta, contribuindo para a manutenção do turgor celular ou do volume.

A análise do perfil do transcriptoma global, sugere que o estresse foi capaz de remodelar diferencialemente o nível da expressão de genes e o perfil do *splicing* alternativo nas folhas de soja. Assim, esse trabalho contribui para a compreensão das bases moleculares estudadas, nas quais envolvem o metabolismo das plantas na busca de respostas eficientes a seca, bem como o aumento da tolerância à escassez hídrica.

REFERÊNCIAS BIBLIOGRÁFICAS

Abdul Jaleel, C.; Manivannan, P.; Kishorekumar, A.; Sankar, B.; Gopi, R.; Somasundaram, R.; Panneerselvam, R. (2007) Alterations in osmoregulations, antioxidant enzymes and indole alkaloid levels in Catharanthusroseus exposed to water deficit. Colloids and Surfaces B: Biointerfaces, v. 59, n. 2.

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome research*, *22*(10), 2008-2017.

Arora, A., Sairam, R. K., & Srivastava, G. C. (2002). Oxidative stress and antioxidative system in plants. Current science, 1227-1238.

Attallah, C. V., Welchen, E., & Gonzalez, D. H. (2007). The promoters of Arabidopsis thaliana genes AtCOX17-1 and-2, encoding a copper chaperone involved in cytochrome c oxidase biogenesis, are preferentially active in roots and anthers and induced by biotic and abiotic stress. *Physiologia plantarum*, *129*(1), 123-134.

Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses, 6(1), 106-136.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.

Bray, E. A. (1993). Molecular responses to water deficit. Plant physiology, 103(4), 1035.

Bray, E. A. (2002). Abscisic acid regulation of gene expression during water-deficit stress in the era of the Arabidopsis genome. Plant, cell & environment, 25(2), 153-161.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, *34*(5), 525.

Casagrande, E. C., Farias, J. R. B., Neumaier, N. O. R. M. A. N., Oya, T. E. T. S. U. J. I., Pedroso, J. Ú. L. I. O., Martins, P. K., ... & Nepomuceno, A. L. (2001). Expressão gênica diferencial durante déficit hídrico em soja. Revista Brasileira de Fisiologia Vegetal, 13(2), 168-184.

Cattivelli, L., Rizza, F., Badeck, F. W., Mazzucotelli, E., Mastrangelo, A. M.,

Chaves, M. M., Maroco, J. P., & Pereira, J. S. (2003). Understanding plant responses to drought—from genes to the whole plant. Functional plant biology, 30(3), 239-264.

Chen, J., Huang, Q., Gao, D., Wang, J., Lang, Y., Liu, T., ... & Chen, C. (2013). Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. *Nature communications*, *4*, 1595.

Cheng, Y.-Q., Liu, J.-F., Yang, X., Ma, R., Liu, C., and Liu, Q. (2013). RNA-seq analysis reveals ethylene-mediated reproductive organ development and abscission in soybean (Glycine max L. Merr.). Plant Mol. Biol. Rep. 31, 607–619. doi: 10.1007/s11105-012-0533-4.

Clement, M., Lambert, A., Herouart, D., and Boncompagni, E. (2008). Identification of new up-regulated genes under drought stress in soybean nodules. Gene 426, 15–22. doi: 10.1016/j.gene.2008.08.016

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., ... & Xue, C. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101.

Ergen, N. Z., & Budak, H. (2009). Sequencing over 13 000 expressed sequence tags from six subtractive cDNA libraries of wild and modern wheats following slow drought stress. *Plant, cell & environment*, *32*(3), 220-236.

Fang, Y., & Xiong, L. (2015). General mechanisms of drought response and their application in drought resistance improvement in plants. Cellular and molecular life sciences, 72(4), 673-689

Fouquet, R., Martin, F., Fajardo, D. S., Gault, C. M., Gómez, E., Tseung, C. W., ... & Settles, A. M. (2011). Maize rough endosperm3 encodes an RNA splicing factor required for endosperm cell differentiation and has a nonautonomous effect on embryo development. *The Plant Cell*, *23*(12), 4280-4297.

Foyer, C. H., & Fletcher, J. M. (2001). Plant antioxidants: colour me healthy. Biologist (London, England), 48(3), 115-120.

Francia, E., ... & Stanca, A. M. (2008). Drought tolerance improvement in crop plants: an integrated view from breeding to genomics. Field Crops Research, 105(1-2), 1-14

Gaspar, T. Franck, T. Bisbis, B, Kevers, C., Jouve, L., Hausma, J F and Dommes, J. (2002),Concepts inplant stress phdiolog. Application to plant tissue cultures. Plant Growth Reglation 37, 26-285

Ge, Y., Li, Y., Zhu, Y. M., Bai, X., Lv, D. K., Guo, D., et al. (2010). Global transcriptome profiling of wild soybean (Glycine soja) roots under $NaHCO_3$ treatment. BMC Plant Biol. 10:153. doi: 10.1186/1471-2229-10-153

Graveley, B. R. (2005). Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell*, *123*(1), 65-73.

Guo, B., Sleper, D., Lu, P., Shannon, J., Nguyen, H., and Arelli, P. (2006). QTLs associated with resistance to soybean cyst nematode in soybean: meta-analysis of QTL locations. Crop Sci. 46, 595–602. doi: 10.2135/cropsci2005.04-0036-2

Guo, L., Zhao, Y., Zhang, S., Zhang, H., & Xiao, K. (2009). Improvement of organic phosphate acquisition in transgenic tobacco plants by overexpression of a soybean phytase gene Sphy1. Frontiers of Agriculture in China, 3(3), 259-265.

Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbrück, S., Lehmann, G., ... & Bork, P. (1999). Alternative *splicing* of human genes: more the rule than the exception?. Trends in Genetics, 15(10), 389-390.

Hasegawa, P. M., Bressan, R. A., Zhu, J. K., & Bohnert, H. J. (2000). Plant cellular and molecular responses to high salinity. Annual review of plant biology, 51(1), 463-499.

Hazen, S. P., Borevitz, J. O., Harmon, F. G., Pruneda-Paz, J. L., Schultz, T. F., Yanovsky, M. J., ... & Kay, S. A. (2005). Rapid array mapping of circadian clock and developmental mutations in Arabidopsis. Plant physiology, 138(2), 990-997.

Heber, S., Alekseyev, M., Sze, S. H., Tang, H., & Pevzner, P. A. (2002). *Splicing* graphs and EST assembly problem. Bioinformatics, 18(suppl_1), S181-S188.

Ingram, J., & Bartels, D. (1996). The molecular basis of dehydration tolerance in plants. Annual review of plant biology, 47(1), 377-403.

Irsigler, A. S., Costa, M. D., Zhang, P., Reis, P. A., Dewey, R. E., Boston, R. S., & Fontes, E. P. (2007). Expression profiling on soybean leaves reveals integration of ER- and osmotic-stress pathways. BMC genomics, 8(1), 431.

Jain, K. K. (2005). Nanotechnology in clinical laboratory diagnostics. Clinica chimica acta, 358(1-2), 37-54.

Kalyna, M., Lopato, S., & Barta, A. (2003). Ectopic expression of atRSZ33 reveals its function in splicing and causes pleiotropic changes in development. *Molecular biology of the cell*, *14*(9), 3565-3577.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology, 14(4), R36.

Kollipara KP, Singh RJ, Hymowitz T (1997) Phylogenetic and genomic relationship in the genus Glycine Willd. Based on sequences from the ITS region of nuclear rDNA. Genome 40:57–68.

Kollipara, K. P., Saab, I. N., Wych, R. D., Lauer, M. J., & Singletary, G. W. (2002). Expression profiling of reciprocal maize hybrids divergent for cold germination and desiccation tolerance. Plant Physiology, 129(3), 974-992.

Komatsu, S., Yamamoto, R., Nanjo, Y., Mikami, Y., Yunokawa, H., & Sakata, K. (2009). A comprehensive analysis of the soybean genes and proteins expressed under flooding stress using transcriptome and proteome techniques. Journal of Proteome Research, 8(10), 4766-4778.

Konarska, M. M., Grabowski, P. J., Padgett, R. A., & Sharp, P. A. (1985). Characterization of the branch site in lariat RNAs produced by *splicing* of mRNA precursors. Nature, 313(6003), 552.

Koncz, C., deJong, F., Villacorta, N., Szakonyi, D., & Koncz, Z. (2012). The spliceosome-activating complex: molecular mechanisms underlying the function of a pleiotropic regulator. *Frontiers in plant science*, *3*, 9.

Koren, E., Lev-Maor, G., & Ast, G. (2007). The emergence of alternative 3′ and 5′ splice site exons from constitutive exons. *PLoS computational biology*, *3*(5), e95.

Ku, Y. S., Au-Yeung, W. K., Yung, Y. L., Li, M. W., Wen, C. Q., Liu, X., & Lam, H. M. (2013). Drought stress and tolerance in soybean. In *A comprehensive survey of international soybean research-genetics, physiology, agronomy and nitrogen relationships*. IntechOpen.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., ... & Barann, M. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501(7468), 506.

Lata, C., Muthamilarasan, M., & Prasad, M. (2015). Drought stress responses and signal transduction in plants. In Elucidation of abiotic stress signaling in plants (pp. 195-225). Springer, New York, NY.

Le, D. T., Nishiyama, R., Watanabe, Y., Tanaka, M., Seki, M., Yamaguchi-Shinozaki, K., et al. (2012). Differential gene expression in soybean leaf tissues at late developmental stages under drought stress revealed by genome-wide transcriptome analysis. PLoS ONE 7:e49522. doi: 10.1371/journal.pone.0049522.

Le, K. Q., Prabhakar, B. S., Hong, W. J., & Li, L. C. (2015). Alternative splicing as a biomarker and potential target for drug discovery. *Acta Pharmacologica Sinica*, *36*(10), 1212.

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research, 18(11), 1851-1858.

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., Franklin, L. D., et al. (2010). An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. Plant J. 63, 86–99. doi: 10.1111/j.1365-313X.2010.04222.x

Licatalosi, D. D., & Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. Nature Reviews Genetics, 11(1), 75.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., ... & Law, M. (2012). Comparison of next-generation sequencing systems. BioMed Research International, 2012.

Lopato, S., Mayeda, A., Krainer, A. R., & Barta, A. (1996). Pre-mRNA splicing in plants: characterization of Ser/Arg splicing factors. *Proceedings of the National Academy of Sciences*, *93*(7), 3074-3079.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet., 9, 387-402.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. Trends in genetics, 24(3), 133-141.

Marguerat, S., Wilhelm, B. T., & Bähler, J. (2008). Next-generation sequencing: applications beyond genomes.

Marrs, K. A., & Walbot, V. (1997). Expression and RNA splicing of the maize glutathione S-transferase Bronze2 gene is regulated by cadmium and other stresses. *Plant physiology*, *113*(1), 93-102.

Martínez, J. P., Silva, H. F. L. J., Ledent, J. F., & Pinto, M. (2007). Effect of drought stress on the osmotic adjustment, cell wall elasticity and cell volume of six cultivars of common beans (Phaseolus vulgaris L.). European Journal of Agronomy, 26(1), 30-38.

Monaghan, P., Metcalfe, N. B., & Torres, R. (2009). Oxidative stress as a mediator of life history trade-offs: mechanisms, measurements and interpretation. *Ecology letters*, *12*(1), 75-92.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods, 5(7), 621.

Nakayama, K., Okawa, K., Kakizaki, T., Honma, T., Itoh, H., & Inaba, T. (2007). Arabidopsis Cor15am is a chloroplast stromal protein that has cryoprotective activity and forms oligomers. Plant physiology, 144(1), 513-523.

Nam, N. H., Chauhan, Y. S., & Johansen, C. (2001). Effect of timing of drought stress on growth and grain yield of extra-short-duration pigeonpea lines. The Journal of Agricultural Science, 136(2), 179-189.

Nobuta, K., Venu, R. C., Lu, C., Beló, A., Vemaraju, K., Kulkarni, K., ... & Meyers, B. C. (2007). An expression atlas of rice mRNAs and small RNAs. Nature biotechnology, 25(4), 473.

Oka, T., Nemoto, T., & Jigami, Y. (2007). Functional analysis of Arabidopsis thaliana RHM2/MUM4, a multidomain protein involved in UDP-D-glucose to UDP-L-rhamnose conversion. *Journal of Biological Chemistry*, *282*(8), 5389-5403.

O'Rourke, J. A., Charlson, D. V., Gonzalez, D. O., Vodkin, L. O., Graham, M. A., Cianzio, S. R., ... & Shoemaker, R. C. (2007). Microarray analysis of iron deficiency chlorosis in near-isogenic soybean lines. Bmc Genomics, 8(1), 476.

Oya, T., Nepomuceno, A. L., Neumaier, N., Farias, J. R. B., Tobita, S., & Ito, O. (2004). Drought tolerance characteristics of Brazilian soybean cultivars. *Plant Production Science*, *7*(2), 129-137.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. Nature reviews genetics, 12(2), 87.

Palusa, S. G., Ali, G. S., & Reddy, A. S. (2007). Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *The Plant Journal*, *49*(6), 1091-1107.

Palusa, S. G., Ali, G. S., & Reddy, A. S. (2007). Alternative splicing of pre-mRNAs of Arabidopsis serine/arginine-rich proteins: regulation by hormones and stresses. *The Plant Journal*, *49*(6), 1091-1107.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative *splicing* complexity in the human transcriptome by high-throughput sequencing. Nature genetics, 40(12), 1413.

Pickrell, J. K., Pai, A. A., Gilad, Y., & Pritchard, J. K. (2010). Noisy *splicing* drives mRNA isoform diversity in human cells. PLoS genetics, 6(12), e1001236.

Pinheiro, G. L., Marques, C. S., Costa, M. D., Reis, P. A., Alves, M. S., Carvalho, C. M., ... & Fontes, E. P. (2009). Complete inventory of soybean NAC transcription factors: sequence conservation and expression analysis uncover their distinct roles in stress response. Gene, 444(1-2), 10-23.

Pitol, C., & Broch, D. L. (2008). Soja mais produtiva e tolerante à seca. Gráfica MARACAJU. Tecnologia de produção: soja e milho, 2009(5).

Pohl, M., Bortfeldt, R. H., Grützmann, K., & Schuster, S. (2013). Alternative splicing of mutually exclusive exons—a review. *Biosystems*, *114*(1), 31-38.

Radford, A. D., Chapman, D., Dixon, L., Chantrey, J., Darby, A. C., & Hall, N. (2012). Application of next-generation sequencing technologies in virology. The Journal of general virology, 93(Pt 9), 1853.

Ranjan, A., & Sawant, S. (2015). Genome-wide transcriptomic comparison of cotton (Gossypium herbaceum) leaf and root under drought stress. 3 Biotech, 5(4), 585-596.

Reddy, A. R., Chaitanya, K. V., & Vivekanandan, M. (2004). Drought-induced responses of photosynthesis and antioxidant metabolism in higher plants. Journal of plant physiology, 161(11), 1189-1202.

Rodrigues Gomes, G. D., Benin, G., Colvara Rosinha, R., Galvan, D., Stefani Pagliosa, E., Pinnow, C., ... & Beche, E. (2012). Produção e qualidade fisiológica de sementes de soja em diferentes ambientes de cultivo. Semina: Ciências Agrárias, 33(1).

Ruskin, B., Krainer, A. R., Maniatis, T., & Green, M. R. (1984). Excision of an intact intron as a novel lariat structure during pre-mRNA *splicing* in vitro. Cell, 38(1), 317-331.

Sahoo, K. K., Tripathi, A. K., Pareek, A., & Singla-Pareek, S. L. (2013). Taming drought stress in rice through genetic engineering of transcription factors and protein kinases. Plant Stress, 7(1), 60-72.

Saltzman, A. L., Pan, Q., & Blencowe, B. J. (2011). Regulation of alternative splicing by the core spliceosomal machinery. *Genes & development*, *25*(4), 373-384..

Schiermeier, Q. (2006). Climate change: A sea change.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature 463, 178–183. doi: 10.1038/nature08670

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., ... & Xu, D. (2010). Genome sequence of the palaeopolyploid soybean. nature, 463(7278), 178.

Scholander, P.F., Hammel, H.J., Bradstreet, A. and Hemmingsen, E.A., 1965. Sap pressure in vascular plants. Science 148 339-346.

Seki, M., Kamei, A., Yamaguchi-Shinozaki, K., & Shinozaki, K. (2003). Molecular responses to drought, salinity and frost: common and different paths for plant protection. Current opinion in biotechnology, 14(2), 194-199.

Serraj R., Sinclair T.R.: Osmolyte accumulation: can it really help increase crop yield under drought conditions? — Plant Cell Environ. 25: 333–341, 2002.

Shao, H. B., Chu, L. Y., Jaleel, C. A., & Zhao, C. X. (2008). Water-deficit stress-induced anatomical changes in higher plants. Comptes rendus biologies, 331(3), 215-225.

Shao, H. B., Chu, L. Y., Jaleel, C. A., & Zhao, C. X. (2008). Water-deficit stress-induced anatomical changes in higher plants. Comptes rendus biologies, 331(3), 215-225.

Shen, S., Park, J. W., Lu, Z. X., Lin, L., Henry, M. D., Wu, Y. N., ... & Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative *splicing* from replicate RNA-Seq data. Proceedings of the National Academy of Sciences, 111(51), E5593-E5601.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nature biotechnology, 26(10), 1135.

Shinozaki K, Yamaguchi- Shinozaki Y (2000) Molecular responses to dehydration and low temperature: difference and cross talk between two stress signaling pathways. Curr Opin Plant Biol 3:217–223.

Shinozaki, K., & Yamaguchi-Shinozaki, K. (1999). Molecular responses to drought stress.

Shinozaki, K., & Yamaguchi-Shinozaki, K. (2000). Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. Current opinion in plant biology, 3(3), 217-223.

Simon, S. A., Zhai, J., Nandety, R. S., McCormick, K. P., Zeng, J., Mejia, D., & Meyers, B. C. (2009). Short-read sequencing technologies for transcriptional analyses. Annual review of plant biology, 60, 305-333.

Spinoni, J., Naumann, G., & Vogt, J. V. (2017). Pan-European seasonal trends and recent changes of drought frequency and severity. Global and Planetary Change, 148, 113-130.

Staiger, D., & Brown, J. W. (2013). Alternative splicing at the intersection of biological timing, development, and stress responses. *The Plant Cell*, *25*(10), 3640-3656

Stokstad, E. (2004). Pollution gets personal.

Sugliani, M., Brambilla, V., Clerkx, E. J., Koornneef, M., & Soppe, W. J. (2010). The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in Arabidopsis. *The Plant Cell*, *22*(6), 1936-1946.
Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., ... & Schmidt, D. (2008). A global view of gene activity and alternative *splicing* by deep sequencing of the human transcriptome. Science, 321(5891), 956-960.

Tanabe, N., Yoshimura, K., Kimura, A., Yabuta, Y., & Shigeoka, S. (2007). Differential expression of alternatively spliced mRNAs of Arabidopsis SR protein homologs, atSR30 and atSR45a, in response to environmental stress. *Plant and Cell Physiology*, *48*(7), 1036-1049.

Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O., & Gilbert, W. (1978). Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain. *Proceedings of the National Academy of Sciences*, *75*(3), 1485-1489.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols, 7(3), 562.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols, 7(3), 562.

Tsukahara, R. Y., de Batista Fonseca, I. C., de Aguiar, M. A., Kochinski, E. G., Neto, J. P., & Suyama, J. T. (2016). Produtividade de soja em consequência do atraso da colheita e de condições ambientais. Pesquisa Agropecuária Brasileira, 51(8), 905-915.

Tuberosa, R., & Salvi, S. (2006). Genomics-based approaches to improve drought tolerance of crops. Trends in plant science, 11(8), 405-412.

TURNER, N.C. Further progress in crop water relations. Adv. Agron., San Diego, v. 58, p. 293-338, 1997.

Umezawa, T., Fujita, M., Fujita, Y., Yamaguchi-Shinozaki, K., & Shinozaki, K. (2006). Engineering drought tolerance in plants: discovering and tailoring genes to unlock the future. Current opinion in biotechnology, 17(2), 113-122.

Vargas, L., Silva, A. D., Borém, A., Rezende, S. D., Ferreira, F. A., & Sediyama, T. (1999). Resistência de plantas daninhas a herbicidas. Viçosa, MG: Universidade Federal de Viçosa.

Varshney, R. K., Nayak, S. N., May, G. D., & Jackson, S. A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends in biotechnology, 27(9), 522-530.

Vogt, J. V., Niemeyer, S., Somma, F., Beaudin, I., & Viau, A. A. (2000). Drought monitoring from space. In Drought and drought mitigation in Europe (pp. 167-183). Springer, Dordrecht.

Wang, D., Zavadil, J., Martin, L., Parisi, F., Friedman, E., Levy, D., ... & Gardner, L. B. (2011). Inhibition of nonsense-mediated RNA decay by the tumor microenvironment promotes tumorigenesis. Molecular and cellular biology, 31(17), 3670-3680.

Wang, J., Zhang, J., Li, K., Zhao, W., & Cui, Q. (2011). SpliceDisease database: linking RNA splicing and disease. *Nucleic acids research*, *40*(D1), D1055-D1059.

Wang, L., Feng, Z., Wang, X., Wang, X., & Zhang, X. (2009). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics, 26(1), 136-138.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 10(1), 57.

Chiu, Y. T., Wong, J. K., Choi, S. W., Sze, K. M., Ho, D. W., Chan, L. K., ... & Wong, C. M. (2016). Novel pre-mRNA splicing of intronically integrated HBV generates oncogenic chimera in hepatocellular carcinoma. *Journal of hepatology*, *64*(6), 1256-1264.

Young, M. D., Wakefield, M. J., Smyth, G. K., & Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome biology, 11(2), R14.

MATERIAL SUPLEMENTAR

Figura S2: O fluxo de trabalho de análise de RNA-Seq que foi utilizado para a obtenção da expressão diferencial de genes. TopHat/Cufflinks.

DAS EMB48	Ortholog_ensg	Description
GLYMA_07G051300	AT1G01160	GRF1-interacting factor 2 [Source:UniProtKB/TrEMBL

GLYMA_03G258600	AT1G01240	At1g01240/F6F3_11 [Source:UniProtKB/TrEMBL
GLYMA_07G073500	AT1G01770	Propionyl-CoA carboxylase [Source:UniProtKB/TrEMBL
GLYMA_07G073500	AT1G01770	Propionyl-CoA carboxylase [Source:UniProtKB/TrEMBL
GLYMA_07G073500	AT1G01770	Propionyl-CoA carboxylase [Source:UniProtKB/TrEMBL
GLYMA_18G252100	AT1G02090	COP9 signalosome complex subunit 7 [Source:UniProtKB/Swiss-Prot
GLYMA_19G209800	AT1G02910	Protein LOW PSII ACCUMULATION 1, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_20G076700	AT1G04790	At1g04790 [Source:UniProtKB/TrEMBL
GLYMA_08G157400	AT1G05570	Callose synthase 1 [Source:UniProtKB/Swiss-Prot
GLYMA_20G248000	AT1G05790	lipase class 3 family protein [Source:TAIR
GLYMA_17G254300	AT1G06060	At1g06060 [Source:UniProtKB/TrEMBL
GLYMA_13G328300	AT1G07960	PDIL5-1 [Source:UniProtKB/TrEMBL
GLYMA_12G083600	AT1G08210	Eukaryotic aspartyl protease family protein [Source:TAIR
GLYMA_06G096400	AT1G11475	NRPE10 [Source:UniProtKB/TrEMBL
GLYMA_08G297200	AT1G14410	Single-stranded DNA-binding protein WHY1, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_08G207400	AT1G15280	CASC3/Barentsz eIF4AIII binding protein [Source:UniProtKB/TrEMBL

GLYMA_15G000900	AT1G16800	P-loop containing nucleoside triphosphate hydrolases superfamily protein [Source:TAIR
GLYMA_13G259200	AT1G17140	RIP1 [Source:UniProtKB/TrEMBL
GLYMA_07G036200	AT1G17590	Uncharacterized protein At1g17590 (Fragment) [Source:UniProtKB/TrEMBL
GLYMA_05G019700	AT1G18610	F25I16.5 protein [Source:UniProtKB/TrEMBL
GLYMA_06G204400	AT1G18620	unknown protein
GLYMA_17G080000	AT1G18620	unknown protein
GLYMA_04G029900	AT1G19650	Phosphatidylinositol/phosphatidylcholine transfer protein SFH4 [Source:UniProtKB/Swiss-Prot
GLYMA_06G030100	AT1G22100	Inositol-pentakisphosphate 2-kinase [Source:UniProtKB/TrEMBL
GLYMA_09G095200	AT1G22690	Gibberellin-regulated protein 9 [Source:UniProtKB/Swiss-Prot
GLYMA_02G033300	AT1G23900	AP-1 complex subunit gamma-1 [Source:UniProtKB/Swiss-Prot
GLYMA_02G033300	AT1G23935	CONTAINS InterPro DOMAIN/s: Apoptosis inhibitory 5 (InterPro:IPR008383)
GLYMA_01G044900	AT1G26940	Peptidyl-prolyl cis-trans isomerase CYP23 [Source:UniProtKB/Swiss-Prot
GLYMA_08G293600	AT1G27170	Transmembrane receptors / ATP binding protein [Source:UniProtKB/TrEMBL
GLYMA_08G293600	AT1G27180	Disease resistance protein (TIR-NBS-LRR class) [Source:UniProtKB/TrEMBL
GLYMA_08G105000	AT1G27320	HK3 [Source:UniProtKB/TrEMBL

GLYMA_15G130500	AT1G30520	2-succinylbenzoate--CoA ligase, chloroplastic/peroxisomal [Source:UniProtKB/Swiss-Prot
GLYMA_01G014400	AT1G32220	Uncharacterized protein At1g32220, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_16G156300	AT1G34380	5'-3' exonuclease family protein [Source:UniProtKB/TrEMBL
GLYMA_08G183100	AT1G43640	Tubby-like F-box protein 5 [Source:UniProtKB/Swiss-Prot
GLYMA_14G150600	AT1G44446	CH1 [Source:UniProtKB/TrEMBL
GLYMA_09G281700	AT1G45100	RNA-binding (RRM/RBD/RNP motifs) family protein [Source:UniProtKB/TrEMBL
GLYMA_20G004000	AT1G45100	RNA-binding (RRM/RBD/RNP motifs) family protein [Source:UniProtKB/TrEMBL
GLYMA_05G004900	AT1G50170	Sirohydrochlorin ferrochelatase, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_18G101400	AT1G50890	TORTIFOLIA1-like protein 1 [Source:UniProtKB/Swiss-Prot
GLYMA_15G032700	AT1G51140	Transcription factor bHLH122 [Source:UniProtKB/Swiss-Prot
GLYMA_08G221800	AT1G51600	GATA transcription factor 28 [Source:UniProtKB/Swiss-Prot
GLYMA_08G211300	AT1G52630	O-fucosyltransferase 13 [Source:UniProtKB/Swiss-Prot
GLYMA_U025600	AT1G53165	Protein kinase superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_06G238000	AT1G53165	Protein kinase superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_12G234300	AT1G53500	RHM2 [Source:UniProtKB/TrEMBL
GLYMA_07G036200	AT1G54160	Nuclear transcription factor Y subunit A-5 [Source:UniProtKB/Swiss-Prot

GLYMA_09G066800	AT1G54217	Ribosomal protein L18ae family [Source:UniProtKB/TrEMBL
GLYMA_08G160500	AT1G54340	Peroxisomal isocitrate dehydrogenase [NADP] [Source:UniProtKB/Swiss-Prot
GLYMA_05G219500	AT1G55090	Glutamine-dependent NAD(+) synthetase [Source:UniProtKB/Swiss-Prot
GLYMA_05G219500	AT1G55090	Glutamine-dependent NAD(+) synthetase [Source:UniProtKB/Swiss-Prot
GLYMA_14G130400	AT1G55300	Transcription initiation factor TFIID subunit 7 [Source:UniProtKB/Swiss-Prot
GLYMA_06G030100	AT1G58643	Inositol-pentakisphosphate 2-kinase [Source:UniProtKB/TrEMBL
GLYMA_06G030100	AT1G58936	Inositol-pentakisphosphate 2-kinase [Source:UniProtKB/TrEMBL
GLYMA_06G030100	AT1G59312	Inositol-pentakisphosphate 2-kinase [Source:UniProtKB/TrEMBL
GLYMA_04G047500	AT1G59520	CW7 [Source:TAIR
GLYMA_02G149800	AT1G59560	E3 ubiquitin-protein ligase SPL1 [Source:UniProtKB/Swiss-Prot
GLYMA_02G033300	AT1G60070	AP-1 complex subunit gamma [Source:UniProtKB/TrEMBL
GLYMA_04G255700	AT1G60830	RNA-binding (RRM/RBD/RNP motifs) family protein [Source:UniProtKB/TrEMBL
GLYMA_04G255700	AT1G60900	Splicing factor U2af large subunit B [Source:UniProtKB/Swiss-Prot
GLYMA_13G219700	AT1G61280	Phosphatidylinositol N-acetylglucosaminyltransferase subunit P [Source:UniProtKB/Swiss-Prot
GLYMA_06G096400	AT1G61700	DNA-directed RNA polymerase subunit 10-like protein [Source:UniProtKB/Swiss-Prot
GLYMA_19G125600	AT1G63210	Transcription elongation factor SPT6-like [Source:UniProtKB/Swiss-Prot

GLYMA_02G149800	AT1G63900	E3 ubiquitin-protein ligase SP1 [Source:UniProtKB/Swiss-Prot
GLYMA_02G088200	AT1G64680	At1g64680 [Source:UniProtKB/TrEMBL
GLYMA_19G110700	AT1G64770	NDH-dependent cyclic electron flow 1 [Source:UniProtKB/TrEMBL
GLYMA_19G125600	AT1G65440	Transcription elongation factor SPT6 homolog [Source:UniProtKB/Swiss-Prot
GLYMA_08G202700	AT1G66590	Cytochrome c oxidase 19-1 [Source:UniProtKB/TrEMBL
GLYMA_20G119100	AT1G68410	Probable protein phosphatase 2C 15 [Source:UniProtKB/Swiss-Prot
GLYMA_10G204900	AT1G68820	Transmembrane Fragile-X-F-associated protein [Source:TAIR
GLYMA_08G202700	AT1G69750	At1g66590 [Source:UniProtKB/TrEMBL
GLYMA_01G028100	AT1G70520	Cysteine-rich receptor-like protein kinase 2 [Source:UniProtKB/Swiss-Prot
GLYMA_07G036200	AT1G72830	Nuclear factor Y, subunit A3 [Source:UniProtKB/TrEMBL
GLYMA_17G122500	AT1G74070	Peptidyl-prolyl cis-trans isomerase CYP26-2, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_05G019700	AT1G74150	Galactose oxidase/kelch repeat superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_06G204400	AT1G74160	LONGIFOLIA protein [Source:UniProtKB/TrEMBL
GLYMA_17G080000	AT1G74160	LONGIFOLIA protein [Source:UniProtKB/TrEMBL
GLYMA_17G247300	AT1G74970	30S ribosomal protein S9, chloroplastic [Source:UniProtKB/Swiss-Prot

GLYMA_04G029900	AT1G75370	Sec14p-like phosphatidylinositol transfer family protein [Source:UniProtKB/TrEMBL
GLYMA_17G259200	AT1G75850	Vacuolar protein sorting-associated protein 35 [Source:UniProtKB/TrEMBL
GLYMA_04G020900	AT1G76050	RNA pseudouridine synthase 2, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_13G259200	AT1G78430	RIP4 [Source:UniProtKB/TrEMBL
GLYMA_12G234300	AT1G78570	Trifunctional UDP-glucose 4,6-dehydratase/UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase/UDP-4-keto-L-rhamnose-reductase RHM1 [Source:UniProtKB/Swiss-Prot
GLYMA_09G252100	AT1G79790	Haloacid dehalogenase-like hydrolase (HAD) superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_08G207400	AT1G80000	Protein MLN51 homolog [Source:UniProtKB/Swiss-Prot
GLYMA_08G297200	AT2G02740	Single-stranded DNA-binding protein WHY3, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_18G120100	AT2G02910	At2g02910 [Source:UniProtKB/TrEMBL
GLYMA_12G188300	AT2G03070	Mediator of RNA polymerase II transcription subunit 8 [Source:UniProtKB/Swiss-Prot
GLYMA_04G080200	AT2G03120	Signal peptide peptidase [Source:UniProtKB/Swiss-Prot
GLYMA_07G066700	AT2G03690	Ubiquinone biosynthesis protein COQ4 homolog, mitochondrial [Source:UniProtKB/Swiss-Prot
GLYMA_13G110300	AT2G06010	OBP3-responsive protein 4 (ORG4) [Source:UniProtKB/TrEMBL

GLYMA_15G058900	AT2G14170	Methylmalonate-semialdehyde dehydrogenase [acylating], mitochondrial [Source:UniProtKB/Swiss-Prot
GLYMA_20G130500	AT2G15000	unknown protein
GLYMA_20G130500	AT2G15000	unknown protein
GLYMA_17G052200	AT2G19385	At2g19385 [Source:UniProtKB/TrEMBL
GLYMA_07G126600	AT2G19800	At2g19800 [Source:UniProtKB/TrEMBL
GLYMA_13G294600	AT2G20585	Nuclear fusion defective 6 [Source:UniProtKB/TrEMBL
GLYMA_11G116600	AT2G21590	Probable glucose-1-phosphate adenylyltransferase large subunit, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_05G186200	AT2G21800	essential meiotic endonuclease 1A [Source:TAIR
GLYMA_05G186200	AT2G22140	Crossover junction endonuclease EME1B [Source:UniProtKB/Swiss-Prot
GLYMA_02G234300	AT2G23140	RING-type E3 ubiquitin transferase [Source:UniProtKB/TrEMBL
GLYMA_19G110300	AT2G23840	At2g23840 [Source:UniProtKB/TrEMBL
GLYMA_06G060300	AT2G24860	At2g24860/F27C12.22 [Source:UniProtKB/TrEMBL
GLYMA_14G116600	AT2G25480	TPX2 (targeting protein for Xklp2) protein family [Source:TAIR
GLYMA_06G070400	AT2G25590	Plant Tudor-like protein [Source:UniProtKB/TrEMBL
GLYMA_06G088500	AT2G25850	PAPS2 [Source:UniProtKB/TrEMBL

GLYMA_08G08130 0	AT2G27340	At2g27340 [Source:UniProtKB/TrEMBL
GLYMA_13G32130 0	AT2G27950	Ring/U-Box superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_11G19340 0	AT2G28130	Actin protein 2/3 complex subunit-like protein [Source:UniProtKB/TrEMBL
GLYMA_10G05390 0	AT2G28260	Putative cyclic nucleotide-gated ion channel 15 [Source:UniProtKB/Swiss-Prot
GLYMA_08G15740 0	AT2G31960	Callose synthase 2 [Source:UniProtKB/Swiss-Prot
GLYMA_12G09690 0	AT2G32530	Cellulose synthase-like protein B3 [Source:UniProtKB/Swiss-Prot
GLYMA_12G09690 0	AT2G32540	Glycosyltransferase (Fragment) [Source:UniProtKB/TrEMBL
GLYMA_12G09690 0	AT2G32610	Cellulose synthase-like protein B1 [Source:UniProtKB/Swiss-Prot
GLYMA_12G09690 0	AT2G32620	Cellulose synthase-like protein B2 [Source:UniProtKB/Swiss-Prot
GLYMA_04G01440 0	AT2G33610	SWI/SNF complex subunit SWI3B [Source:UniProtKB/Swiss-Prot
GLYMA_05G06770 0	AT2G35190	Novel plant SNARE 11 [Source:UniProtKB/Swiss-Prot
GLYMA_10G07120 0	AT2G36305	CAAX prenyl protease 2 [Source:UniProtKB/Swiss-Prot
GLYMA_13G14460 0	AT2G37025	TRF-like 8 [Source:UniProtKB/TrEMBL
GLYMA_19G18000 0	AT2G37150	RING/U-box superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_05G19300 0	AT2G37720	Protein trichome birefringence-like 15 [Source:UniProtKB/Swiss-Prot
GLYMA_02G08740 0	AT2G39250	AP2-like ethylene-responsive transcription factor SNZ [Source:UniProtKB/Swiss-Prot

GLYMA_13G219700	AT2G39445	Phosphatidylinositol N-acetylglucosaminyltransferase, GPI19/PIG-P subunit [Source:UniProtKB/TrEMBL
GLYMA_18G024100	AT2G40316	Autophagy-like protein [Source:UniProtKB/TrEMBL
GLYMA_04G222600	AT2G40640	U-box domain-containing protein 63 [Source:UniProtKB/Swiss-Prot
GLYMA_12G120100	AT2G42400	VOZ2 [Source:UniProtKB/TrEMBL
GLYMA_10G296000	AT2G43240	CMP-sialic acid transporter 2 [Source:UniProtKB/Swiss-Prot
GLYMA_18G227200	AT2G45740	Peroxisomal membrane protein 11D [Source:UniProtKB/Swiss-Prot
GLYMA_16G023700	AT2G46520	Exportin-2 [Source:UniProtKB/Swiss-Prot
GLYMA_03G258600	AT2G46550	Expressed protein [Source:UniProtKB/TrEMBL
GLYMA_19G261500	AT2G46880	Probable inactive purple acid phosphatase 14 [Source:UniProtKB/Swiss-Prot
GLYMA_11G012300	AT3G03140	T17B22.17 protein [Source:UniProtKB/TrEMBL
GLYMA_16G151100	AT3G03500	T21P5.8 protein [Source:UniProtKB/TrEMBL
GLYMA_09G043900	AT3G06880	Transducin/WD40 repeat-like superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_04G002700	AT3G07580	At3g07580 [Source:UniProtKB/TrEMBL
GLYMA_08G304100	AT3G10260	Reticulon family protein [Source:TAIR
GLYMA_10G104300	AT3G10410	Carboxypeptidase [Source:UniProtKB/TrEMBL
GLYMA_08G161000	AT3G12670	Emb2742 [Source:UniProtKB/TrEMBL

GLYMA_11G015300	AT3G12990	Exosome complex component RRP45A [Source:UniProtKB/Swiss-Prot
GLYMA_03G027700	AT3G13440	AT3g13440/MRP15_7 [Source:UniProtKB/TrEMBL
GLYMA_16G006000	AT3G13990	AT3g13990/MDC16_11 [Source:UniProtKB/TrEMBL
GLYMA_07G036200	AT3G14020	Nuclear transcription factor Y subunit A-6 [Source:UniProtKB/Swiss-Prot
GLYMA_07G043100	AT3G14200	Chaperone DnaJ-domain superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_12G234300	AT3G14790	Rhamnose biosynthesis 3 [Source:UniProtKB/TrEMBL
GLYMA_13G271800	AT3G15050	IQ-domain 10 [Source:UniProtKB/TrEMBL
GLYMA_U025600	AT3G15220	MAP kinase [Source:UniProtKB/TrEMBL
GLYMA_06G238000	AT3G15220	MAP kinase [Source:UniProtKB/TrEMBL
GLYMA_13G359600	AT3G16010	Pentatricopeptide repeat-containing protein At3g16010 [Source:UniProtKB/Swiss-Prot
GLYMA_07G244700	AT3G16800	Probable protein phosphatase 2C 41 [Source:UniProtKB/Swiss-Prot
GLYMA_15G187500	AT3G17450	HAT dimerization domain-containing protein [Source:UniProtKB/TrEMBL
GLYMA_09G168500	AT3G17690	Putative cyclic nucleotide-gated ion channel 19 [Source:UniProtKB/Swiss-Prot
GLYMA_09G168500	AT3G17700	Probable cyclic nucleotide-gated ion channel 20, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_02G021900	AT3G17900	AT3g17900/MEB5_12 [Source:UniProtKB/TrEMBL
GLYMA_08G296300	AT3G17900	AT3g17900/MEB5_12 [Source:UniProtKB/TrEMBL

GLYMA_09G115100	AT3G18290	Zinc finger protein BRUTUS [Source:UniProtKB/Swiss-Prot
GLYMA_01G008800	AT3G18760	Translation elongation factor EF1B/ribosomal protein S6 family protein [Source:UniProtKB/TrEMBL
GLYMA_12G044700	AT3G18830	PMT5 [Source:UniProtKB/TrEMBL
GLYMA_02G133400	AT3G19190	Autophagy-related protein 2 [Source:UniProtKB/Swiss-Prot
GLYMA_13G074600	AT3G20320	Protein TRIGALACTOSYLDIACYLGLYCEROL 2, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_08G221800	AT3G21175	GATA transcription factor 24 [Source:UniProtKB/Swiss-Prot
GLYMA_10G004800	AT3G23430	Phosphate transporter PHO1 [Source:UniProtKB/Swiss-Prot
GLYMA_12G083600	AT3G42550	Eukaryotic aspartyl protease family protein [Source:UniProtKB/TrEMBL
GLYMA_09G151400	AT3G45860	Cysteine-rich receptor-like protein kinase 4 [Source:UniProtKB/Swiss-Prot
GLYMA_04G180100	AT3G48050	Protein SUO [Source:UniProtKB/TrEMBL
GLYMA_04G180100	AT3G48060	BAH and TFIIS domain-containing protein [Source:UniProtKB/TrEMBL
GLYMA_09G281700	AT3G48835	Polynucleotide adenylyltransferase domain/RNA recognition motif protein [Source:UniProtKB/TrEMBL
GLYMA_20G004000	AT3G48835	Polynucleotide adenylyltransferase domain/RNA recognition motif protein [Source:UniProtKB/TrEMBL
GLYMA_06G068900	AT3G49500	RNA-dependent RNA polymerase 6 [Source:UniProtKB/Swiss-Prot

GLYMA_04G074300	AT3G49600	Ubiquitin carboxyl-terminal hydrolase 26 [Source:UniProtKB/Swiss-Prot
GLYMA_05G105700	AT3G49880	Glycosyl hydrolase family protein 43 [Source:UniProtKB/TrEMBL
GLYMA_16G144200	AT3G50380	Vacuolar protein sorting-associated protein, putative (DUF1162) [Source:UniProtKB/TrEMBL
GLYMA_16G144200	AT3G50380	Vacuolar protein sorting-associated protein, putative (DUF1162) [Source:UniProtKB/TrEMBL
GLYMA_11G089000	AT3G51500	At3g51500 [Source:UniProtKB/TrEMBL
GLYMA_16G115700	AT3G51520	Diacylglycerol O-acyltransferase 2 [Source:UniProtKB/Swiss-Prot
GLYMA_01G116000	AT3G51880	high mobility group B1 [Source:TAIR
GLYMA_09G062900	AT3G52840	Beta-galactosidase [Source:UniProtKB/TrEMBL
GLYMA_05G040200	AT3G54170	FIP37 [Source:UniProtKB/TrEMBL
GLYMA_16G102300	AT3G54230	suppressor of abi3-5 [Source:TAIR
GLYMA_02G087400	AT3G54990	AP2-like ethylene-responsive transcription factor SMZ [Source:UniProtKB/Swiss-Prot
GLYMA_08G081300	AT3G58130	At3g58130 [Source:UniProtKB/TrEMBL
GLYMA_10G296000	AT3G59360	CMP-sialic acid transporter 3 [Source:UniProtKB/Swiss-Prot
GLYMA_12G052600	AT4G15290	Glycosyltransferase family protein [Source:UniProtKB/TrEMBL
GLYMA_01G074500	AT3G60415	At3g60420 [Source:UniProtKB/TrEMBL

GLYMA_01G074500	AT3G60420	Phosphoglycerate mutase family protein [Source:UniProtKB/TrEMBL
GLYMA_01G074500	AT3G60440	Phosphoglycerate mutase family protein [Source:UniProtKB/TrEMBL
GLYMA_01G074500	AT3G60450	Phosphoglycerate mutase family protein [Source:UniProtKB/TrEMBL
GLYMA_11G015300	AT3G60500	Exosome complex component RRP45B [Source:UniProtKB/Swiss-Prot
GLYMA_13G278200	AT3G60710	Putative FBD-associated F-box protein At3g60710 [Source:UniProtKB/Swiss-Prot
GLYMA_18G220400	AT3G60800	S-acyltransferase [Source:UniProtKB/TrEMBL
GLYMA_18G227200	AT3G61070	PEX11E [Source:UniProtKB/TrEMBL
GLYMA_19G222100	AT3G62080	SNF7 family protein [Source:UniProtKB/TrEMBL
GLYMA_07G051300	AT4G00850	GRF1-interacting factor 3 [Source:UniProtKB/Swiss-Prot
GLYMA_07G170000	AT4G01880	Methyltransferase [Source:UniProtKB/TrEMBL
GLYMA_15G042000	AT4G03090	Nodulin homeobox [Source:UniProtKB/Swiss-Prot
GLYMA_09G151400	AT4G05200	cysteine-rich RLK (RECEPTOR-like protein kinase) 25 [Source:TAIR
GLYMA_03G085700	AT4G13020	Protein kinase superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_08G043700	AT4G13070	At4g13070 [Source:UniProtKB/TrEMBL
GLYMA_13G202400	AT4G15030	Folate-sensitive fragile site protein [Source:UniProtKB/TrEMBL
GLYMA_10G259000	AT4G15090	Protein FAR-RED IMPAIRED RESPONSE 1 [Source:UniProtKB/Swiss-Prot

GLYMA_20G132100	AT4G15090	Protein FAR-RED IMPAIRED RESPONSE 1 [Source:UniProtKB/Swiss-Prot
GLYMA_12G096900	AT4G15290	Glycosyltransferase (Fragment) [Source:UniProtKB/TrEMBL
GLYMA_12G096900	AT4G15320	cellulose synthase-like B6 [Source:TAIR
GLYMA_05G060500	AT4G17360	Formyltetrahydrofolate deformylase 2, mitochondrial [Source:UniProtKB/Swiss-Prot
GLYMA_06G034800	AT4G18230	Glycosyltransferase [Source:UniProtKB/TrEMBL
GLYMA_09G029700	AT4G20760	NAD(P)-binding Rossmann-fold superfamily protein [Source:TAIR
GLYMA_08G129600	AT4G21120	CAT1 [Source:UniProtKB/TrEMBL
GLYMA_13G247400	AT4G21430	Protein B160 [Source:UniProtKB/TrEMBL
GLYMA_09G151400	AT4G23130	cysteine-rich RLK (RECEPTOR-like protein kinase) 5 [Source:TAIR
GLYMA_09G151400	AT4G23140	cysteine-rich RLK (RECEPTOR-like protein kinase) 6 [Source:TAIR
GLYMA_09G151400	AT4G23150	Cysteine-rich receptor-like protein kinase 7 [Source:UniProtKB/Swiss-Prot
GLYMA_09G151400	AT4G23180	Cysteine-rich receptor-like protein kinase 10 [Source:UniProtKB/Swiss-Prot
GLYMA_09G151400	AT4G23230	Cysteine-rich receptor-like protein kinase 15 [Source:UniProtKB/Swiss-Prot
GLYMA_09G151400	AT4G23270	cysteine-rich RLK (RECEPTOR-like protein kinase) 19 [Source:TAIR
GLYMA_09G151400	AT4G23280	Putative cysteine-rich receptor-like protein kinase 20 [Source:UniProtKB/Swiss-Prot
GLYMA_09G151400	AT4G23310	Putative cysteine-rich receptor-like protein kinase 23 [Source:UniProtKB/Swiss-Prot

GLYMA_02G290800	AT4G25130	Peptide methionine sulfoxide reductase A4, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_08G114400	AT4G25650	Protochlorophyllide-dependent translocon component 52, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_09G066800	AT4G26060	At4g26060 [Source:UniProtKB/TrEMBL
GLYMA_09G062900	AT4G26140	beta-galactosidase 12 [Source:TAIR
GLYMA_07G126600	AT4G26260	MIOX4 [Source:UniProtKB/TrEMBL
GLYMA_04G122600	AT4G26310	Elongation factor P (EF-P) family protein [Source:UniProtKB/TrEMBL
GLYMA_18G101400	AT4G27060	Microtubule-associated protein TORTIFOLIA1 [Source:UniProtKB/Swiss-Prot
GLYMA_09G193600	AT4G28860	Casein kinase 1-like protein 4 [Source:UniProtKB/Swiss-Prot
GLYMA_09G193600	AT4G28860	Casein kinase 1-like protein 4 [Source:UniProtKB/Swiss-Prot
GLYMA_09G193600	AT4G28880	Ckl3 [Source:UniProtKB/TrEMBL
GLYMA_09G193600	AT4G28880	Ckl3 [Source:UniProtKB/TrEMBL
GLYMA_17G057700	AT4G29890	Choline monooxygenase, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_17G057800	AT4G29900	Calcium-transporting ATPase 10, plasma membrane-type [Source:UniProtKB/Swiss-Prot
GLYMA_07G196700	AT4G30930	NFD1 [Source:UniProtKB/TrEMBL

GLYMA_15G205500	AT4G30993	Calcineurin-like metallo-phosphoesterase superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_06G066500	AT4G32140	AT4g32140/F10N7_50 [Source:UniProtKB/TrEMBL
GLYMA_14G116600	AT4G32330	Protein WVD2-like 5 [Source:UniProtKB/Swiss-Prot
GLYMA_06G070400	AT4G32440	Plant Tudor-like RNA-binding protein [Source:UniProtKB/TrEMBL
GLYMA_06G088500	AT4G32850	nuclear poly(a) polymerase [Source:TAIR
GLYMA_20G130500	AT4G34265	unknown protein
GLYMA_20G130500	AT4G34265	unknown protein
GLYMA_11G121200	AT4G34360	At4g34360 [Source:UniProtKB/TrEMBL
GLYMA_04G255700	AT4G36690	U2 snRNP auxiliary factor large subunit [Source:UniProtKB/TrEMBL
GLYMA_01G019900	AT4G37920	Uncharacterized protein At4g37920 [Source:UniProtKB/Swiss-Prot
GLYMA_05G156400	AT4G38500	AT4g38500/F20M13_60 [Source:UniProtKB/TrEMBL
GLYMA_12G024600	AT4G38760	Protein of unknown function (DUF3414) [Source:TAIR
GLYMA_11G116600	AT4G39210	Glucose-1-phosphate adenylyltransferase [Source:UniProtKB/TrEMBL
GLYMA_08G134300	AT4G39970	Haloacid dehalogenase-like hydrolase domain-containing protein At4g39970 [Source:UniProtKB/Swiss-Prot
GLYMA_04G222600	AT5G05230	U-box domain-containing protein 62 [Source:UniProtKB/Swiss-Prot

GLYMA_02G290800	AT5G07460	PMSR2 [Source:UniProtKB/TrEMBL
GLYMA_02G290800	AT5G07470	PMSR3 [Source:UniProtKB/TrEMBL
GLYMA_05G023200	AT5G07830	Heparanase-like protein 1 [Source:UniProtKB/Swiss-Prot
GLYMA_06G209000	AT5G08190	NF-YB12 [Source:UniProtKB/TrEMBL
GLYMA_06G209000	AT5G08190	NF-YB12 [Source:UniProtKB/TrEMBL
GLYMA_09G160400	AT5G08370	Alpha-galactosidase 2 [Source:UniProtKB/Swiss-Prot
GLYMA_20G214600	AT5G08450	Zinc finger CCCH domain protein [Source:UniProtKB/TrEMBL
GLYMA_05G195900	AT5G08720	At5g08720 [Source:UniProtKB/TrEMBL
GLYMA_07G181500	AT5G09540	At5g09540 [Source:UniProtKB/TrEMBL
GLYMA_04G216400	AT5G09840	Emb [Source:UniProtKB/TrEMBL
GLYMA_08G145400	AT5G09880	Splicing factor, CC1-like protein [Source:UniProtKB/TrEMBL
GLYMA_08G280600	AT5G10440	Cyclin-D4-2 [Source:UniProtKB/Swiss-Prot
GLYMA_04G084800	AT5G11630	Uncharacterized protein T22P22_20 [Source:UniProtKB/TrEMBL
GLYMA_06G001300	AT5G12230	Mediator of RNA polymerase II transcription subunit 19a [Source:UniProtKB/Swiss-Prot
GLYMA_06G001300	AT5G12230	Mediator of RNA polymerase II transcription subunit 19a [Source:UniProtKB/Swiss-Prot

GLYMA_18G284500	AT5G13100	AT5g13100/T19L5_60 [Source:UniProtKB/TrEMBL
GLYMA_03G221200	AT5G14900	Helicase associated (HA2) domain-containing protein [Source:UniProtKB/TrEMBL
GLYMA_18G043000	AT5G16560	Transcription repressor KAN1 [Source:UniProtKB/Swiss-Prot
GLYMA_16G151100	AT5G17570	At5g17570 [Source:UniProtKB/TrEMBL
GLYMA_17G016000	AT5G19180	NEDD8-activating enzyme E1 catalytic subunit [Source:UniProtKB/Swiss-Prot
GLYMA_06G001300	AT5G19480	Probable mediator of RNA polymerase II transcription subunit 19b [Source:UniProtKB/Swiss-Prot
GLYMA_06G001300	AT5G19480	Probable mediator of RNA polymerase II transcription subunit 19b [Source:UniProtKB/Swiss-Prot
GLYMA_15G033700	AT5G19620	Outer envelope protein 80, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_12G083600	AT5G22850	Eukaryotic aspartyl protease family protein [Source:UniProtKB/TrEMBL
GLYMA_15G207200	AT5G23050	Probable acyl-activating enzyme 17, peroxisomal [Source:UniProtKB/Swiss-Prot
GLYMA_06G209000	AT5G23090	AT5G23090 protein [Source:UniProtKB/TrEMBL
GLYMA_06G209000	AT5G23090	AT5G23090 protein [Source:UniProtKB/TrEMBL
GLYMA_19G065700	AT5G24170	Vesicle transport protein [Source:UniProtKB/TrEMBL
GLYMA_06G074300	AT5G25560	CHY-type/CTCHY-type/RING-type Zinc finger protein [Source:UniProtKB/TrEMBL
GLYMA_13G109600	AT5G35560	DENN (AEX-3) domain-containing protein [Source:UniProtKB/TrEMBL

GLYMA_08G157400	AT5G36870	Callose synthase 4 [Source:UniProtKB/Swiss-Prot
GLYMA_14G210900	AT5G37290	ARM repeat superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_16G071500	AT5G40450	unknown protein
GLYMA_09G281700	AT5G41690	RNA-binding (RRM/RBD/RNP motifs) family protein [Source:UniProtKB/TrEMBL
GLYMA_20G004000	AT5G41690	RNA-binding (RRM/RBD/RNP motifs) family protein [Source:UniProtKB/TrEMBL
GLYMA_15G158500	AT5G42400	SET domain protein 25 [Source:TAIR
GLYMA_06G030100	AT5G42810	Inositol-pentakisphosphate 2-kinase [Source:UniProtKB/Swiss-Prot
GLYMA_05G060500	AT5G47435	Formyltetrahydrofolate deformylase 1, mitochondrial [Source:UniProtKB/Swiss-Prot
GLYMA_08G265400	AT5G47860	AT5g47860/MCA23_20 [Source:UniProtKB/TrEMBL
GLYMA_04G002700	AT5G48335	At5g48335 [Source:UniProtKB/TrEMBL
GLYMA_10G118800	AT5G51230	Polycomb group protein EMBRYONIC FLOWER 2 [Source:UniProtKB/Swiss-Prot
GLYMA_20G155600	AT5G52100	CRR1 [Source:UniProtKB/TrEMBL
GLYMA_17G057800	AT5G53010	Calcium-transporting ATPase [Source:UniProtKB/TrEMBL
GLYMA_14G161200	AT5G53150	DnaJ heat shock amino-terminal domain protein [Source:UniProtKB/TrEMBL
GLYMA_04G231700	AT5G53570	Ypt/Rab-GAP domain of gyp1p superfamily protein [Source:UniProtKB/TrEMBL

GLYMA_06G254000	AT5G54310	ADP-ribosylation factor GTPase-activating protein AGD5 [Source:UniProtKB/Swiss-Prot
GLYMA_07G126600	AT5G56640	Inositol oxygenase 5 [Source:UniProtKB/Swiss-Prot
GLYMA_09G062900	AT5G56870	Beta-galactosidase 4 [Source:UniProtKB/Swiss-Prot
GLYMA_09G066800	AT5G57060	60S ribosomal L18a-like protein [Source:UniProtKB/TrEMBL
GLYMA_17G057800	AT5G57110	Calcium-transporting ATPase [Source:UniProtKB/TrEMBL
GLYMA_11G163900	AT5G58450	N-terminal acetyltransferase B complex auxiliary subunit NAA25 [Source:UniProtKB/Swiss-Prot
GLYMA_U034500	AT5G60100	Pseudo-response regulator 3 [Source:UniProtKB/TrEMBL
GLYMA_U034500	AT5G60100	Pseudo-response regulator 3 [Source:UniProtKB/TrEMBL
GLYMA_13G328100	AT5G60410	DNA-binding protein with MIZ/SP-RING zinc finger, PHD-finger and SAP domain [Source:TAIR
GLYMA_11G147400	AT5G60570	F-box/kelch-repeat protein At5g60570 [Source:UniProtKB/Swiss-Prot
GLYMA_11G147400	AT5G60570	F-box/kelch-repeat protein At5g60570 [Source:UniProtKB/Swiss-Prot
GLYMA_05G023200	AT5G61250	GUS1 [Source:UniProtKB/TrEMBL
GLYMA_02G290800	AT5G61640	peptidemethionine sulfoxide reductase 1 [Source:TAIR
GLYMA_17G094100	AT5G62580	TORTIFOLIA1-like protein 3 [Source:UniProtKB/Swiss-Prot
GLYMA_19G261500	AT5G63140	Probable inactive purple acid phosphatase 29 [Source:UniProtKB/Swiss-Prot

GLYMA_04G194200	AT5G63380	4-coumarate--CoA ligase-like 9 [Source:UniProtKB/Swiss-Prot
GLYMA_05G193000	AT5G64020	Protein trichome birefringence-like 14 [Source:UniProtKB/Swiss-Prot
GLYMA_07G181500	AT5G64360	AT5G64360 protein [Source:UniProtKB/TrEMBL
GLYMA_04G216400	AT5G64710	Emb [Source:UniProtKB/TrEMBL
GLYMA_08G133200	AT5G65290	LMBR1-like membrane protein [Source:UniProtKB/TrEMBL
GLYMA_17G222400	AT5G66240	Transducin/WD40 repeat-like superfamily protein [Source:UniProtKB/TrEMBL
GLYMA_02G052400	AT5G66380	Folate transporter 1, chloroplastic [Source:UniProtKB/Swiss-Prot
GLYMA_02G234300	AT5G67340	U-box domain-containing protein 2 [Source:UniProtKB/Swiss-Prot
GLYMA_05G105700	AT5G67540	Arabinanase/levansucrase/invertase [Source:TAIR