UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

BRUNA PIERECK MOURA

ANÁLISE DE ELEMENTOS TRANSPONÍVEIS NO TRANSCRIPTOMA (RNA-SEQ) DE *Vigna unguiculata* EM SITUAÇÕES DE ESTRESSE

RECIFE
2019

BRUNA PIERECK MOURA

ANÁLISE DE ELEMENTOS TRANSPONÍVEIS NO TRANSCRIPTOMA (RNA-SEQ) DE *Vigna unguiculata* EM SITUAÇÕES DE ESTRESSE

Tese apresentada ao Programa de Pós-Graduação em Genética da Universidade Federal de Pernambuco como parte dos requisitos exigidos para obtenção do título de Doutor em Genética.
Área de concentração: Biologia Molecular

Orientadora: Profa. Dra. Ana Maria Benko Iseppon

Coorientadora: Profa. Dra. Ana Christina Brasileiro Vidal

RECIFE

2019

Catalogação na Fonte:
Bibliotecário Bruno Márcio Gouveia, CRB4/1788

Moura, Bruna Piereck
Análise de elementos transponíveis no transcriptoma (RNA-SEQ) *Vigna unguiculata* em situação de stresse / Bruna Piereck Moura. – 201
9.
.

73 f. : il.

Orientadora: Profa. Dra. Ana Maria Benko Iseppon.
Coorientadora: Profa. Dra. Ana Christina Brasileiro Vidal.
Tese (doutorado) – Universidade Federal de Pernambuco. Centro de Biociências. Programa de Pós-graduação em Genética, Recife, 2019.
Inclui referências, apêndices e anexos.

1. Genética vegetal. 2. Bioinformática. 3. Leguminosa. I. Benko Iseppon, Ana Maria (orientador). II. Vidal, Ana Christina Brasileiro (coorientadora). III. Título.

581.35 CDD (22.ed.) UFPE/CB – 2022-196

BRUNA PIERECK MOURA

Análise de elementos transponíveis no transcriptoma (RNA-Seq) de *Vigna unguiculata* em situações de estresse

Aprovado em 24 / 09 / 2019

Banca Examinadora:

Profª Drª Ana Maria Benko Iseppon
Universidade Federal de Pernambuco

Dr. Gabriel da Luz Wallau
Instituto Aggeu Magalhães - FIOCRUZ IAM

Dr. Tercilio Calsa Junior
Universidade Federal de Pernambuco

Dra. Flávia Aburjaile
Universidade Federal de Pernambuco

Dra. Valesca Padolfi
Universidade Federal de Pernambuco

AGRADECIMENTOS

A concretização de um projeto como esse não se deve apenas a uma pessoa, mas antes, a todos aqueles que o viabilizaram e nele se envolveram. Partilharam-se dúvidas, incertezas, conquistas e muito, muito aprendizado.

Agradeço, em primeiro lugar, às minhas duas orientadoras, as Anas, que me ajudaram a construir o caminho que percorri desde a idealização até à concretização deste projeto, confiando e investindo na minha formação. Agradeço também aos professores e instituições que me receberam de braços abertos e foram essências para o meu amadurecimento: Professores Reinhard Schneider e Adriano Barbosa da universidade de Luxemburgo, onde estive por quase oito meses; Professores Guillaume Borque e David Morais, na McGill, onde estive por seis meses e suas respectivas equipes.

Também gostaria de deixar um agradecimento especial às instituições de pesquisa CAPES, FACEPE e CNPq por possibilitar a execução e as parcerias deste trabalho científico importantes para minha formação profissional. Sem deixar de lado a Universidade Federal de Pernambuco e o Programa de Pós-graduação em Genética (PPGG) que me acolhem desde o mestrado, colaborando com o desenvolvimento intelectual e científico dentro do ramo que, por amor, decidi seguir.

Seria, no entanto, impossível sobreviver aos intensos dias de trabalho e às inúmeras dúvidas científicas e existências que surgem ao longo do processo de desenvolvimento de uma tese de doutorado, que, além de um desenvolvimento científico, nos obriga a passar por um processo maior de autoconhecimento e exercício de resiliência, se não fossem a família e os amigos. Por isso deixo meu forte agradecimento aos amigos que seguiram comigo nessa jornada, aos amigos que conheci durante os meus intercâmbios que me ajudaram ativamente, com discussões científicas, aulas de programação, apoio emocional e terapias tanto de risada, quanto de choque (quando necessário).

Por último, e não menos importante, agradeço à minha família, por me aguentar entre a euforia da nova descoberta e o mau-humor dos imprevistos; por segurar firme na minha mão, sem me deixar desistir desse sonho e dessa jornada, quando duvidei de mim mesma; por perdoarem, muito compreensivamente, das fugas nas reuniões de família para poder terminar o

meu trabalho. Agradeço a todo amor que recebi em forma de ensino, puxões de orelha, apoio e carinho.

Obrigada a todos por tornarem esse projeto possível.

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes.”
Marthin Luther King -

RESUMO

Estresses ambientais provocam perdas significativas na produção agrícola. Para minimizá-las, é necessário compreender os mecanismos de resposta vegetal, incluindo o papel de elementos transponíveis (TEs). Duas abordagens foram usadas visando entender os processos envolvidos na resposta a estresses que podem estar relacionados aos elementos transponíveis em *Vigna unguiculata*. Na primeira, a mineração de texto foi realizada com o pipeline LAITOR4HPC (Capítulo I), baseado na ferramenta online PESCADOR. O trabalho permitiu o processamento eficiente de aproximadamente 31 milhões de resumos em seis dias, diminuindo drasticamente o tempo original estimado (dois anos). Com o resultado, foi possível mapear na literatura interações super- e sub-representadas em soja, bem como construir e enriquecer uma via de regulação de defensina em *Arabidopsis thaliana*. Contudo, uma busca com a mesma abordagem sobre TEs evidenciou a necessidade de um dicionário específico. Sabe-se que TEs são capazes de provocar alterações estruturais e epigenéticas, sendo modulados e responsivos em plantas sob estresse. Portanto, na segunda abordagem (Capítulo II) TEs diferencialmente expressos (TE-DEGs) foram identificados e anotados no transcriptoma de *V. unguiculata* com o intuito de compreender seu papel na resposta tolerante/resistente desta leguminosa. A análise incluiu bibliotecas de folhas inoculadas com dois vírus de RNA e de raízes submetidas a desidratação. Foram identificadas 12 superfamílias de TEs-DEGs, além de TEs cuja classificação é desconhecida (*unk*) e *nested*-TEs. Elementos transponíveis abrangeram entre 7% e 20% do total de TE-DEGs para cada tratamento. Houve especificidade de isoformas e de categorias de ontologia gênica (GO) comparando-se os tratamentos e tempos. A presença TE-DEGs com anotação GO de fatores de transcrição e domínios completos associados à regulação de transcrição indicam um possível papel desses elementos na resposta aos estresses em tela, demandando validação experimental.

Palavras-chave: bioinformática; estresse abiótico; estresse biótico; expressão diferencial; mineração de texto; montagem de transcriptoma.

ABSTRACT

Environmental stresses lead to significant losses in crop production. To minimize them, it is necessary to understand plant response mechanisms, including the role of transposable elements (TEs). Two approaches were used aiming to understand the TE-related stress responses in *Vigna unguiculata*. The first used text mining analysis with LAITOR4HPC (Chapter I), based on an improvement of PESCADOR online tool. The former enabled processing approximately 31 million abstracts in six days, drastically decreasing the original estimated time (two years). The result, allowed to map super- and sub- represented interactions in soybean, as much as to build a defensin regulatory network in *Arabidopsis thaliana*. However, a search with the same approach targeting TEs indicated the need for a specific dictionary. It is known that TEs are capable of causing structural and epigenetic changes, being modulated and responsive in plants under stress. Thus, in the second approach (Chapter II), differentially expressed TEs (TEs-DEGs) were identified and annotated in RNA-Seq data of *V. unguiculata* transcriptome, aiming to understand its role at tolerant/resistant response in this legume. The study comprised leaf libraries inoculated with two RNA virus and roots subjected to dehydration. Twelve TE-DEGs superfamilies were identified, as well as TEs of unknown classification (*unk*) and *nested*-TEs. Transposons embraced between 7% and 20% of all TE-DEGs. Some isoforms and gene ontology (GO) categories are specific when comparing treatments and times. The presence of TE-DEGs with transcription factor GO annotation and complete domains associated with transcription regulation indicate a possible role of these elements in the response to the targeted stresses, requiring experimental validation.

Key words: bioinformatics; abiotic stress; biotic stress; differential expression; text mining; transcriptome assembly.

LISTA DE FIGURAS

TESE

ARTIGO 1 - EXPRESSÃO DIFERENCIAL DE ELEMENTOS TRANSPONÍVEIS NO TRANSCRIPTOMA (RNA-SEQ) DO FEIJÃO-CAUPI EM SITUAÇÕES DE ESTRESSE

LISTA DE QUADROS

TESE

LISTA DE TABELAS

ARTIGO 1 - EXPRESSÃO DIFERENCIAL DE ELEMENTOS TRANSPONÍVEIS NO TRANSCRIPTOMA (RNA-SEQ) DO FEIJÃO-CAUPI EM SITUAÇÕES DE ESTRESSE

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

Item	Definição Inglês	Definição Português
CRE	*Cis-Acting Regulatory Element*	Elemento CIS de Regulação
CGKB	*Cowpea Genespace/Genomics Knowledge Base*	Base de Conhecimento Genômico/Espaço Gênico de Feijão-Caupi
DCL	*DICER-Like protein*	Proteína DICER-Like
dsDNA	*double strand DNA*	DNA dupla fita
dsRNA	*double strand RNA*	RNA dupla fita
EN	*Endonuclease*	Endonuclease
epsiRNAs	*Epigenetic-Derived* (Tes-Derived) *siRNAs*	siRNA Derivados do controle epigenético (Derivado de TEs)
FAO	*Food and Agricultural Organization*	Organização De Agricultura E Alimentos
FPKM	*Fragments Per Kilobase of Feature Sequence Per Million Fragments Mapped*	Fragmentos por Milhões de Quilobases
IR	*Information Retrieval*	Recuperação da Informação
LINEs	*Long Interspersed Nuclear Elements*	Longos Elementos Intercalares Nucleares
LTRs	*Long Terminal Repeats*	Longas Repetições Terminais
miRNA	*micro RNA*	micro RNA
MITEs	*Miniature Inverted Repeats*	Miniaturas Invertidas Repetidas
mRNAs	*messenger RNA*	RNA mensageiro

NCBI	*National Center for Biotechnology Information*	Centro Nacional para Informação Biotecnológica
NGS	*Next Generation Sequencing*	Sequenciamento de Nova Geração
NLP	*Natural Language Process*	Processo de Linguagem Natura
PLEs	*Penelope-Like Elements*	Elementos *Penelope-Like*
QTL	*Quantitative Trace Locus*	*Locus* de Característica Quantitativa
RSEM	*RNA-Seq by Expectation-Maximization*	RNA-Seq por Expectativa de Maximização
RDR	*RNA Dependent RNA Polimerase*	RNA Polimerase Dependente de RNA
RT	*Reverse Transcriptase*	Transcriptase Reversa
SINEs	*Small Interspersed Nuclear Elements*	*Pequenos Elementos Intercalares Nucleares*
siRNA	*Small Interference RNA*	Pequenos RNAs De Interferência
SNP	*Single Nucleotide Polymorphism*	Polimorfismo de Base Única
sRNA	*Small RNAs*	Pequenos RNAs
ssDNA	*Single strand* DNA	DNA fita simples
ssRNA	*Single strand RNA*	RNA Fita Simples
Super-SAGE	*Super-Serial Analysis of Gene Expression*	Análise Super Seriada da Expressão Gênica
TEs	*Transposable Element*	Elementos Transponíveis
TEs-MIR	*TE-Derived miRNA Genes*	Genes de miRNA Derivados de TEs
TIRs	*Terminal Inverted Repeats*	*Repetições Terminais Invertidas*

TNP	*Transposase*	Transposase
TSS	*Transcription Starting Site*	Sítio de Iniciação da Transcrição

SUMÁRIO

1 INTRODUÇÃO

Estresses ambientais, sejam de origem biótica ou abiótica, provocam perdas significativas na produção agrícola, por isso, demandando um melhor entendimento dos mecanismos de resposta do vegetal, incluindo aqueles envolvendo espécies de interesse para a agricultura na região Nordeste do Brasil, como no caso feijão-caupi [*Vigna unguiculata* (L.) Walp.]. Os mecanismos de resposta podem ser estudados a partir da construção de vias de interação ou dados de RNA-Seq que permitem avaliar a expressão gênica.

O feijão-caupi tem origem africana e representa uma das principais leguminosas produtoras de grãos em áreas de baixa pluviosidade. No Brasil vem sendo produzido nas regiões Norte e Nordeste, onde o feijão comum (gênero *Phaseolus*) não é cultivável. Apesar de sua elevada plasticidade, a referida cultura apresenta produtividade limitada, especialmente devido a fatores ambientais (bióticos e abióticos). Visando entender processos moleculares do feijão-caupi em resposta a diferentes situações de estresse biótico e abiótico, foram gerados transcriptomas (RNA-Seq) com tratamentos envolvendo a cultura citada, além da realização de uma busca por mineração de texto de processos moleculares previamente descritos.

Entre os alvos moleculares, destacam-se elementos transponíveis (TEs), que possuem um papel crucial na evolução dos genomas e na modulação da expressão gênica, sendo em muitos casos transcricionalmente ativos. Devido à sua natureza móvel, TEs são capazes de provocar alterações no genoma a partir de mecanismos, como: (i) ativação/inativação de genes; (ii) geração de novos genes e pseudogenes; (iii) atuação ou cooptação como promotores; (iv) geração de elementos-cis e pequenos RNAs de interferência (siRNAs), entre outros. Seus efeitos podem ser intensificados quando o organismo hospedeiro se encontra sob estresse. No entanto, sabe-se que sua expressão pode ser cultivar, tecido e condição específica. Alguns TEs se expressam preferencialmente durante o estresse, funcionando também como genes responsivos no processo de defesa das plantas.

Adicionalmente, a análise da atividade de TEs pode ser feita mediante sua inclusão na construção de vias metabólicas, auxiliando no entendimento de sua atividade e sua influência durante um determinado estresse ou condição. Para tal, o

uso da mineração de texto pode auxiliar na detecção e destaque de TEs atuantes como promotores, ativadores e geradores de micro RNAs (miRNA) que tenham sido previamente descritos. A mineração de texto se baseia na identificação de entidades biológicas (proteínas, DNA, RNA) em um texto, seguida do reconhecimento de termos de interação. O seu resultado pode ser traduzido em uma imagem representativa das interações moleculares identificadas, ou seja, numa rede de interação.

Dessa forma, o presente trabalho visou identificar elementos transponíveis relacionados com a resposta a estresse a partir das duas abordagens: (I) Usando a mineração de texto em busca de interações previamente descritas e (II) Acessando o transcriptoma a fim de elucidar as respostas moduladas pela planta sob estresse, selecionando transcritos de TEs de interesse.

2 REVISÃO DA LITERATURA

2.1 TES: IMPORTÂNCIA E PAPEL NOS GENOMAS DAS PLANTAS

Bárbara McClintock identificou os *loci* mutáveis ou genes instáveis, hoje conhecidos como elementos transponíveis (TEs), enquanto buscava compreender as alterações no padrão de cores do tegumento dos grãos de milho. Em 1950, McClintock publicou sua primeira descrição e alertou: “As observações acumuladas sobre esses numerosos *loci* mutáveis são tão extensivas, que nenhum breve relato daria suficiente informação para preparar o leitor para um julgamento independente sobre a natureza desse fenômeno” (MCCLINTOCK, 1950). Seu trabalho foi ignorado até ser reconhecido e agraciado com o prêmio Nobel em 1983, mais de 30 anos depois (MCCLINTOCK, 1984). A descoberta dos TEs quebrou dogmas e abriu as portas para uma nova visão sobre a estrutura, organização dos genomas.

Os TEs eram inicialmente apontados como os principais responsáveis pela falta de correlação entre o tamanho dos genomas e a complexidade das espécies, pois compõem uma grande proporção nos genomas disponíveis (GEMAYEL et al., 2012), variando entre cerca de 10% do genoma de *Arabidopsis thaliana* L. (ARABIDOPSIS GENOME INITIATIVE, 2000) a 85% do genoma de *Zea mays* (CHÉNAIS et al., 2012). Atualmente, sabe-se que apesar dos resultados deletérios os TEs podem possuir um papel crucial na evolução dos genomas de todos os grupos de organismos, com ênfase para vegetais (BENNETZEN; WANG, 2014; PLATT; VANDEWEGE; RAY, 2018) e são capazes de atuar de formas variadas: (i) ativação/inativação de genes; (ii) geração de novos genes ou pseudogenes; (iii) atuação ou cooptação como promotores; (iv) geração de elementos-cis e de pequenos RNAs (sRNAs), entre outros. Seus efeitos podem ser ainda mais intensos quando o organismo hospedeiro encontra-se sob estresse (BENNETZEN, 2000; BUNDOCK; CHÉNAIS et al., 2012; HOOYKAAS, 2005). Apesar da inserção de TEs ocorrer de forma aleatória, em alguns casos, há algumas famílias que se inserem em alvos específicos ou apresentam regiões preferenciais de inserção (ELLIOTT et al., 2013; GUÉRILLOT et al., 2014; LIU et al., 2009).

2.1.1 TEs e sua classificação

Como consequência de sua abrangência e diversidade, surgiu a necessidade de organizar os TEs em um sistema de classificação que levasse em conta sua diversidade e complexidade. Várias têm sido as propostas de classificação, não havendo ainda unanimidade. Duas propostas de classificação serão abordadas no presente trabalho. A primeira de Wicker et al. (2007), chamada de “Proposta de Wicker”, que se baseia no mecanismo básico de transposição e sua composição enzimática. A segunda, mais recente, foi proposta por Piégu et al. (2015), sendo chamada de “Proposta de Cúrcio e Derbyshire”, por ser baseada em artigo publicado em 2003 pelos autores citados (CÚRCIO; DERBYSHIRE, 2003), avaliando similaridades e diferenças nos mecanismos das transposases, bem como sua origem evolutiva.

2.1.1.1 Proposta de Wicker

Wicker et al. (2007) separam os TEs em duas grandes classes, com base em seus intermediários de DNA ou RNA. Os elementos classe I (Retrotransposons) se transpõem a partir de um intermediário de RNA, que, após ser traduzido em cDNA por uma transcriptase reversa, gera uma cópia a qual é transportada e inserida por uma transposase/integrase ao material genético. Por sua vez, os elementos classe II (DNA transposons) transpõem sua sequência de DNA integralmente, podendo possuir uma ou duas transposases (TNP), as quais são enzimas capazes de quebrar as fitas de DNA, mover a sequência e inserir elementos em um novo local.

Dentro deste sistema (WICKER et al., 2007), as classes I e II são subdivididas em subclasses, ordens, superfamílias e famílias, de acordo com os critérios descritos a seguir: (i) subclasses distinguem elementos que criam cópias para serem inseridas ou reintegradas em regiões diferentes; (ii) ordens, as quais são definidas de acordo com o mecanismo de inserção; (iii) superfamílias, que compartilham a estratégia de replicação e são agrupadas de acordo com a uniformidade de sua estrutura proteica, presença de domínios não codificantes e duplicações de sítio alvo; (iv) famílias se diferenciam pela similaridade de sequência,

sendo necessária no mínimo 80% de similaridade para ser considerada da mesma família; e (v) subfamílias são distinguidas com base filogenética.

A classe I (Retrotransposons) está dividida em cinco ordens (LTRs, LINEs, SINEs, DIRS e PLEs), com um total de 17 superfamílias (Quadro 1). Por sua vez, a classe II (DNA-transposons) é dividida em duas subclasses (I e II) com duas ordens cada e um total de 12 superfamílias (Quadro 2) (WICKER et al., 2007).

Quadro 1 - Classificação dos Retrotransposons (classe I), proposta de Wicker et al. (2007), considerando ordens e superfamílias.

Ordem	Superfamília
LTRs (*long terminal repeats*)	*Copia*
	Gypsy
	Bel-Pao
	Retrovírus
	VER
SINEs (*Short interspersed nuclear elements*)	*tRNA*
	7SL
	5S
PLEs (*Penelope-like elements*)	*Penelope*
LINEs (Long interspersed nuclear elements)	*R2*
	RTE
	Jockey
	L1
	I
DIRS-Like	*DIRS*
	Ngaro
	VIPER

Fonte: Wicker et al. (2007).

No entanto, a classificação de Wicker et al. (2007) apresenta dois problemas: (I) não inclui claramente os elementos *MITE* (*Miniature Inverted-Repeats Transposable Elements*), geralmente classificados como elementos da classe II por

serem flanqueados por TIRs (*Terminal Inverted Repeats*), embora possuam grande número de cópias de modo semelhante aos elementos da classe I (WICKER et al., 2007; YE; JI; LIANG, 2016); (II) exclui todos os elementos transponíveis existentes em microrganismos, como realçado pelo título de seu artigo “Sistema de classificação unificada para elementos transponíveis em eucariotos” (PIÉGU et al., 2015).

Quadro 2 - Classificação dos DNA transposons (classe II), proposta de Wicker et al. (2007) considerando as diferentes subclasses, ordens e superfamílias.

Subclasse	Ordem	Superfamília
Subclasse I	***TIR*** (*Terminal inverted repeat*)	*CACTA*
		Mutator
		Merlin
		Transib
		P
		hAT
		PiggyBac
		PIF/harbinger
		Tc1-mariner
	Crypton	*Crypton*
Subclasse II	***Helitron***	*Helitron*
	Maverick	*Maverick*

Fonte: Wicker et al. (2007).

2.1.1.2 Proposta de Cúrcio e Derbyshire

Piégu et al. (2015) propuseram um novo sistema de classificação baseado em três critérios: (1) semelhanças no fenótipo da transposição; (2) convergências evolutivas, e (3) mecanismos similares que podem ter processos evolutivos diferentes. A proposta visou criar um sistema unificado para procariotos e eucariotos, baseado principalmente em uma revisão sobre os mecanismos de transposição propostos por Cúrcio e Derbyshire (2003). Contudo, Piégu et al. (2015) enfatizaram

que a nova proposta de classificação também possui falhas. Para minimizá-las sugeriram a criação de um comitê internacional para discutir e unificar o sistema de classificação de TEs, usando todas as propostas existentes.

A proposta de Cúrcio e Derbyshire (2003) apresenta classes definidas em três grandes grupos: (I) TEs que usam o maquinário do hospedeiro na sua transposição (Quadro 3); (II) não LTRs (Quadro 4) e, (III) de acordo com o tipo de nuclease/recombinase (DDE-transposon, Y1-transposon, Y2-transposon, S-transposon, ou classificação pendente) (Tabela 5). As ordens são definidas de acordo com sua endonuclease (EN) ou mecanismo de transposição, e as superfamílias separadas de acordo com a relação filogenética baseada na estrutura de sua nuclease, recombinase, endonuclease e/ou transcriptase reversa (RT). De modo semelhante a proposta de Wicker et al. (2007), Piégu et al. (2015) propuseram oito classes, incluindo 25 ordens e mais de 50 superfamílias até o momento. Em ambas as propostas, as ordens e as superfamílias são escritas sempre em itálico (WICKER et al., 2007; PIÉGU et al., 2015).

Quadro 3 - Classificação dos elementos transponíveis, proposta de Cúrcio e Derbyshire (2015), grupo que usa o hospedeiro para transposição.

Classe	Ordem	Superfamília
Intein	LAGLIDADG, inteins, HNH inteins	Nome do gene em que se inserem especificamente
Grupo I intron (G1i)	LAGLIDADG, G1i, HNH G1i, His Cys G1i, PD (D/E)XK G1i, Vsr G1i	Nome do sítio específico de inserção

Fonte: Cúrcio e Derbyshire (2015).

Quadro 4 - Classificação dos elementos transponíveis, proposta de Cúrcio e Derbyshire (2015), grupo não LTR. 1EN – Endonuclease.

Classe	Ordem	Superfamília
Retrotransposons	LINEs (Long interspersed nuclear elements)	LINEs com EN[1] apurinica/apirimidinica LINEs com motivo PD-(D/E)XK em LINEs com ambas características
	PLEs (Penelope-like elements)	Athena, Caprina, Neptune, Penelope
	Grupo II introns	Grupo II intron, Mobile lariat intron, Intron-like

Fonte: Cúrcio e Derbyshire (2015).

Quadro 5 - Classificação dos elementos transponíveis, proposta de Cúrcio e Derbyshire (2015), cujo grupo contém Nuclease/ Recombinase.

Classe	Ordem	Superfamília
DDE-transposon	***DDE transposons*** (*copy-in*)	*Um*
		Tn3
	***DDE transposons* tipo 1** (dsDNA linear → *cut-out/ past-in*)	*IS (1, 3, 4, 6, 21, 30, 66, 110, 701, 982, 1380, 1182, 1634)*
		ISH3
		ISAs1
		ISL3
		IS630/ Tc1-mariner/Zator
		IS1595/ PIF-harbinger
		IS256/ MuDR/ Mutator/ Rehavkus
		IS1380/ PiggyBac
		Academ
		CACTA/Mirage/Chapev (CMC)
		Dada
		hAT
		Kolobok
		P
		Sola
		Transib
	***DDE/D transposon* tipo 2** (dsDNA → *cut-out/ past-in*)	*Tn7*
	DDE/D transposons (dsDNA circular → *copy-out/ past-in*)	*IS3*
	LTR retrotransposon (*Copy-out/ past-in*)	*Copia*
		Gypsy
		BEL
		ERV1
		ERV2
		ERV3
Y1	***Y1 transposon*** (dsDNA circular → *Cut-out/ past-in*)	*IS (200,605)*
		Tn916
		CTnDOT
		Crypton
	Y1 transposon (dsDNA circular → *Copy-out/ past-in*)	*DIRS*
		Ngaro

		VIPER
Y2	***Y2 transposon*** (ssDNA circular → *copy-in or -out/ past-in*)	*IS91*
		Helitrons
S	***S-transposon*** (dsDNA circular → *cut-out/ past-in*)	*IS607*
		Tn5397
Classificação Pendente	***?***	*ISAs1*
	?	*Fanzor*
	Politons/ Mavericks (*copy-in or -out/ copy-in*)	*Marvirus*
		Politons/Mavericks
		Tlr1
	Transposase provavelmente relacionada às integrases do *LTR retrotransposon*	*Ginger (1 e 2)*
	***DDE transposons* tipo 3**	*P*
	Zisupton	*Zisupton*

Fonte: Cúrcio e Derbyshire (2015). dsDNA – DNA fita dupla; ssDNA – DNA fita simples; ? – Ordens ainda sem nomes sugeridos

2.1.2 Expressão de TEs

O estresse é considerado o principal causador da regulação transcricional de TEs. Por esta razão, diversos estudos visam compreender o seu comportamento diante das adversidades. Em arroz inoculado com vírus, por exemplo, 12 a 14% dos TEs foram diferencialmente expressos, com indução proporcional ao tempo de estresse (CHO et al., 2015). Por outro lado, em *Pinus halepensis* tolerante à seca, observou-se que durante tratamento de seca, a maioria dos TEs encontrava-se reprimido, contudo, após reiniciar a rega, os TEs induzidos tornaram-se mais abundantes (FOX et al., 2018). De modo semelhante, durante desidratação, 2-4% dos TEs foram modulados e majoritariamente reprimidos em *Arabidopsis* (GÖBEL et al., 2018).

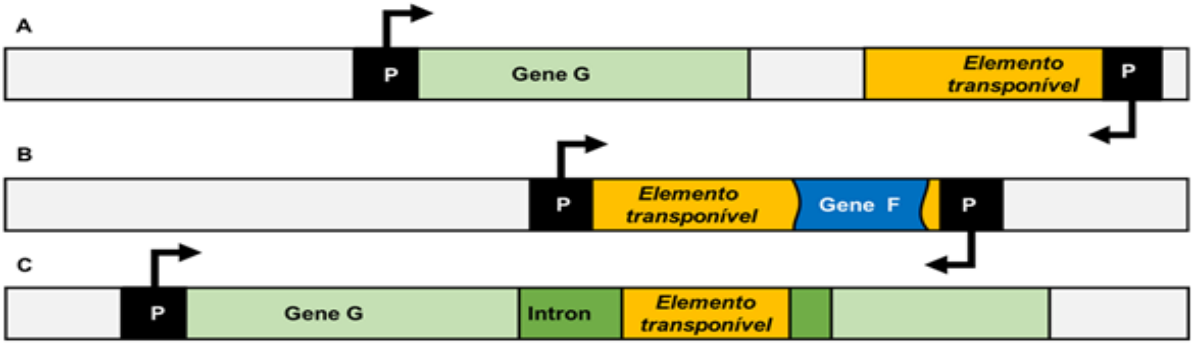
Anteriormente, acreditava-se que o silenciamento epigenético dos TEs era influenciado por situações de estresse, consideradas os principais agentes causais do aumento na atividade de transposição (SLOTKIN; MARTIENSSEN, 2007; ZEH; ZEH; ISHIDA; 2009). Ao contrário do que se pensava anteriormente, já se sabe que

nem sempre a mobilidade de TEs é induzida durante o estresse e que a baixa metilação nem sempre é suficiente para ativar a sua transcrição (CAVRAK et al., 2014). Sabe-se também que promotores de TEs geralmente são bastante específicos, com promotores de TEs de arroz apresentando alta similaridade com promotores responsivos a estresse, enquanto em milho continham sequências de TFs (FINATTO et al., 2015; MAKAREVITCH et al., 2015). Além disso, a cooptação de seus promotores pode ser responsável pela especificidade de expressão de alguns genes, como por exemplo na planta revivesceste *B. hygrometrica*, onde um *nested-TE* de *LTRs* (*Copia* e *Gypsy*) foi identificado carregando o gene OAR1 (*osmotic and alcaline resistance*) expresso a partir de um promotor na LTR, aumentando a tolerância à desidratação e o *fitness* (taxa de sobrevivência, crescimento, nível de integridade de membrana e eficiência fotoquímica) (LE et al., 2014; WU et al., 2018; ZHAO, Y. et al., 2014). Atualmente se assume a importância da regulação epigenética dirigida para e por TEs, além de ressaltar que o ressequenciamento de dados e as tecnologias NGS (*Next Generation Sequencing*, Sequenciamento de Nova Geração) devem favorecer o entendimento mais aprofundado do papel desses elementos (CHO, 2018).

Adicionalmente, a geração de pequenos RNAs de interferência (siRNAs) e micro RNAs (miRNAs) a partir da expressão de TEs (TE-MIRs e epsiRNAs; capazes de modular os próprios elementos e os genes) representa um mecanismo flexível e rápido de adaptação dos organismos (LISCH; BENNETZEN, 2011). Geralmente TE-MIRs e epsiRNAs são originados de um único TE (LI, Y.; LI; JIN, 2011; MCCUE; SLOTKIN, 2012), embora haja casos de elementos transponíveis inseridos em outros TEs (*nested*-TEs) que formam TE-MIRs (PIRIYAPONGSA et al., 2008) deixando as possibilidades ainda mais complexas.

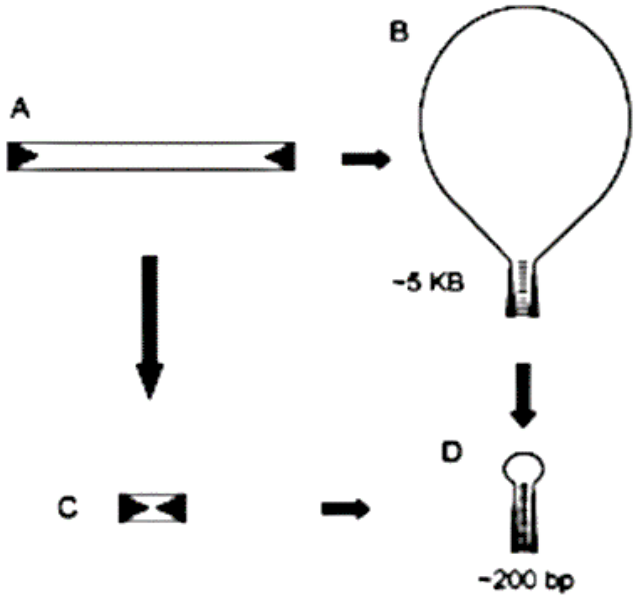
A biogênese dos TE-MIRs e epsiRNAs ocorre por: (i) formação de RNA fita dupla (dsRNA), ou pela expressão de sequências complementares, ou ainda pela ação da RNA polimerase dependente de RNA (RDR) (Figura 1A) (LI; LI; JIN, 2011); (ii) pela captura e inversão de sequências gênicas, cujo fragmento capturado pode ser complementar ao do gene parental (Figura 1B) (BORGES et al., 2018); e (iii) a partir da estrutura de terminações complementares, por exemplo as TIRs, elementos que ao serem expressos favorecem a formação de grampos (*hairpins*) (Figura 1C e 2) (PIRIYAPONGSA; JORDAN, 2008). As ditas estruturas quando reconhecidas pela DCL geram TEs-MIR ou epsiRNAs.

Figura 1 - Formação de transcrito complementar por elementos trasnponíveis (TEs).



Fonte: A autora (2019). Na imagem, em cinza claro o genoma, em verde o gene G, em azul o fragmento do gene F, em preto o promotor (P) sinalizando a direção da transcrição e em amarelo o TE. A Quando transcrito, o TE pode também induzir a transcrição da fita complementar do gene G gerando um RNA fita dupla (dsRNA) do gene. B Alguns elementos possuem promotor nas duas terminações, sendo capazes de transcrever sequências complementares tanto ao elemento, quanto do gene capturado. O transcrito do Gene F pode ser complementar ao do gene parental não capturado. C TEs cujas terminações são complementares (por exemplo, elementos TIR; terminal inverted repeat) podem formar estrutura de grampo (harpin) ou pela transcrição ou por splicing quando inseridos em introns.

Figura 2 - Formação de hairpin por Elementos transponíveis TEs.



Fonte: Piriyapongsa e Jordan (2008). Na figura a TIR (repetição terminal invertida) é representada por triângulos pretos. Em A esta a estrutura de Tc1-mariner completo se dobra em forma de grande hairpin pelo pareamento das TIRs como mostrado em B; Em C um elemento MITE (miniature inverted repeat) está representado, seguido por D, formando sua estrutura de hairpin.

Apesar dos indícios da interação entre TEs e genes, como proposto por McClintock (1950, 1953), só recentemente seus mecanismos de ação e abrangência têm sido revelados. Os TEs deixaram de ser considerados DNA-lixo e atualmente são aceitos como estruturas fundamentais para os organismos (GEMAYEL et al., 2012). Em espécies do gênero *Brassica* aproximadamente 70% dos sítios de ligação ao fator de transcrição E2F, que inclui ativadores e repressores de várias etapas do ciclo celular, estão dentro de estruturas de TEs (HÉNAFF et al., 2014). Também, a indução de elementos da ordem *LTR* tem sido relacionada à tolerância a estresses abióticos em diversas espécies, como em café (LOPES et al., 2013), cevada (SUONIEMI; NARVANTO; SCHULMAN, 1996) e alfafa (IVASHUTA et al., 2002), geralmente coincidindo com a presença de motivos conservados nos promotores (GRANDBASTIEN, 2015). Além disso, vem sendo destacada a expressão de TEs em tecidos, acessos e condições específicas (BUNDOCK; HOOYKAAS, 2005; FINATTO et al., 2015; TOVKACH et al., 2013; WU et al., 2018), como observado, por exemplo, em *A. thaliana* (WU et al., 2018), arroz (HANADA et al., 2009; ZHAO, D. et al., 2018) e milho (RAIZADA; BREWER; WALBOT, 2001) para as duas classes de TEs.

Pouco enfoque tem sido dado a TEs expressos constitutivamente, como revisado por Grandbastien (2015) para as *LTRs*. Em seu trabalho, a autora relata que existe uma grande proporção de *LTRs* sendo expressos constitutivamente em tecido meristemático. Em folhas, raízes e flores de ervilha (TOVKACH et al., 2013), bem como raiz de trigo (NEUMANN; POZÁRKOVÁ; MACAS, 2003), embora sua função não esteja esclarecida para ervilha, no trigo há indícios de que sejam responsáveis pela tolerância a alumínio. Vale salientar que a expressão constitutiva do elemento classe II, *hAT*, foi observada em gramíneas em relação a uma transposase truncada cuja função ainda precisa ser mais bem elucidada (MUEHLBAUER et al., 2006).

2.2 *Vigna unguiculata*

2.2.1 Importância socioeconômica

A espécie *V. unguiculata* (família Fabaceae) é conhecida popularmente como feijão-caupi. Possui origem africana, com cultivo reportado em mais de 97

países (FREIRE FILHO, 1988; FREIRE FILHO et al., 2011; PASSOS et al., 2007). Seus grãos são considerados uma excelente fonte de proteínas (20-25% de sua composição), sendo também ricos em fibras e aminoácidos essenciais, normalmente escassos em dietas à base de cereais (AKIOBODE; MAREDIA, 2012).

Entre 2016 e 2017, o feijão-caupi apresentou produção mundial anual de aproximadamente 74 milhões de toneladas, com destaque para a África responsável por 94.8% da sua produção. As Américas são responsáveis por cerca de 1,6% - 76,9 mil toneladas (FAO, 2019). Contudo, dados da produção brasileira não estão incluídos, devido à dificuldade na distinção entre o feijão comum e o feijão-caupi, pela falta de bases de dados e estatísticas distintas, fato que afeta países como Brasil (FREIRE FILHO et al., 2011) e a Índia (AKIOBODE; MAREDIA, 2011). No entanto, de acordo com estimativas da Empresa Brasileira de Pesquisa Agropecuária - Embrapa, em 2016 apenas o Brasil produziu cerca de 345 mil toneladas de feijão-caupi (CNPAF – EMBRAPA, 2021).

No Brasil o feijão-caupi é comercializado na forma de grão seco ou de grão verde. Este segundo, em especial, exige muito trabalho manual que acaba sendo suprido por agricultores familiares. A produção de feijão-caupi tem se expandido para regiões de cerrado gerando bom rendimento para pequenos agricultores em virtude do baixo custo e da plasticidade. Adicionalmente, houve aumento da produção no país devido à demanda mundial de proteínas de baixo custo (BASTOS, 2016). Contudo, embora as leguminosas contabilizem a segunda categoria mais cultivada do mundo, esse grupo não tem recebido a devida atenção e investimento, motivo pelo qual as culturas vegetais desse grupo não têm alcançado todo o seu potencial de produção (AKIOBODE; MAREDIA, 2011).

2.2.2 Fatores limitantes da produção

Apesar da elevada plasticidade fenotípica do feijão-caupi, seu cultivo apresenta baixa produtividade quando comparado ao de outros feijões (BOUKAR et al., 2015; FREIRE FILHO et al., 2011; SINGH, 2006). São diversos os estresses bióticos que acomete a produção do feijão-caupi, como por exemplo: pragas (BOUKAR et al., 2015); fungos e vírus (BENCHIMOL et al., 2017; THOTTAPPILLY; ROSSEL, 1992); parasitas de sementes (OMOIGUI et al., 2016), entre outros.

A ocorrência de déficit hídrico também é muito comum nas principais áreas de cultivo do feijão-caupi, em especial na região semiárida do Nordeste brasileiro (FRITCHE NETO; BORÉM, 2011; MENDES et al., 2007; NASCIMENTO et al., 2011). Apesar de estudos recentes provarem que alguns acessos desta cultura são capazes de recuperar rapidamente as funções fisiológicas na reidratação após estresse hídrico (FREITAS et al., 2016), sendo tolerante à falta ou à baixa disponibilidade de água, este estresse ainda é um dos principais fatores limitantes da produtividade da cultura (BOUKAR et al., 2015; FREITAS et al., 2016).

A escassez de água, especialmente durante a floração e o enchimento dos grãos, afeta negativamente a planta (GUIMARÃES; STONE; BRUNINI, 2006), culminando na redução da produtividade. No entanto, a baixa produtividade pode ser minimizada através emprego de tecnologias adequadas (NASCIMENTO et al., 2011), do plantio de variedades melhoradas tolerantes ao déficit hídrico (MENDES et al., 2007), bem como o estudo de variedades crioulas (não melhoradas) que apresentem fenótipo resistente ou tolerante ao estresse em questão (FREITAS et al., 2016; FRITCHE NETO; BORÉM, 2012). Nesse sentido, diversas pesquisas têm sido desenvolvidas visando à compreensão dos mecanismos de resistência/tolerância, bem como à identificação de genes responsivos a estresses em cultivares selecionadas, a fim de viabilizar o melhoramento genético e aumento da produtividade (BOUKAR et al., 2015; FATOKUN; BOUKAR; MURANAKA, 2012).

2.2.3 Bancos de dados de V. unguiculata

Em 2007, surgiu a primeira iniciativa de sequenciamento e montagem de sequências genômicas de feijão-caupi, totalizando 298.848 sequências obtidas a partir de filtragem por metilação, disponíveis no banco CGKB (‘*Cowpea Genespace/Genomics Knowledge Base’*). O banco focou em anotação funcional acessada a partir de análise BLAST (*Basic Local Alignment Search Tool*), contra os bancos GenBank (SAYERS et al., 2018) e UniProt (CHEN et al., 2007). Apenas recentemente foi publicado um sequenciamento genômico de alta densidade e disponibilizado no banco de dados Phytozome (LONARDI et al., 2019) como comentado mais adiante.

Apesar da grande importância socioeconômica do feijão-caupi, até recentemente a disponibilidade de dados “ômicos” (genoma, transcriptoma, etc.) estava principalmente limitado ao âmbito da transcriptômica, com iniciativas como a da rede NordEST com dados de EST (*Expressed Sequence Tag*) e de Super-SAGE (*Super-Serial Analysis of Gene Expression*; KIDO et al., 2011), que foi criada com o intuito de estudar as respostas de resistência e tolerância de plantas de importância socioeconômica, como o feijão-caupi e originou, o Cowpea Genome Consortium (CpGC; Jesus-Pires et al. 2019) que contém mais de 500 milhões de transcritos referentes a tratamentos com acessos de feijão-caupi, incluindo experimentos de salinidade, desidratação, resistência a vírus e avaliação da resposta à injúria foliar. Considerando ambas as iniciativas, os transcritos gerados incluíram 453.952.833 transcritos de RNAseq (100 bp Illumina HiSeq TruSeq); 46.582.833 HT-SuperSAGE tags (26 bp, Solexa-Illumina); 298.119 SuperSAGE tags (26 bp, 454 Life Sciences/Roche); 32.084 LongSAGE tags (19-21 bp); e 314.765 ESTs (49,820) (JESUS-PIRES et al., 2019).

Em 2018 o mapa com 16 QTLs (*Quantitative trate locus*) e informação relacionada sobre SNP (*Single nucleotide polimorfism*) foi publicado e disponibilizado no Phytozome. Entre as 16 QTLs estão características como cor da flor, tamanho da folha, peso da semente, número de grãos na vagem, etc. (LO et al., 2018). Mas foi no início de 2019, que a primeira versão do sequenciamento de genoma completo do feijão-caupi (*Vigna unguiculata V.1.0*) foi publicada (LONARDI et al., 2019)., Disponível nos bancos de dados Phytozome (phytozome.jgi.doe.gov) e NCBI (https://www.ncbi.nlm.nih.gov), o genoma disponibilizado compreende 519.4 Mb de sequências referentes aos 11 cromossomos e um total de 29.773 loci gênicos, incluindo a caracterização de sequências repetitivas como elementos transponíveis e microssatélites (LONARDI et al., 2019).

Considerando os dados genômicos e após aplicação de um pipeline de anotação automatizado, LONARDI et al. (2019) estimaram que 49,5% do genoma do feijão-caupi seja composto dos seguintes elementos repetitivos: 39,2% de elementos transponíveis (TEs), 4% de SSRs (*Simple Sequence Repeat*) e 5,7% sequências não identificadas de baixa complexidade. Os retrotransposons, ou Classe I TEs, compreenderam 84,6% dos TEs por cobertura de sequência e 82,3% por número. Dos Retrotransposons (classe I), da ordem *LTR*, os elementos da superfamília *Gypsy* foram 1,5 vezes mais abundantes que os elementos *Copia*. Por sua vez, os

elementos *TRIM,* não autônomos, parecem ser muito raros, com apenas 57 encontrados. Os *LINEs* e os *SINEs*, compreendendo os retrotransposons *não LTR*, somaram apenas 0,4% do genoma. Os DNA-transposons (classe II), compreenderam 6,1% do genoma, sendo o *CACTA* (5,7%), o *hAT* (3,5%) e o *MuDR* (2,4%) os principais grupos da ordem *TIR*. A superfamília *Helitron* do tipo "*rolling-circle*" foi relativamente abundante, compreendendo 1,3% do genoma e 7013 elementos individuais. Apenas 6,4% das sequências TE não puderam ser classificadas.

Além do dado genômico, até o presente momento, o NCBI inclui sequências mais de 370.000 ESTs de *V. unguiculata* que abrangem diversos órgãos e condições e podem ser acessados livremente.

2.3 BIOINFORMÁTICA

2.3.1 Transcriptômica

2.3.1.1 Montagem: métodos e ferramentas

O transcriptoma representa o conjunto de transcritos (RNAs não estruturais) a partir dos genes contidos em uma célula ou tecido sob uma determinada condição (FINOTELLO; DI CAMILLO, 2015). Geralmente o transcriptoma possui baixo número de sequências repetitivas, o que facilita sua montagem (MARTIN; WANG, 2011). Tal abordagem tornou-se significativamente mais informativa com o advento do sequenciamento NGS (*Next Generation Sequencing*) que resultou na técnica de RNA-Seq a partir de RNAs mensageiros maduros, tratando-se de uma técnica muito informativa e aplicável também a organismos que não possuem genoma de referência (WANG; GERSTEIN; SNYDER, 2009). Tais análises permitem a montagem sem referência (denominada montagem '*de novo*'), além da montagem a partir de mapeamento no genoma (chamada de 'montagem por referência') (HAAS et al., 2013; JAIN, 2012).

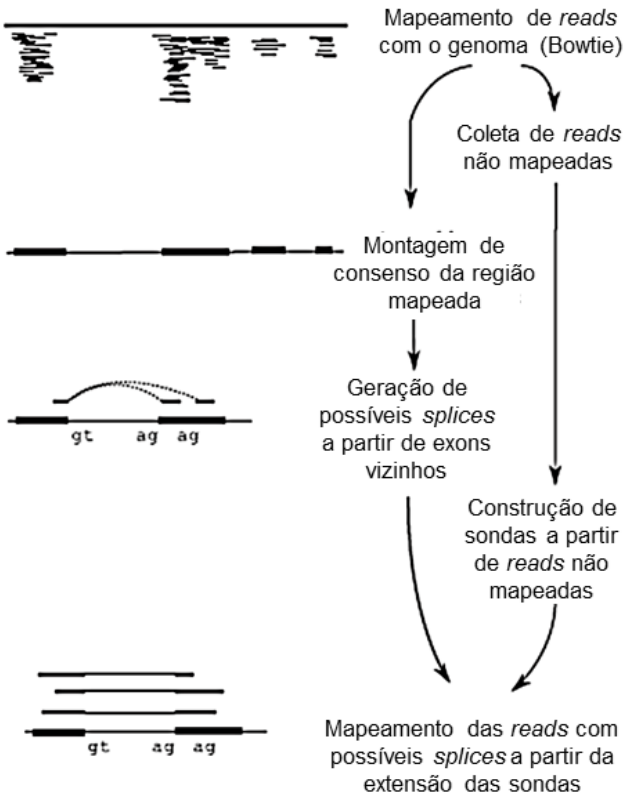
Diferentemente da montagem de genomas, o processamento do transcriptoma produz um transcrito para cada isoforma, como resultado, cada *locus* pode ser associado a mais de uma isoforma. Portanto, cada transcrito apresentará uma cobertura distinta, dependendo do seu nível de expressão, além disso a

expressão de um gene não pode ser deduzida a partir de apenas uma de suas isoformas (HAAS et al., 2013).

Para atingir bons resultados se faz necessária uma grande profundidade de sequenciamento, permitindo assim detectar transcritos e isoformas raras. As plataformas como a HiSeq da Illumina (JAIN, 2012) têm sido as mais recomendadas. Os fragmentos gerados possuem tamanhos que variam entre 35 e 500 pb (MARTIN; WANG, 2011). A plataforma da Illumina é a mais utilizada, gerando *reads* entre 25-250 pb (FABBRO et al., 2013). Independentemente da plataforma escolhida, a etapa mais desafiadora é a montagem, na qual adaptadores precisam ser eliminados, as *reads* precisam ser processadas e combinadas em *contigs*, que representam os transcritos (completos ou não), da forma mais fiel possível. Logo, para assegurar a qualidade, é essencial um pré-processamento seguido da montagem e acesso aos dados de expressão. Tais procedimentos geralmente ocorrem em cinco etapas: (i) controle de qualidade das sequências (trimagem); (ii) alinhamento com o genoma de referência (se houver); (iii) montagem propriamente dita (formação de *contigs*); (iv) normalização da contagem de reads; (v) análise de expressão diferencial. Fica sugerida como sexta etapa a (vi) anotação das sequências, essencial para a interpretação dos dados (JAIN, 2012; MARTIN; WANG, 2011).

Primeiramente, o controle de qualidade das sequências elimina as extremidades ou grupos de *reads* que não atingem a qualidade mínima (JAIN, 2012). Essa etapa é realizada por diversos programas, destacando-se Trimmomatic (BOLGER; LOHSE; USADEL, 2014) por também remover adaptadores e ser o único com possibilidade de paralelização (BOLGER; LOHSE; USADEL, 2014; FABBRO et al., 2013). Na segunda etapa (alinhamento com o genoma de referência), quando disponível, é necessário alinhar as *reads* antes da montagem. O TopHat (TRAPNELL; PACHTER; SALZBERG, 2009) faz o alinhamento das sequências de RNA-Seq contra o genoma, ao mesmo tempo que busca por sítios de *splicing*. A busca é realizada em duas etapas. Na primeira etapa é realizada a indexação das *reads* usando o Bowtie (LANGMEAD et al., 2009) e a geração de consensos. Em seguida, as *reads* não indexadas são usadas para estender as sequências e buscar por possíveis junções de *splicing* (Figura 3).

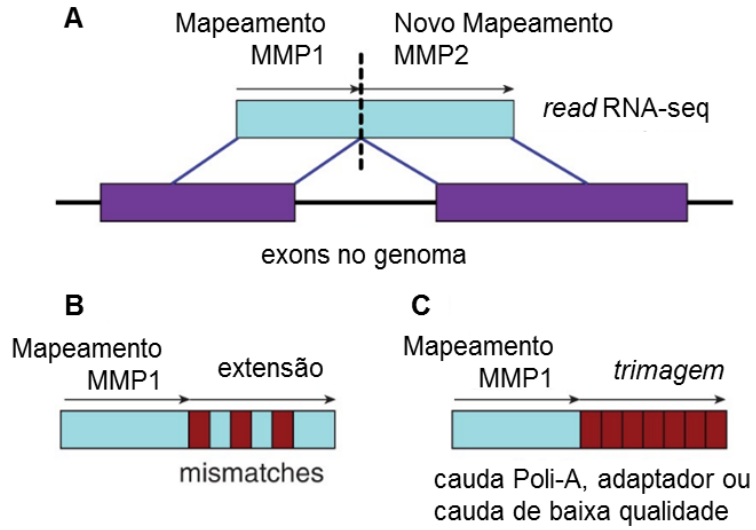
Figura 3 - TopHat pipeline.



Fonte: Adaptado de Tranell et al. (2009). Sequências de RNA-Seq são mapeadas contra o genoma de referência, as sequências não mapeadas são separadas em um grupo à parte. Das regiões mapeadas são feitas sequências consenso. As sequências não mapeadas são alinhadas e indexadas às regiões que apresentam potencial de junção de *splicing*.

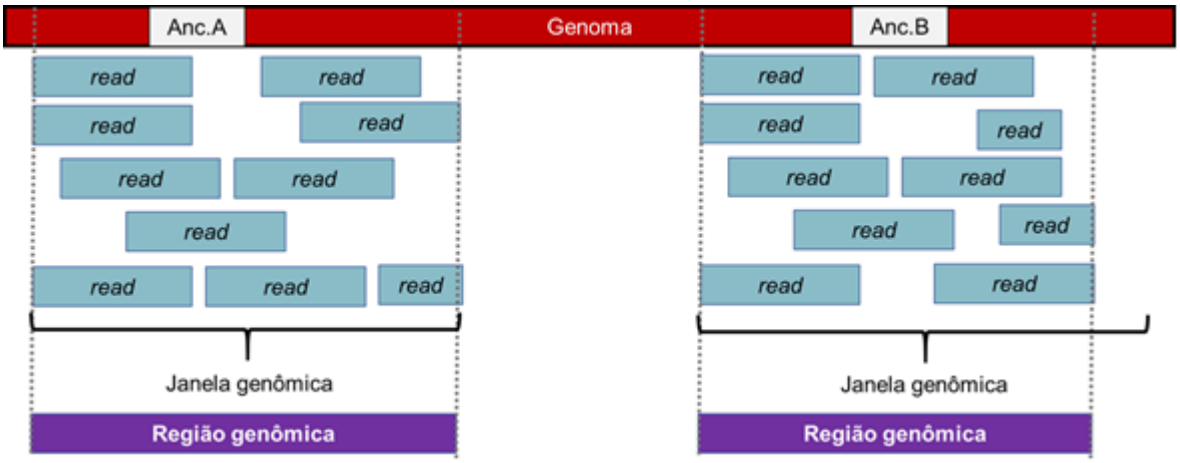
O programa STAR (DOBIN et al., 2013) também tem a proposta de alinhar *reads* contra o genoma de referência. É um programa mais rápido e mais preciso, com algoritmo que identifica transcritos quiméricos. O STAR funciona em duas etapas: (i) busca de sondas, que busca o MMP (*Maximal Mappable Prefix*) no qual as partes de uma *read* são alinhadas respeitando o tamanho máximo definido para um íntron, permitindo identificar as junções (Figura 4); (ii) agrupamento, unificação e penalidades, as sequências alinhadas na etapa anterior são separadas em grupos a partir de uma sequência “âncora” e unidas para definir a região genômica (*locus*) relacionada (Figura 5). Neste processo, são definidas pontuações de qualidade para os alinhamentos.

Figura 4 - STAR, representação esquemática da etapa I: busca de sonda.



Fonte: Adaptado de Dobin et al. (2013). A Representa o caso de alinhamento perfeito, na qual a *read* está representada em azul e seu alinhamento com o genoma em roxo (exons). A porção entre os exons representa o intron e permite identificar a junção de exons. B Caso onde a *read* a ser alinhada apresenta *mismatches*, acarretando em extensão do alinhamento. C Caso um adaptador, uma cauda poli-A ou sequência de baixa qualidade seja identificada, as mesmas serão trimadas.

Figura 5 - Representação esquemática da etapa II: agrupamento, unificação e penalidades.

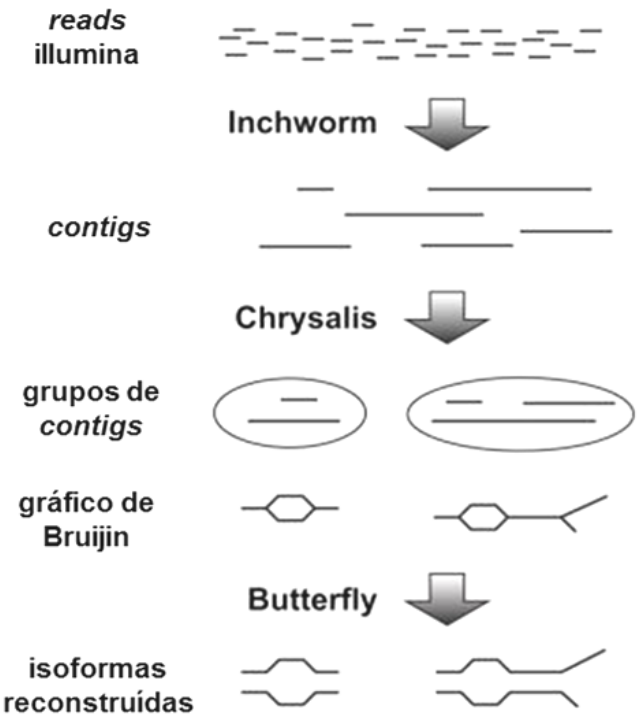


Fonte: Adaptado de Dobin et al. (2013). Em vermelho está representado o genoma, sobreposto por sequências “âncoras” (Anc). A partir da sequência âncora, as *reads* agrupadas serão alinhadas ao redor da sequência âncora, dentro da “janela” definida pelo usuário a fim de definir a sequência com base na região genômica.

A terceira etapa é a montagem de *contigs* propriamente dita, que ocorre a partir do mapeamento resultante (JAIN, 2012), quando disponível. Alternativamente, quando é feita uma montagem *de novo*, a etapa acima é ignorada. Para isso, estudos independentes têm defendido a ferramenta Trinity como altamente efetiva,

flexível e associada a uma comunidade ativa de desenvolvedores (HAAS et al., 2013). A ferramenta Trinity permite a montagem com genoma de referência e a montagem *de novo*, embora seja mais utilizada para esta última. A ferramenta possui parâmetros ajustáveis e ferramentas integradas para cálculo de expressão diferencial, consistindo em três etapas (Figura 6): (I) *Inchworm*, gera um *contig* dominante a partir do k-mer das *reads*; (II) *Chrysalis*, agrupa os *contigs* gerados formando um gráfico de Bruijin para cada agrupamento; (III) *Butterfly*, avalia as possibilidades do gráfico de Bruijin, definindo isoformas alternativas (GRABHERR et al., 2011; HAAS et al., 2013).

Figura 6 - Pipeline do Trinity.



Fonte: Adaptado de Grabherr et al. (2011). A figura mostra as três etapas de montagem. Inicialmente com o *Inchworm*, construindo os *contigs* a partir dos dados de sequenciamento Illumina. Em seguida, o *Chrysalis* agrupa os *contigs* e gera os gráficos de Bruijin para cada agrupamento. Por último, o *Butterfly* analisa individualmente cada uma das possibilidades dos gráficos de Bruijin.

A ferramenta *Cufflinks* (TRAPNELL et al., 2012) é atualmente a mais utilizada para montagem de RNA-Seq com genoma de referência (GHOSH; CHAN, 2016). Disponível no pacote Tuxedo, juntamente com o TopHat, a ferramenta parte da análise de um arquivo de alinhamento que pode ser gerado tanto pelo TopHat

quanto pelo STAR. Considerando as *reads* mapeadas o *cufflinks* monta os *contigs* a partir de uma abordagem parcimoniosa. Após essa etapa, com a estimativa de abundância feita pelo *cuffdiff*, os transcritos imaturos são automaticamente excluídos.

A quarta etapa do processo de montagem acessa os dados de RNA-Seq para normalização da contagem das *reads*. Visando representar a verdadeira abundância de cada transcrito, são consideradas a profundidade e a cobertura do sequenciamento por ferramentas estatísticas (JAIN, 2012). Um problema encontrado neste processo é que isoformas, bem como duplicações gênicas, vão mapear igualmente com múltiplos alvos. Para minimizar esta dificuldade, o método estatístico RSEM (*RNA-Seq by Expectation-Maximization*) é aplicado para calcular a probabilidade de cada *read* ser de fato derivada de um determinado transcrito (GRABHERR et al., 2011; JAIN, 2012).

Tanto o Trinity (GRABHERR et al., 2011; HAAS et al., 2013) quanto o pacote *Cufflinks* (TRAPNELL et al., 2012) possuem etapas de normalização que geram dados de FPKM (*Fragments per Kilobase Million*). No caso do Trinity é possível usar também o RSEM. Já o pacote *Cufflinks* inclui o programa *Cuffnorm*, que permite uma etapa extra de normalização, visando escalas de expressão mais acuradas.

A quinta etapa, análise da expressão diferencial, é o resultado que desperta maior interesse entre os cientistas, sendo indicado o uso de no mínimo três réplicas biológicas para cada amostra. Uma revisão recente comparou diversos programas voltados para avaliação da expressão diferencial, dentre eles destacaram-se o Edge-R e o DESeq como os mais eficientes (SCHURCH et al., 2016), ambos disponíveis no Trinity. Já o pacote *cufflinks*, dispõe do programa *cuffdiff*, que apresenta alta acurácia e seus resultados de expressão, considerando cada gene e cada isoforma, permitem também juntar isoformas pelo TSS (*transcript start site)* e distinguir isoformas geradas a partir de um mesmo pré-mRNA. Esta etapa será discutida em mais detalhes na próxima sessão (tópico 2.4.2.2).

Sugere-se como sexta etapa, a anotação funcional dos transcritos, que promove a identificação e tradução das ORFs (*Open Read Frame*) para anotação de função a partir de sequências de nucleotídeos e proteínas. Para tal, é realizada a tradução das sequências, com ferramentas como o ORFfinder (NCBI), TranSeq (EMBL) e o TransDecoder (https://github.com/TransDecoder). As duas primeiras são ferramentas online com limite de sequências, enquanto o TransDecoder é uma

ferramenta programática sem limite de sequências para análise. Esta última adota como parâmetros, além do tamanho mínimo de ORF de 100 nt, um score de similaridade considerando homologia com bancos do BLAST ou do PFAM (banco de dados de proteínas curadas, representadas por alinhamentos múltiplos e modelos ocultos de Markov ou HMM, *Hidden Markov Models*). Em seguida são realizadas anotações funcionais automáticas com o Trinotate (http://trinotate.github.io/), as quais são desenvolvidas especialmente para montagens transcriptômicas '*de novo'* (HAAS et al., 2013).

Como alternativa de automatização, há o conjunto de *pipelines* MUGQIC (BOURGEY et al., 2019) desenvolvido pela Universidade McGill e pelo *Genome Quebec Innovation Center*. Em seu site estão disponíveis opções de montagem '*de novo'* ou com referência para dados de RNA-Seq, além de abordagens de DNA-Seq e ChIP-seq entre outras. Todos os *pipelines* são escritos em linguagem de programação Python 2.7. Para montagem de RNA-Seq '*de novo'* o *pipeline* baseado no Trinity sumariza todas as etapas de montagem e anotação. Complementarmente, são feitas análises exploratórias, para acessar a homogeneidade dos dados, efeitos de variáveis de experimento bem como potenciais erros de rotulação. Todas as informações e acesso podem sem encontradas no site Mugqic–GenPipe: https://bitbucket.org/mugqic/genpipes.

2.3.1.2 Análise de expressão diferencial via RNA-Seq

A análise de RNA-Seq oferece a oportunidade de estudar a genômica funcional de espécies que possuem menor investimento, abrangendo assim organismos não modelo, como aqueles de importância ecológica e evolutiva (HAAS et al., 2013). Adicionalmente, seu alto rendimento e acurácia na avaliação da expressão são considerados superiores aos da técnica de *microarray* (JAIN, 2012; WANG; GERSTEIN; SNYDER, 2009), oferecendo vantagens como abordagens *de novo*, grande nível de reprodutibilidade, além de não ser limitada por saturação de sinal (FINOTELLO; DI CAMILLO, 2015). Consequentemente, o método de RNA-Seq é a principal escolha para estudos de expressão gênica, podendo ser usada para capturar a dinâmica em diferentes tecidos e em diferentes condições (JAIN, 2012; WANG; GERSTEIN; SNYDER, 2009). A referida análise tornou possível a geração de uma visão global sem precedentes do transcriptoma e sua organização (WANG;

GERSTEIN; SNYDER, 2009), especialmente por permitir avaliar um conjunto representativo de RNAs mensageiros ou de RNAs não codificantes (HRDLICKOVA, TOLOUE; TIAN, 2017). Recentemente, um estudo com 48 réplicas biológicas avaliou o desempenho de 11 programas de cálculo da expressão diferencial. Dentre estes, cinco programas mostraram excelente performance: EBSeq, edgeR, DESeq, DESeq2 e *limma*. O autor ainda reporta que edgeR e DESeq são os mais usados na literatura, e discute que, entre as ferramentas avaliadas com baixo número de réplicas (n ≤ 12), edgeR e DESeq2 foram os que obtiveram melhores resultados, sendo o edgeR o que melhor controla a taxas de falsas descobertas (SCHURCH et al., 2016). Quanto melhor e mais segura a estatística aplicada nesta etapa, mais coerentes e consistentes são os resultados quando comparados nos testes de validação.

Diversas são as áreas de aplicação para estudos de transcriptoma e expressão diferencial e tem permitido a descoberta de mecanismos de determinação de sexo em plantas (JAIN, 2012); identificação de diferenças de resposta a estresses (HÜMBNER; KOROL; SCHMID, 2015; KAKUMANU et al., 2012; KUNDU et al., 2017); estudos de SNP (*single nucleotide polymorphism*) e de transcritos alternativos (ZENONI et al., 2010).

Complementarmente, a técnica de RNA-Seq permitiu o enriquecimento do atlas integrado de expressão para humanos, animais e plantas (PETRYSZAK et al., 2016), além da geração de atlas transcricionais (STELPFLUG et al., 2015; YAO et al., 2016). Em humanos, a análise de transcriptoma completo e avaliação da expressão diferencial tornou-se ferramenta fundamental na compreensão funcional sobre a caracterização genética de diversas doenças. Com grande impacto nas pesquisas sobre câncer (COSTA et al., 2013; NG et al., 2013).

2.3.2 Mineração de texto e Biologia de sistemas

2.3.2.1 Mineração de texto: importância e aplicação

De acordo com Andrade e Bork (2000), ainda no século 19 a informação científica era divulgada entre colegas de trabalho apenas e a partir de cartas, somente ao final do século 19 os jornais científicos foram criados tomando para si

esta responsabilidade e profissionalizando as publicações e descobertas científicas. Foi então nos anos 1960, com o intuito de facilitar as buscas, que foram criadas listas de publicações divulgadas mensalmente e indexadas por palavras-chave. Em seguida, foi iniciada a transição para a mídia eletrônica. Nos anos 1980, a revista Current Contents inovou com o lançamento da “CC on diskette”, tornando as publicações científicas acessíveis em computadores (ANDRADE; BORK, 2000).

Com o advento da internet, as possibilidades de armazenamento e divulgação foram revolucionadas especialmente pela criação dos bancos de artigos, sendo o principal o MEDLINE, com artigos compilados desde 1966. Andrade e Bork (2000) afirmam em seu artigo: “Com o aumento da distribuição das revistas no formato eletrônico, o acesso a textos completos de qualquer artigo estará à distância de um clique”, e completa dizendo que se faz necessário o uso de técnicas de seleção de artigos mais relevantes, bem como de extração automática da informação com métodos que podem ser baseados em dicionários, amostras pré-selecionadas e uso de NLP (*Natural Language Processing*).

Após apenas 19 anos, já é um fato consolidado pela comunidade científica a impossibilidade humana de atualização completa sobre um tema através de meios convencionais. Com mais de cinco mil jornais indexados até novembro de 2017, o MEDLINE representa o componente primário do PubMed, banco de artigos do NCBI.

Atualmente, de acordo com Sayers et al. (2019), a taxa de crescimento do banco de artigos científicos completos é de 11,9% ao ano, além de dados publicados na forma de livros e resumos. Até junho de 2017, o MEDLINE já apresentava mais de 31 milhões de resumos de artigos disponíveis. Para os profissionais que devem estar constantemente se atualizando, tornou-se necessário, quase imprescindível, o uso das novas técnicas de mineração de texto para que ocorra a integração de toda a informação.

As ferramentas de mineração de texto são compostas por técnicas computacionais que envolvem três etapas gerais de processamento, descritas a seguir de acordo com Krallinger e Valencia (2005): (I) conhecida como IR (*Information retrieval*), baseia-se em encontrar e selecionar artigos; (II) identificação de entidades biológicas ou “tagueamento”, e (III) extração de termos de associações. Para a etapa de IR, a busca pode se dar por combinação de palavras-chave ou por busca baseada em documento, na qual um artigo ou lista de artigos são usados como “*queries*” (KRALLINGER; VALENCIA, 2005).

O “tagueamento” é uma etapa extremamente complicada devido à falta de padronização na nomenclatura de proteínas e genes, por apresentarem variações tipográficas, citação com símbolos, além de termos sinônimos e ambiguidade de nomenclatura (KRALLINGER; VALENCIA, 2005; MIKA; ROST, 2004). A busca é dificultada ainda mais, pois algumas proteínas e genes possuem nomes compostos por nomes não científicos da língua inglesa, como por exemplo *white, wing* e *bizarre*. Além disso, existe a necessidade de distinguir proteínas de mesmo nome em organismos diferentes. Outro complicador desta etapa é o constante surgimento de novos nomes, tornando ineficiente a abordagem apenas baseada em dicionários (MIKA; ROST, 2004).

Mika e Rost (2004) afirmam que o primeiro programa a usar aprendizado de máquina (*machine learning*) para mineração de texto foi o NLProt, uma ferramenta para busca por nomes de proteínas. O NLProt identifica e classifica informações textuais biológicas e não biológicas, fazendo uso de contextualização a partir de dois dicionários. Um dicionário de proteínas e outro dicionário contendo palavras não científicas da língua inglesa. O segundo foi expandido com termos médicos, nomes de espécies, tipos de tecido e fórmulas minerais. O programa funciona com versão online (com restrição de 50.000 caracteres) ou comando de linhas (nenhuma restrição, na versão LINUX).

Diversas ferramentas integram as três etapas de mineração de texto como STRING (JENSEN et al., 2009; SNEL et al., 2000; SZKLARCZYK et al., 2017), iHOP (FERNÁNDEZ; HOFFMANN; VALENCIA, 2007) e PESCADOR (BARBOSA-SILVA et al., 2011), entre outras.

O STRING é uma ferramenta criada em 2000 com o intuito de encontrar associações funcionais entre proteínas. Iniciado a partir de um gene ou de seu ortólogo identificados por BLAST, o programa busca por coocorrências e em seguida por interações (SNEL et al., 2000). Em uma de suas atualizações, o programa foi incrementado com filtros de mineração de texto (JENSEN et al., 2009). Ainda mais recente, em 2017, foram adicionadas além da ferramenta online, funções programáticas (SZKLARCZYK et al., 2017).

O iHOP funciona a partir de palavra-chave, como nome do gene ou da proteína, fazendo uma busca e seleção de artigos. Os mesmos recebem um *score* baseado no impacto de sua publicação, significância, data de publicação bem como sintaxe. O iHOP tenta diminuir problemas de ambiguidade a partir de métodos

heurísticos e seu resultado é retornado na forma de uma descrição resumida das interações encontradas (FERNÁNDEZ; HOFFMANN; VALENCIA , 2007).

O PESCADOR, é uma ferramenta online que permite a utilização de um banco selecionado de artigos para a extração de interações, além de buscar por novos artigos relevantes a partir do MedlineRanker. Diferente do iHOP e do STRING, ele possui filtros (BARBOSA-SILVA et al., 2011), embora a atualização do STRING tenha adicionado esta usabilidade em 2017 (SZKLARCZYK et al., 2017). O PESCADOR tem como resultado interações que podem ser observadas graficamente (BARBOSA-SILVA et al., 2011). Após gerar uma lista de artigos com o tema ou a entidade taxonômica (ou ambos) buscados, os abstracts são submetidos à segunda e à terceira etapa de mineração de texto. Com o processamento dos artigos, as interações são classificadas em quatro tipos distintos (BARBOSA-SILVA et al., 2010):

a) Nomes de proteínas na mesma sentença, com um termo de biointeração no meio, e identificação de conceitos;
b) Nomes de proteínas com termo de biointeração em qualquer posição na sentença, e identificação de conceitos;
c) Nomes de proteínas na mesma sentença, sem que seja necessária a identificação de uma biointeração;
d) Todos os termos do *abstract* são levados em consideração, independentemente de estarem na mesma sentença. Nesse caso, todas as interações acima também são analisadas.

Com o intuito de fornecer informação integrada, sucinta e de fácil compreensão as ferramentas de mineração de texto apresentam grande potencial para impulsionar o estudo da biologia de sistemas, através da identificação de termos biológicos tais como DNA, RNA e principalmente proteínas e suas interações descritas em um ou mais artigos.

2.3.2.2 Biologia de sistema e redes de interação

Para entender a biologia dentro de toda sua complexidade, se faz necessário o estudo e a compreensão integrada de dados biológicos envolvendo atores moleculares, estrutura, função e dinâmica em nível celular, de tecido e de

organismo. Com a evolução das técnicas de sequenciamento genômico e transcriptômico ficou mais fácil coletar dados que agreguem este tipo de informação, impulsionando assim, o estudo da biologia de sistemas (KITANO, 2002b), principalmente quando integrados a dados de proteômica, interferômica e interatoma, entre outros (BENKO-ISEPPON et al., 2017).

A biologia de sistemas visa mudar o ponto focal de estudo, das proteínas e genes em sua unidade para entender a sua dinâmica. No entanto, é crucial a realização de um estudo multidimensional. Na pesquisa, o desenho exaustivo de redes de interação é que impulsiona o primeiro passo dentro do processo de estudo da biologia de sistemas. Assim, será viabilizado o conhecimento de como módulos de interação interferem uns sobre os outros (KITANO, 2002b). A biologia de sistemas é em sua maioria baseada em estudos moleculares e genéticos que envolvem as “ômicas”. Entre seus principais desafios estão a vasta quantidade de dados dispersos, o que se torna ainda mais desafiador na era de ‘*Big Data*’, com heterogeneidade dos dados e variedade de formatos, identificadores e organização (GHOSH et al., 2011; PAVLOPOULOS et al., 2015).

Neste contexto, a bioinformática desempenha papel crucial, com o manejo sofisticado dos dados e análises de integração dos dados, tornando possível a predição de comportamentos biológicos diante de diferentes condições e tempos. Consequentemente, a compreensão dos mecanismos moleculares e de suas respostas será aprimorada (GHOSH et al., 2011).

A publicação científica é o canal-chave da distribuição da informação na ciência (REBHOLZ-SCHUHMANN; OELLRICH; HOEHNDORF, 2012) e (como descrito em maior detalhe no tópico anterior) a mineração de texto busca integrar toda a informação dispersa na literatura a partir da extração de informações relacionadas a genes e proteínas no contexto de cada organismo, célula, tecido específico ou estágio de desenvolvimento (BARBOSA-SILVA et al., 2010, 2011; REBHOLZ-SCHUHMANN; OELLRICH; HOEHNDORF, 2012). Por isso, essa abordagem torna-se parte da biologia de sistemas.

Para a construção de conhecimento na visão da biologia de sistemas deve-se realizar uma abordagem *multilayer*, desta forma auxiliando a responder perguntas como: Quais circuitos são conservados evolutivamente? Quais funções e por que são conservados? Qual a sua relação com evolução gênica? Desta forma, são necessários estudos com abordagens variadas (GHOSH et al., 2011; KITANO,

2002a), tornando-se crucial a disponibilidade de repositórios para compartilhamento e acesso aos interatomas, de forma que os dados se tornem cada vez mais integrados e completos (KITANO, 2002a). Dito isso, são diversas as iniciativas do estudo da biologia de sistemas, nos mais variados campos e com diversas abordagens experimentais ou *in silico* (GRECO et al., 2012; VANHOLME et al., 2012; VERDIER et al., 2013; ZHU et al., 2013).

Um exemplo bastante completo é o PD-map, que representa a combinação de dados experimentais, bancos de dados e mineração de texto para gerar o interatoma da doença de Parkinson (PD). O PD-map é uma rede de interação completa, que inclui a interações metabólicas, regulação gênica e processos de sinalização, bem como pontos de interação medicamentosa. Trata-se de um grande exemplo do que se pode alcançar dentro da biologia de sistemas, com dados de disfunção mitocondrial e sináptica, neuroinflamação entre outros. Este mapa representa uma valiosa fonte de dados gerados a partir de estudos no contexto da doença de Parkinson (FUJITA et al., 2014).

3 OBJETIVOS

3.1 OBJETIVO GERAL

Usar duas abordagens, uma generalista e outra específica, no estudo sobre o papel de TEs na resposta transcricional de plantas sob estresse. Na primeira, usar a mineração de texto para avaliar dados da literatura científica. Em seguida, acessar dados de RNA-Seq de *Vigna unguiculata* para identificar elementos transponíveis diferencialmente expressos e sua possível relação com a resposta a estresses.

3.2 OBJETIVOS ESPECÍFICOS

a) Avaliar o conjunto de bioentidades relacionadas a resposta a estresse biótico que sejam super- e sub- representadas na literatura científica;
b) Usar a mineração de texto no desenho de uma via de interação de resposta a estresse biótico;
c) Testar diferentes formas de identificação da ação de TEs em plantas sob estresse, incluindo mineração de texto;
d) Determinar qualitativa e quantitativamente TEs diferencialmente expressos em diferentes estresses (desidratação radicular e inoculação viral) e tecidos (raízes e folhas);
e) Determinar quais superfamílias de TEs possuem interação com genes responsivos a estresse, avaliando sua expressão. Determinar a distribuição dos transcritos de TEs nos pseudocromossomos do feijão-caupi, inferindo sobre sua distribuição e mecanismos de expansão;

4 ARTIGO 1 - EXPRESSÃO DIFERENCIAL DE ELEMENTOS TRANSPONÍVEIS NO TRANSCRIPTOMA (RNA-SEQ) DO FEIJÃO-CAUPI EM SITUAÇÕES DE ESTRESSE

Bruna Piereck[1], David Anderson de Lima Morais[2], João Pacífico Bezerra-Neto[1], Gabriela Frosi[3], Guillaume Bourque[4], Ana Christina Brasileiro-Vidal[1] e Ana Maria Benko-Iseppon[1]

[1]Departamento de Genética, Laboratório de Genética e Biologia Vegetal (LGBV), Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil.
[2]Centre de Calcul Scientifique (CCS), Université de Sherbrook, Sherbrook, Canadá.
[3]Departamento de Botânica, Laboratório de Fisiologia Vegetal (LFV), Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil.
[4]Genome Quebec Innovation Center, McGill University, Montreal, Canadá.

Mobile DNA

Expressão diferencial de elementos transponíveis no transcriptoma (RNA-Seq) do feijão-caupi em situações de estresse

Bruna Piereck[1], David Anderson de Lima Morais[2], João Pacífico Bezerra-Neto[1], Gabriela Frosi[3], Guillaume Bourque[4], Ana Christina Brasileiro-Vidal[1] e Ana Maria Benko-Iseppon[1]

[1]Departamento de Genética, Laboratório de Genética e Biologia Vegetal (LGBV), Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil.
[2]Centre de Calcul Scientifique (CCS), Université de Sherbrook, Sherbrook, Canadá.

[3]Departamento de Botânica, Laboratório de Fisiologia Vegetal (LFV), Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil.
[4]Genome Quebec Innovation Center, McGill University, Montreal, Canadá.

E-mails: piereck.bruna@gmail.com, david.a.morais@gmail.com, pacifico.joao@gmail.com, gabriella.frosi@gmail.com, guil.bourque@mcgill.ca, brasileirovidal.ac@gmail.com, ana.iseppon@gmail.com

*** Correspondência:**

Profa. Dra. Ana Maria Benko Iseppon

ana.iseppon@gmail.com

Resumo

Introdução: As plantas são frequentemente expostas a diversos estresses que causam impacto na produtividade agrícola. Para resistir aos impactos causados, precisam ativar diferentes mecanismos que dependem da regulação transcricional. Dentre os transcritos que sofrem regulação, estão os elementos transponíveis (TEs) que, em decorrência de sua natureza móvel, podem gerar mudanças estruturais e epigenéticas. No presente trabalho, a modulação da expressão de TEs foi avaliada no transcriptoma do feijão-caupi (*Vigna unguiculata*), uma leguminosa de importância socioeconômica, que a despeito da existência de genótipos adaptados a condições adversas, não tem atingido seu potencial produtivo devido a estresses bióticos (vírus) e abióticos (seca).

Resultados: Ao analisar o transcriptoma (RNA-Seq) de feijão-caupi sob dois estresses bióticos (C*owpea Severe Mosaic Virus*, CPSMV; *Cowpea Aphid-Born Mosaic Virus,* CABMV) e um abiótico (desidratação radicular, RD) foram identificadas 12 superfamílias de TEs diferencialmente expressas (TE-DEGs). Os elementos *Copia* e *hAT* são os mais representativos e estão relacionados a domínios de resposta a estresse, apresentando maior potencial para estudos de validação. Além disso, foram observados aproximadamente 500 TE-DEGs de classificação desconhecida, sendo. Parte (23-30%) dos transcritos apresentou similaridade significativa com mais de 10 elementos (geralmente da mesma superfamília), o que pode indicar elevado número de cópias ou atividade recente nesta leguminosa. A anotação funcional baseada em ontologia gênica revelou para os três acessos a modulação de TEs relacionados à resposta a estresse.

Conclusão: O total de TEs diferencialmente expressos dos acessos inoculados com vírus tende a diminuir ao longo do tempo e apresentam expressão estresse, tempo e acesso específica. Por outro lado, no acesso sob desidratação, o total de TEs modulados aumenta ao longo do tempo, seguido de uma maior proporção de elementos reprimidos. A quantidade, de transcritos (23-30%) que apresenta similaridade com dezenas de TEs da mesma superfamília é um indício de conservação entre elementos, o que deve ser avaliado com análises complementares.

Palavras-chave: *Cowpea Severe Mosaic Virus* (CPSMV); *Cowpea Aphid-Born Mosaic Virus* (CABMV); desidratação radicular, Fabaceae.

Tipo de artigo: Research Article

1. Introdução

As plantas são frequentemente expostas a diversos tipos de estresse que podem afetar seu crescimento e desenvolvimento, resultando em impactos na sua produtividade. Para resistir ou tolerar estresses bióticos e abióticos, os vegetais podem ativar diferentes mecanismos fisiológicos [1-3], os quais envolvem cascatas de transdução de sinais, que alteram o conteúdo de transcritos, proteínas e metabólitos. Entre os transcritos que sofrem regulação sob estresse, estão os elementos transponíveis (TEs) que, assim como os genes, também podem ser modulados por fatores de transcrição (TFs) [4-7].

Os TEs compreendem sequências móveis de DNA classificadas de acordo com seu intermediário no processo de transposição, que pode ser uma molécula de RNA (classe I, Retrotransposons) ou de DNA (classe II, DNA-transposon). Em plantas, os TEs estão distribuídos em oito e sete superfamílias nas classes I e II, respectivamente [8]. Apesar de apresentarem estruturas variadas, as duas classes possuem elementos CIS que mimetizam promotores, acarretando na especificidade de expressão de alguns elementos [9-11]. Adicionalmente, durante o processo de movimentação no genoma, ambas as classes de TEs podem capturar fragmentos ou até sequências gênicas inteiras. Assim, quimeras (genes associados aos TEs) geradas podem ser cooptadas e neofuncionalizadas interferindo no padrão de expressão gênica [12-14].

Em angiospermas, os elementos móveis estão frequentemente associados à captura de genes de resposta ao estresse, considerando-se que podem contribuir ativamente na resposta adaptativa regulando genes do hospedeiro, como observado para *Boea hygrometrica* e *A. thaliana* onde o gene que confere tolerância à seca e um gene responsivo à infecção por *Fusarium* sp., respectivamente, são regulados por promotores de TEs. Além de atuarem mediando a regulação epigenética, a partir da geração de pequenos RNAs de interferência (siRNA, *small interfering RNA*) e da indução da metilação [15-21]. Contudo, estudos que avaliam dados de RNA-Seq quanto à expressão diferencial de TEs sob estresse em plantas ainda são escassos.

O genoma completo do feijão-caupi [*Vigna unguiculata* (L.) Walp.] foi recentemente publicado. Compreende 39,2% de TEs, os quais são os principais

responsáveis pelo seu maior genoma (11-12%) comparativamente a *V. radiata* e *V. angularis*. Dentre os TEs identificados no feijão-caupi, os elementos *Gypsy* contribuíram fortemente na expansão do genoma de *V. unguiculata* que pode ter ocorrido devido a amplificações recentes ou retenção de elementos antigos [27]. No presente trabalho, a modulação da expressão de TEs foi avaliada no transcriptoma do feijão-caupi, uma leguminosa importante para a agricultura de subsistência e a geração de empregos em regiões semiáridas e tropicais, sendo importante fonte de proteína para populações da África e da América do Sul. A despeito de sua importância e da existência de genótipos adaptados a condições adversas, o feijão-caupi não tem atingido todo seu potencial de produtividade [22].

Especificamente na América do Sul, as viroses se destacam como fatores bióticos limitantes da agricultura, com destaque para dois vírus de RNA: o *Cowpea Severe Mosaic Virus* (CPSMV) e o *Cowpea Aphid-Born Mosaic Virus* (CABMV), ambos transmitidos por insetos. Plantas acometidas apresentam redução do limbo foliar, distorção de brotos e folhas, bem como subdesenvolvimento da planta, ocasionando grandes perdas na produção [24, 26]. Por sua vez, a seca é um dos principais fatores que limitam a produtividade em regiões semiáridas do Brasil e da região Norte da África, sendo muito comum nas principais áreas de cultivo, onde causa perdas drásticas caso ocorra no período de floração [25,26].

Considerando-se a modulação transcricional de TEs sob condições de estresse, verificou-se a necessidade de avaliar o conteúdo gênico capturado por TEs, bem como a sua expressão diferencial nos transcriptomas gerados pelo ‘Cowpea Genomics Consortium’ (CpGC) que incluem três genótipos de feijão-caupi submetidos a estresses bióticos (folhas inoculadas com CPSMV e CABMV) e abiótico (desidratação radicular; RD) [28]. Considera-se que a disponibilidade de transcritos dos três tratamentos é uma excelente oportunidade para avaliação das principais superfamílias de TEs moduladas em feijão-caupi, do comportamento (induzido ou reprimido) destes nos diferentes estresses, bem como para fazer a primeira análise funcional de TEs nessa leguminosa cultivada, contribuindo para o conhecimento associado à ativação dessa categoria molecular sob estresse em Fabaceae.

2. Resultados

2.1 Métricas de qualidade e montagem do transcriptoma

Um total de 1.206.283.236 reads foram geradas pelos três experimentos (ver Metodologia). Após trimagem, a taxa de *reads* sobreviventes variou entre 99,9% e 100%. O mapeamento contra o genoma de referência (*Vigna unguiculata* v1.1) viabilizou o alinhamento de 93,11% das reads, entre as quais 86,19% em regiões únicas e apenas 13,8% com duplicações (não relacionadas ao RNAr). A taxa de reads alinhadas à região de exons foi de 64% (IT85F-2687/CPSMV), 65,5% (BR14-Mulato/CABMV) e 72,1% (Pingo de Ouro/RD), com uma média de 29% e 3,8% de reads mapeadas em regiões de íntrons e entre genes, respectivamente (Anexo B, Material Suplementar 1 e 2).

A análise de componente principal (PCA) mostrou para IT85F-2687/CPSMV e BR14-Mulato/CABMV, que os acessos resistentes apresentaram resposta distinta. Apesar da proximidade entre tratamento e controle, que pode indicar que as réplicas estão se agrupando em relação ao horário de coleta, ainda é possível observar genes de resposta a estresse regulados diferencialmente (Figura 1A). Para os dados de Pingo de Ouro, tratamento e controles, bem como duração do estresse apresentaram-se distintos (Figura 1B).

Em relação aos dados gerais da montagem, foram gerados 21.702 transcritos únicos, sendo 8.553 compartilhados pelos três acessos, enquanto 1.108 (IT85F-2687/CPSMV), 1.058 (BR14-Mulato/CABMV) e 10.983 (Pingo de Ouro/RD) foram exclusivos para os acessos citados. Entre os transcritos diferencialmente expressos (DEGs, *Differentially Expressed Transcripts*) foram identificados para cada acesso 1.665 (IT85F-2687/CPSMV), 1.689 (BR14-Mulato/CABMV) e 6.591 (Pingo de Ouro/RD), sendo 634, 623 e 5.386 exclusivos respectivamente. Comparando a expressão por órgão (folha/raiz), a folha dos acessos inoculados apresentaram uma maior quantidade de DEGs exclusivos no tempo precoce, a qual tendeu a diminuir após 16 h de estresse (Figura 2A e 2B). O contrário foi observado na desidratação radicular, comparando-se os tempos de 25 min (menor indução) e 150 min (maior indução) (2C e 2D).

2.2 Elementos transponíveis (TEs) expressos no genoma de *V. unguiculata*

2.2.1 Identificação de elementos transponíveis diferencialmente expressos (TE-DEGs)

Para a identificação de TEs os 21.702 DEGs do genoma foram analisados via RepeatMasker (RM). Para minimizar o número de elementos não classificados disponíveis nos dados genômicos, foi realizada uma anotação dos TEs do genoma contra o banco de Eudicotiledôneas do RepBase Update, resultando na reanotação de 354 (1,42%) de 24.876 sequências (Anexo B, Material Suplementar 3). Utilizando o banco contendo anotações atualizadas, foi feita a busca por TEs no transcriptoma de *V. unguiculata* com todas as cópias de TEs do genoma.

Entre os transcritos de TEs identificados, foi observado que alguns apresentavam características de mais de uma superfamília, todas respeitando os critérios determinados (ver Metodologia). Este tipo de ocorrência indica uma estruturação de TEs, que se encontram inseridos em outros TEs, sendo, portanto, denominados *nested*-TEs. O número de *matches* da mesma superfamília identificados pelo RM em determinados transcritos indica um provável nível de conservação, que pode ser devido a transposição/excisão recente, embora análises complementares sejam necessárias. Apenas quando os *matches* são referentes a superfamílias diferentes, considerou-se que os transcritos expressam *nested*-TEs.

Identificação de TE-DEGs em *V. unguiculata* (IT85F-2687) após inoculação com CPSMV

No tempo inicial de 1 h (CPSMV-1h) após imposição do estresse, amostras das folhas de IT85F-2687 expressaram diferencialmente 113 TE-DEGs (10% do total de DEGs para este tempo), as quais apresentaram similaridade com 711 cópias de TEs (Anexo B, Material Suplementar 4). Dentre os transcritos, 50 apresentaram apenas TEs *Unk* (*Unknown*, desconhecidos), enquanto os demais foram classificados em Retrotransposons (24), DNA-transposons (14) e *nested*-

TEs (25) (Tabela 1). TEs que continham elementos *Unk* e pelo menos um elemento anotado foram classificados de acordo com o elemento conhecido.

Após 1 h de estresse, foram modulados entre os Retrotransposons dois transcritos da ordem *LINE* (superfamília L1 e Tad1) e 22 *LTRs* (12 *Copia* e 10 *Gypsy)*. Entre os transcritos da classe II que foram modulados, um elemento da ordem *Crypton* (superfamília de mesmo nome) e 13 da ordem TIR (*CACTA, hAT e PIF-harbinger*) foram anotados (Tabela 1).

Após 16 h o total de TE-DEGs diminuiu para 67 (9,6% das DEGs para esse tratamento). Apesar do número reduzido de TE-DEGs , a proporção de retrotransposons reprimidos ficou maior, com 1/3 reprimido no tempo precoce e pouco mais da metade reprimido após 16 h. Elementos da ordem *LTR* somaram 22 transcritos (14 *Copia* e 8 *Gypsy*), a ordem TIR (*CACTA, hAT, MULE e PIF-harbinger*) apresentou apenas seis transcritos e os *nested*-TEs acumularam 15. Após 16 h de estresse os elementos *LINE* e *Crypton* não foram mais encontrados, indicando uma provável especificidade destes no tempo precoce, sob estímulo de inoculação viral, dado que o mesmo ocorreu para a cultivar BR14-Mulato também sob estresse biótico por inoculação de vírus de RNA.

Tabela 1 - Distribuição de elementos transponíveis diferencialmente expressos (TE-DEGs) por superfamília no acesso IT85F-2687 de *Vigna unguiculata* após inoculação com CPSMV (Cowpea Severe Mosaic Virus).

		Classificação	TE-DEGs 1 h	TE-DEGs 16 h	Tempo 1: 1 h			Tempo 2: 16 h			
					Up	Down		Up		Down	
					T	D	T	D	T	D	T
Retrotransposon		Superfamília desconhecida	-	--	-	-	-	-	-	-	-
	LTR	Copia	7	12	3	4	4	4	16	8	9
		Copia + *Unk*	5	2	12	1	6	1	9	1	4
		Gypsy	7	3	13	1	7	3	8	-	-
		Gypsy + *Unk*	3	5	17	2	16	1	14	4	29
		Retrovirus	-	-	-	-	-	-	-	-	-
		Retrovirus + *Unk*	-	-	-	-	-	-	-	-	-
	LINE	L1	1	-	1	-	-	-	-	-	-
		L1 + *Unk*	-	-	-	-	-	-	-	-	-

		Tad1	1	-		1	-	-	-	-	-	-
		Tad1 + *Unk*										
		CR1 + *Unk*	-	-		-	-	-	-	-	-	-
DNA-transposon	**TIR**	**CACTA**	1	1	1	1	-	-	-	-	1	8
		CACTA + *Unk*	3	3	1	7	2	12	2	28	1	35
		hAT	5	1	3	3	2	2	-	-	1	1
		hAT + *Unk*	3	1	2	15	1	10	1	4	-	-
		MULE	-	2	-	-	-	-	1	1	1	1
		MULE + *Unk*	-	-	-	-	-	-	-	-	-	-
		PIF	-	1	-	-	-	-	1	1	-	-
		PIF + *Unk*	1	1	-	-	1	14	1	3	-	-
		TC-mariner	-	-	-	-	-	-	-	-	-	-
		TC-mariner + *Unk*	-	-	-	-	-	-	-	-	-	-
		Merlin	-	-	-	-			-	-	-	-
		Crypton	1	-	1	1	-	-	-	-	-	-
		Unk	50	20	3	79	9	53	1	39	7	1
		***nested*-TEs**	25	15	6	19		15	1	86		3
		Total	**13**	**67**		**711**				**564**		
				DEGs				Transposons				

Fonte: A autora (2019). Número de TE-DEGs (D) e de cópias de TEs (T) identificados após 1 e 16 h de estresse biótico (vírus). Os dados estão agrupados por classe, ordem e superfamília. Um único transcrito pode apresentar mais de uma estrutura de TE. Quando um mesmo transcrito apresentou um TE desconhecido (*Unk*) e uma superfamília conhecida (superfamília + *Unk*), ele foi classificado de acordo com este segundo. Transcritos que apresentaram mais de um TE de superfamílias distintas são chamados *nested*-TEs e estão destacados separadamente. A distribuição dos transcritos é distinguida em induzidos (UP) e reprimidos (DOWN).

Identificação de TEs em *V. unguiculata* (BR14-Mulato) após inoculação com CABMV

Após 1 h da inoculação com o vírus (CABMV-1h) a variedade BR14-Mulato expressou 239 TE-DEGs (18,2% dos DEGs desse tempo) que incluíram 1.945 transposons, entre os quais um *LINE* (L1), 62 *LTR* (*Copia e Gypsy*), um *Crypton*, 33 de *TIR* (*CACTA, hAT, MULE e PIF*) e 52 *nested*-TEs (Tabela 2).

Após 16 h, nenhum *LINE* e nenhum *Crypton* foi identificado, mantendo um padrão similar ao acesso IT85F-2687 (CPSMV), com menor número de transcritos (34 TE-DEGs; 6,8% dos DEGs deste tempo), com oito *LTRs* (*Copia e Gypsy*), sete *TIRs* (*CACTA, hAT, MULE*) e 13 *nested-TE*s, sendo apenas três reprimidos (dois *Gypsy* e um *CACTA*).

Identificação de TEs em *V. unguiculata* **(Pingo de Ouro) sob desidratação radicular**

Após 25 min de exposição à desidratação radicular (RD-25min), o acesso Pingo de Ouro considerado tolerante à seca (vide Jesus-Pires et al., 2019) expressou 689 transcritos TE (15,7% dos DEGs desse tempo), compostos por 5.838 transposons, 13 *LINEs* (*L1, Tad1, CR1*), 212 *LTRs* (*Copia, Gypsy, Retrovirus*), três *Crypton,* 89 *TIRs* (*CACTA, hAT, MULE, PIF,TC-mariner*) e 151 *nested*-TEs (Tabela 3).

Neste estresse, a resposta tardia (RD-150min) apresentou maior quantidade de TEs modulados. Foram identificados 798 transcritos (18,6% dos DEGs deste tempo) associados às raízes de Pingo de Ouro, compreendendo 6.297 transposons, classificados em 12 *LINE*s (*L1, Tad1, CR1*), 224 *LTRs* (*Copia, Gypsy, Retrovírus*), 118 *TIRs* (*CACTA, hAT, MULE, PIF,TC-mariner*) e 205 *nested*-TEs.

Tabela 2 - Distribuição de elementos transponíveis diferencialmente expressos (TE-DEGs) por superfamília no acesso BR14-Mulato de *Vigna unguiculata* após inoculação com CABMV (Cowpea Aphid-Born Mosaic Virus).

Classificação			DEGs 1 h	DEGs 16 h	Tempo 1: 1 h				Tempo 2: 16 h			
					Up		Down		Up		Down	
Retrotransposon		**Desconhecida**	**-**	**-**	-	-	-	-	-	-	-	-
	LTR	**Copia**	**22**	**2**	10	28	12	39	2	5	-	-
		Copia + *Unk*	**15**	**2**	12	123	3	56	2	26	-	-
		Gypsy	**13**	**3**	9	31	4	15	1	14	2	3
		Gypsy + *Unk*	**11**	**1**	4	80	7	56	1	2	-	-
		Retrovirus	**-**	**-**	-	-	-	-	-	-	-	-
		Retrovirus + *Unk*	-	**-**	-	-	-	-	-	-	-	-

DNA-transposon	LINE	**L1**	**1**	**-**	1	1	-	-	-	-	-	-
		L1 + *Unk*	**-**	**-**	-	-	-	-	-	-	-	-
		Tad1	**-**	**-**	-	-	-	-	-	-	-	-
		Tad1 + *Unk*	**-**	**-**	-	-	-	-	-	-	-	-
		CR1 + *Unk*	**-**	**-**	-	-	-	-	-	-	-	-
	TIR	**CACTA**	**3**	**-**	2	8	1	1	-	-	-	-
		CACTA + *Unk*	**1**	**2**	1	2	-	-	1	24	1	10
		hAT	**12**	**1**	7	14	5	5	1	1	-	-
		hAT + *Unk*	**8**	**2**	7	65	1	8	2	17	-	-
		MULE	**2**	**2**	2	2	-	-	2	2	-	-
		MULE + *Unk*	**2**	**-**	1	14	1	11	-	-	-	-
		PIF	**-**	**-**	-	-	-	-	-	-	-	-
		PIF + *Unk*	**5**	**-**	5	42	-	-	-	-	-	-
		TC-mariner	**-**	**-**	-	-	-	-	-	-	-	-
		TC-mariner + *Unk*	**-**	**-**	-	-	-	-	-	-	-	-
		Merlin	**-**	**-**	-	-	-	-	-	-	-	-
		Crypton	**1**	**-**	1	1	-	-	-	-	-	-
		Unk	**91**	**15**	75	210	16	36	8	26	7	7
		***nested*-TEs**	**52**	**4**	39	765	13	332	4	39	-	-
		Total	**239**	**34**			**1945**				**176**	
			DEGs						Transposons			

Fonte: A autora (2019). Número de transcritos de TE-DEGs (D) e de cópias de TEs (T) identificados após 1 e 16 h de estresse biótico (vírus). Os dados estão agrupados por classe, ordem e superfamília. Um único transcrito pode apresentar mais de uma estrutura de TE. Quando um mesmo transcrito apresentou um TE desconhecido (*Unk*) e uma superfamília conhecida (superfamília + *Unk*), ele foi classificado de acordo com este segundo. Transcritos que apresentam mais de um TE de superfamílias distintas são chamados *nested*-TEs e estão destacados separadamente. A distribuição dos transcritos é distinguida em induzidos (UP) e reprimidos (DOWN).

Tabela 3 - Distribuição de elementos transponíveis diferencialmente expressos (TE-DEGs) por superfamília nas raízes do acesso Pingo de Ouro de *Vigna unguiculata*.

		Classificação	DEGs 25 min	DEGs 150 min	Tempo 1: 25 min				Tempo 2: 150 min			
					Up		Down		Up		Down	
					D	T	D	T	D	T	D	T
Retrotransposon	LTR	**Desconhecida**	**1**	**-**	-	-	1	10	-	-	-	-
		Copia	**70**	**69**	30	73	40	97	48	134	21	32
		Copia + *Unk*	**59**	**52**	23	118	36	472	35	339	17	202
		Gypsy	**47**	**49**	22	62	25	39	29	66	20	101
		Gypsy + *Unk*	**41**	**50**	16	323	25	157	29	206	21	314
		Retrovirus	**2**	**2**	-	-	2	7	2	2	-	-
		Retrovirus + *Unk*	**2**	**2**	1	6	1	2	2	8	-	-
	LINE	**L1**	**8**	**6**	1	1	7	15	6	12	-	-
		L1 + *Unk*	**3**	**-**	2	39	1	7	-	-	-	-
		Tad1	**1**	**3**	1	1	-	-	1	1	3	2
		Tad1 + *Unk*	**-**	**-**	-	-	-	-	-	-	1	2
		CR1 + *Unk*	**1**	**2**	1	3	-	-	1	3	1	3
DNA-transposon	TIR	**CACTA**	**7**	**10**	2	2	5	6	4	4	6	26
		CACTA + *Unk*	**14**	**11**	4	51	10	106	5	74	6	42
		hAT	**18**	**28**	11	22	7	7	18	30	10	10
		hAT + *Unk*	**10**	**22**	5	34	5	39	14	110	8	67
		MULE	**17**	**21**	7	7	10	16	9	9	12	12
		MULE + *Unk*	**14**	**11**	7	85	7	59	8	85	3	26
		PIF	**3**	**5**	2	3	1	1	2	2	3	4
		PIF + *Unk*	**6**	**8**	3	26	3	26	3	13	5	41
		TC-mariner	**-**	**-**	-	-	-	-	-	-	-	-
		TC-mariner + *Unk*	**-**	**1**	-	-	-	-	1	4	-	-
		Merlin	**-**	**1**	-	-	-	-	-	-	1	1
		Crypton	**3**	**4**	3	3	-	-	2	2	2	2
		Unk	**223**	**235**	102	402	119	374	139	360	96	281
		***nested*-TEs**	**151**	**205**	75	1599	76	1538	128	2239	77	1426
		TOTAL	**689**	**798**			**5828**				**6297**	
					DEGs				Transposons			

Fonte: A autora (2019). Número de transcritos de TE-DEGs e de cópias de TEs identificados após 25 e 150 min de estresse abiótico (desidratação radicular; RD) para o acesso Pingo de Ouro. Os dados estão agrupados em classe, ordem e superfamília. Um único TE-DEG pode apresentar mais de uma estrutura de TE. Quando um mesmo transcrito apresentou um TE desconhecido (Unk) e uma

superfamília conhecida (superfamília + Unk), ele foi classificado de acordo com este segundo. Transcritos que apresentam mais de um TE de superfamílias distintas são chamados nested-TEs e estão destacados separadamente. A distribuição dos transcritos é distinguida em induzidos (UP) e reprimidos (DOWN).

2.2.2 Anotação funcional de TE-DEGs baseada em GO

Nos três experimentos, em todos os tempos, TE-DEGs que continham sequências referentes à categoria componente celular foram minoria, abrangendo entre 10% e 15% das sequências anotadas, com destaque para componentes de membrana (componente integral de membrana, componente de membrana reticular, nuclear, vacuolar) (Anexo B, Material Suplementar 5). Entre as sequências anotadas, a maioria (Tabela 4) é referente a transcritos *Unk*. É possível que a dificuldade em classificar tais elementos seja devida ao alto grau de divergência destas sequências com TEs previamente descritos e consequente especificidade para *V. unguiculata,* ou que representem elementos cooptados cujas estruturas não permitem identificar a superfamília ou classe de TE de origem.

Tabela 4 - Número de transcritos de TE diferencialmente expressos anotados pelo GO e BLASTx (UniProt) em acessos de *Vigna unguiculata* sob diferentes condições.

	Total anotado	*Unk*	*Unk* %
CPSMV 1h	59	24	40,67
CPSMV 16h	39	17	43,58
CABMV 1h	142	57	40,14
CABMV 16h	14	7	50,00
RD 25 min	375	107	28,53
RD 150 min	465	143	29,48

Fonte: A autora (2019). Acesso IT85F-2687 inoculado com CPSMV (Cowpea Severe Mosaic Virus), acesso BR14-Mulato inoculado com CABMV (Cowpea Aphid-Born Mosaic Virus) e cultivar Pingo de Ouro sob desidratação radicular. Os elementos cuja classificação é desconhecida (Unk) compreenderam a maioria para todos os acessos.

Anotação funcional em TE-DEGs de *V. unguiculata* (IT85F-2687) inoculado com CPSMV

Apenas 52,21% (59) dos TE-DEGs em CPSMV-1h foram anotados pelo GO, alguns associados a mais de um termo, sendo 41 termos referentes a processo biológico, 37 referentes à função molecular e 15 a componente celular (Tabela 5).

Observou-se que algumas anotações de transcritos induzidos apontaram relação com a infecção como: 'transdução de sinal' (elemento *Unk*), 'regulação do crescimento' junto à 'sinalização mediada por ácido giberélico' (nested-TE), 'ligação com íon de cálcio' (nested-TE), bem como seis transcritos relacionados ao 'processo metabólico' de carboidratos (dois *Unk*, dois *nested*-TEs, *hAT e Gypsy*). Neste último, os dois *nested*-TEs contiveram sequências de *Gypsy* enquanto um teve também similaridade com *hAT*.

Para CPSMV-16h, 60,93% (39) dos TE-DEGs foram anotados, sendo 37 termos referentes a função molecular, 19 a processos biológicos e 10 de componente celular (Tabela 5). Entre os TE-DEGs induzidos estavam os transcritos de 'resposta a feromônios" (*Unk*) e de 'ligação com íon de cálcio' (*Copia*). Tanto para 16 h quanto para 1 h, o transcrito de íon de cálcio estava sendo carregado por *LTR*s, dado que o *nested*-TEs em CPSMV-1h foi composto por *Copia* e *Gypsy*. Dois transcritos ligados ao processo biológico de 'resposta de defesa' estavam classificados como elemento *Copia*, porém estão reprimidos (Anexo B, Material Suplementar 5).

Anotação funcional em TEs de *V. unguiculata* (BR14-Mulato) inoculado com CABMV

Um total de 59,41% (142) dos TE-DEGs foram anotados contra o GO para CABMV-1h, com 233 termos relacionados à função molecular, 139 a processos biológicos e 29 a componente celular (Tabela 5). A função de 'atividade de fator de transcrição (TF) com ligação ao DNA' foi observada em 12 transcritos (três *Copia,* dois um *L1,* dois *hAT*, dois *nested*-TEs e quatro *Unk*), oito deles induzidos. Com 138 termos relacionados a processos biológicos, destacaram-se três transcritos *Copia* de

‘resposta de defesa’ (dois deles como ‘resposta a estímulo biótico’), um *nested*-TE (*Copia, Gypsy e hAT*) de ‘resposta a hormônio’, 10 transcritos relacionados a ‘transdução de sinal’ (*Copia, hAT,* dois *nested*-TEs e seis *Unk*), três de ‘biossíntese de celulose’, entre outros (Anexo B, Material Suplementar 5).

Com apenas 14 transcritos anotados (21,87%), BR14-Mulato após 16 h de inoculação de CAMBV foi o tratamento com menor número de TE-DEGs identificados e menor proporção de transcritos anotados. Apesar disso, com 20 termos associados a função molecular, 9 a processos biológicos e 3 a componente celular, um termo com função predita de ‘integração de DNA’ (processo biológico) foi anotado contra um elemento *Copia* induzido. Este transcrito alinhou contra 10 elementos *Copia* na identificação via RepeatMasker, o que fortalece a hipótese de que pode ter ocorrido atividade de transposição relativamente recente (Anexo B, Material Suplementar 4 e 5). Foi encontrado também um elemento *hAT* anotado como ‘transdução de sinal’.

Anotação funcional em TE-DEGs de *V. unguiculata* (Pingo de Ouro), desidratação radicular

Com o maior número de TE-DEGs expressos, o acesso Pingo de Ouro sob condição de desidratação radicular teve 53,69% (371) dos transcritos de RD-25min anotados e 54,88% (438) de RD-150min. Entre os 1007 termos (577, função molecular; 327, processos biológicos; 103, componentes de membrana) descritos em RD-25min, considerando-se os transcritos induzidos com termos relacionados a estresses, observaram-se anotações como: ‘resposta a estímulo luminoso’ (*Unk*), que condiz com a exposição à luz sofrida pela raiz durante o processo de desidratação (ver detalhe na Metodologia), bem como resposta a estresse osmótico e privação de água (*Copia*). Além disso, 23 DEGs (*LTRs, TIRs nested*-TEs e *Unk*) estavam relacionados a ‘atividade TF com ligação ao DNA’ sendo 15 induzidos.

Tabela 5 - Anotação de termos de ontologia gênica em elementos transponíveis diferencialmente expressos (TE-DEGs), para cada um dos tratamentos em *Vigna unguiculata*.

	Categoria GO / Tratamento	Número de TE-DEGs	Número de TE-DEGs
	IT85F-2687 / CPSMV	**1 h**	**16 h**
Processo Biológico	Outros	17	12
	Proteína de fosforilação	8	0
	Processo de oxiredução	6	4
	Processo metabólico de carboidratos	6	0
	Processos metabólicos	4	1
Componente Celular	Outros	6	3
	Membrana	4	1
	Componente integral de membrana	4	5
Função Molecular	Outros	14	29
	Ligação com ATP	8	2
	Atividade de proteína quinase	8	2
	Ligação *heme*	4	4
	Atividade de oxidoredutase, atuando na ligação com doadores, com incorporação ou redução de moléculas de oxigênio	3	0
	BR14-Mulato / CABMV	**1 h**	**16 h**
Processo Biológico	Outros	53	3
	Fosforilação de proteínas	30	0
	Processo de oxidação-redução	23	2
	Regulação da transcrição, DNA-*templated*	15	0
	Transdução de sinal	9	1
	Processo metabólico	8	3
Componente Celular	Outros	6	1
	Componente integral de membrana	11	2
	Membrana	7	0
	Parede celular	3	0
Função Molecular	Outros	142	17
	Ligação com ATP	32	0
	Atividade de proteína quinase	29	0
	Ligação *heme*	15	1
	Proteína de ligação	15	2
Pingo de Ouro / Desidratação Radicular		**25 min**	**150 min**
Processo Biológico	Outros	171	221

	Fosforilação de proteínas	57	80
	Processo de oxidação-redução	48	58
	Regulação da transcrição, DNA-*templated*	30	32
	Processo metabólico	23	33
Componente Celular	Outros	42	39
	Componente integral de membrana	29	33
	Núcleo	18	25
	Membrana	8	21
	Intracelular	2	6
Função Molecular	Outros	341	433
	Ligação com ATP	73	95
	Proteína de ligação	61	62
	Atividade de proteína quinase	57	80
	Atividade catalítica	26	23
	Ligação *heme*	24	24

Fonte: A autora (2019). Acesso IT85F-2687 inoculado com CPSMV (1 e 16 h), BR14-Mulato inoculado com CABMV (1 e 16 h) e Pingo de Ouro submetido à desidratação radicular (25 e 150 min). Maiores detalhes no Material Suplementar 5.

Além desses, dois transcritos induzidos foram anotados como ‘resposta a estresse oxidativo’ além de um como ‘resposta a infecção bacteriana’, todos pertencentes à superfamília *hAT* (Tabela 4, Material Suplementar 5).

Após 150 min de desidratação radicular, com 1273 termos (720, função molecular; 429, processos biológicos; 124 componentes celular), foram observadas também anotações de processos envolvendo ‘integração de DNA’ (*Unk*) e ‘desenvolvimento pós-embrionário de raiz’ (*CACTA*), ambos reprimidos. Já os transcritos relacionados a ‘crescimento de raiz lateral’ (*Copia*) e ‘catabolismo de parede celular’ encontraram-se induzidos e foram exclusivos dos TE-DEGs desse tempo. Identificada em 25 transcritos, a ‘atividade de TF com ligação ao DNA’ está presente em *LTRs, TIRs, nested-TEs e Unk,* dos quais 17 são induzidos (4 *TIRs*, 6 *LTRs*, 3 *nested*-TEs e 7 *Unk*).

2.2.3 Anotação de domínios de TE-DEGs baseada no CD-search

Anotação de domínios de TE-DEGs de *V. unguiculata* (IT85F-2687) inoculado com CPSMV

Em CPSMV-1h, 60 (53,10%) dos TE-DEGs foram anotados com o CD-search, totalizando 15 de classe I, seis de classe II, 15 *nested*-TEs e 24 Unk. Transcritos com domínio de transdução de sinal que apresentam domínios completos capturados por um TE unk encontram-se induzidos. Por outro ladoos capturados por um PIF-harbinger, estão induzidos. Apenas um fator de transcrição (TF) da família bZIP foi identificado, sendo este completo e induzido. Sua anotação revelou estar relacionado ao desenvolvimento da semente e à regulação positiva da fotomorfogenêse. O dois transcritos de transporte e metabolismo de carboidratos com domínio completo compreenderam *nested*-TEs e também mostraram-se induzidos. Transcritos Toll/interleukin-1 receptor (STKc_IRAK) e um AAI_LTSS (Alpha-Amylase Inhibitors, Lipid Transfer and Seed Storage) envolvidos da resposta a estresse biótico mostraram domínios completos em todos os TE-DEGs em que foram identificados, sendo reprimidos quando capturados por DNA-transposons (*hAT* e *PIF*) e induzidos em *Copia*, *nested*-TEs e Unk.

Foram anotados 41 (61,20%) TE-DEGs para CPSMV-16h, sendo 18 de classe I, três de classe II, sete *nested*-TEs e 13 unk (Material Suplementar S6). Apenas um transcrito que participa do mecanismo de transdução de sinal mostrou-se completo, sendo capturado por um *MULE* (induzido). O mesmo se aplicou ao transcrito envolvido no transporte e metabolismo de carboidratos (*Copia*). Entre TE-DEGs de resposta a estresse biótico dois STKc_IRAK foram identificados, sendo um induzido e um reprimido (*nested*-TEs). As respostas hormonais como os dois transcritos Pyrabactin resistance 1 (PYR1) like (PYR1-like; capturado por *Copia*) responsáveis pela mediação da sinalização do ácido abscísico (ABA) e um de reconhecimento de promotor responsivo a etileno (AP2; *Gypsy*) estavam eprimidos, enquanto o TE-DEGs Unk com AP2 mostrou-se induzido.

Anotação de domínios de TE-DEGs de *V. unguiculata* (BR14-Mulato) inoculado com CABMV

Com 168 (70,30%) transcritos anotados , pelo CD-search, CABMV-1h possui 50 TE-DEGs de classe I, 23 de classe II, 33 *nested*-TEs e 64 *unk* (Material Suplementar S6). Entre transcritos com domínio completo estão domínios de transdução de sinal, os quais geralmente inseridos em Retrotransposons e induzidos; TFs estão presentes nas duas classes (I e II de TEs), como o WRKY, o MYB-like e o TCP induzidos, ao contrário de bZip, APETALA2-like e TF responsivo a auxina questão e reprimidos. Entre os domínios responsáveis por transporte e metabolismo de carboidratos, apenas dois estão completos sendo um induzido (Gypsy) e um reprimido (Copia). Domínios relacionados a resposta hormonal, como acido jasmônico, etileno e ácido abscísico (ABA), estão inseridos em Unk ou retrotransposons, induzidos, com exceção de um Copia que encontra-se reprimido. Além desses, outros domínios de resposta a patógenos como STKc_IRAK (Gypsy, induzido), CC-NB-LRR (Copia, reprimido) e calmodulinas induzidas (*CACTA* e Unk) que podem induzir a resposta de defesa da planta.

Foram anotados apenas 11 (16,42%) transcritos para CABMV-16h (Material suplementar 6). Dentre eles, seis Unk, dos quais um domínio completo de ligação ao cálcio (reprimido). Entre os elementos *hAT*, uma STKc_IRAK (induzida) e um AAI_LTSS (induzido) ambos relacionados com resposta a estresse biótico.

Anotação de domínios de TE-DEGs de *V. unguiculata* (Pingo de Ouro) sob desidratação radicular

Em RD-25min 405 (58,79%) transcritos foram anotados pelo CD-search, (Material suplementar 6), sendo 133 classe I, 65 classe II, 84 *nested*-TEs e 123 Unk. A maioria (63,64%) dos TE-DEGs com domínios que indicam participação em mecanismos de transdução de sinal apresentaram-se reprimidos, independente da classe (I ou II), sendo 52,73% dos TE-DEGs tiveram domínio incompleto. Quatro TFs (um WRKY, um bZIP, um TF histona-like) tinham domínio completo e estavam induzidos, em contraste com os TFs R2R3-MYB (responsivo a auxina), o TBP (proteína de ligação ao TATA box) e o SBP (responsável pelo controle de floração)

que mostraram-se reprimidos. Os três domínios GRAS (um *Copia* e dois Unk) relacionados ao desenvolvimento, crescimento da planta e componentes na sinalização da giberelina apresentaram-se reprimidos, o que pode ser justificado devido ao choque sofrido com a desidratação radicular. Todos os TE-DEGs descritos pelo CD-search como responsivos ou sinalizadores de seca e salinidade mostraram-se reprimidos independente da sua estrutura completa ou incompleta, assim como os domínios AP2, EIN3, EDR1 responsivos ao estímulo do hormônio etileno e capturados por retrotransposon, exceto Tad1. Já os mesmos domínios quando presentes em um elemento *hAT* ou um *nested*-TE contendo *hAT*, mostraram-se induzidos.

Com 494 (62,87%) transcritos anotados pelo CD-search, RD-150min teve 148 TE-DEGs anotados para classe I, 76 para classe II, 124 para *nested*-TEs e 146 Unk. Ao contrário de RD-25min, em RD-150min a maioria (74,70%) dos TE-DEGs relacionados a transdução de sinal estão induzidos sendo um total de 20% com domínio completo. Apenas três de 15 TF estão reprimidos e 13 estão completos. O domínio GRAS apareceu apenas em retrotransposons, sendo três induzidos e dois reprimidos e todos completos. A Glutathione S-transferase (GST), também completa e capturada por elemento *hAT* encontra-se induzida e responde a estresses no solo, auxina e citoquinas. Outros hormônios, como o acido jasmônico (domínio TIFY), o etileno (domínios AP2, EIN3 e EDR1 completos) também provocam a indução de TE-DEGs . Adicionalmente, diferente de RD-25min, o tempo tardio teve todos, exceto um, dos domínios responsivos a seca induzidos.

2.2.3 Expressão Diferencial de Elementos Transponíveis

Nos três acessos, sob três diferentes condições e em tecidos de dois órgãos diferentes, os elementos do tipo *Copia* (26 CPSMV, 41 CABMV, 240 RD) e *hAT* (10 CPSMV, 23 CABMV, 78 RD) foram identificados em maior número, além de serem parte de vários *nested*-TEs, por vezes no mesmo transcrito. Apesar disso, o número de TE-DEGs induzidos e reprimidos para cada superfamília em um mesmo acesso e tempo foi mais similar para os estresses bióticos (viroses) do que o abiótico (desidratação radicular; vide Tabelas 1, 2 e 3).

Entre os TE-DEGs do tempo precoce e tardio de CPSMV e CABMV, apenas um e três transcritos, respectivamente, foram comuns aos dois tempos, indicando alto nível de especificidade na expressão das isoformas das superfamílias encontradas em relação a estresse por vírus de RNA.

Em RD, 137 transcritos (10,2%) foram compartilhadas entre os dois tempos, dos quais 68 (reprimidas em RD-25min e induzidas em RD-150min) e 53 (induzidas em RD-25min e reprimidas em RD-150min) com respostas opostas. Com apenas 2% dos transcritos apresentando o mesmo comportamento nas duas bibliotecas, pode-se dizer que a expressão de TEs classe I ou II nas cultivares estudadas apresenta expressão bastante específica. Este padrão também pôde ser constatado ao contrastar a expressão de isoformas nos diferentes tempos de estresse (Figuras 3, 4 e 5), onde ficou evidente a formação de agrupamentos de TEs cuja expressão foi contrastante (induzido/reprimido) ou apresentou-se com alteração gradativa entre os tempos.

Expressão diferencial de TEs de *V. unguiculata* (IT85F-2687) inoculado com CPSMV

No tempo precoce (CPSMV-1h), o elemento *Copia* teve quantidade parecida de elementos induzidos (sete) e reprimidos (cinco), assim como *hAT* (cinco induzidos, três reprimidos), *nested*-TEs (aqueles que incluem superfamílias diferentes na mesma estrutura) foram em sua maioria induzidos (16) (Tabela 1).

No conjunto de TE-DEGs identificados, alguns estavam relacionados com a resposta ao estresse, de acordo com as anotações. O transcrito Vu.r_00191366 (1,77 log2FC) de *hAT* referiu-se a ‘processamento de carboidratos’, contudo seu domínio mostrou-se incompleto. Além desse, outros cinco merecem menção, dois *nested*-TEs, dois *Unk* e um *Gypsy* (*nested*-TEs contendo também sequências de *hAT*). O *nested*-TE Vu.r_00093198 (3.19 log2FC) carregava sequência relacionada à ‘regulação do crescimento’ e ‘sinalização do ácido giberélico’ e o Vu.r_00093925 foi anotado contra função molecular de ligação de íon de cálcio, processo importante na transdução de sinal, além de apresentar o domínio de matriz de metaloproteinase dependente de zinco e envolvido em reparo e apoptose. Este foi induzido em 3,75

log2FC. Já o DEG Vu.r_00100829 (*Unk*) estava induzido em 3,38 log2FC, expressando sequência relacionada à proteólise com domínios completos.

Em CPSMV-16h um elemento *Gypsy* associado a uma sequência específica de ligação com DNA com atividade de TF (domínios incompleto) foi reprimido (log2FC = -1,68). Contudo, o TE-DEG Vu.r_00174381 (*Unk*) com domínios completos de AP2 (TF responsivo ao etileno) mostrou-se induzido (log2FC = 7,93) com possível atuação como TF na resposta ao estresse. Ao contrário do esperado, os dois *elementos Copia* (PYR1-like, Vu.r_00209229, Vu.r_00209230) relacionados à resposta de defesa estavam reprimidos. Já o transcrito Vu.00159051 (classificado como unk) foi induzido (log2FC = 1.71), incluindo domínio completo de receptor de membrana e função predita de resposta de feromônio.

Expressão diferencial de TEs de *V. unguiculata* (BR14-Mulato) inoculado com CABMV

No tempo inicial (CABMV-1h), três elementos *Copia* induzidos e com domínios completos (PYR1-like, Vu.r_00209230, log2FC = 2,38; PYR1-like, Vu.r_00209229, log2FC = 2,66; proteína integral de membrana Mlo, Vu.r_ 00216721, log2FC = 3,27) foram caracterizados em ‘resposta de defesa’, enquanto dois deles foram anotados também em ‘resposta a estímulo biótico’. Adicionalmente, o elemento *Unk* (Vu.r_ 00020373) induzido (log2FC = 6,56) compreendeu uma sequência de transdução de sinal com domínios completos. Por sua vez, dentre os doze transcritos com atividade de ligação a TF, oito mostraram-se induzidos (entre 5,15 e 1,61 Log2FC, respectivamente), com um bZIP completo e quatro WRKY completos, sendo o mais induzido um elemento WRKY capturado por um *hAT* (Material Suplementar 5 e 7). O *nested*-TE (Vu.r_00075017) composto por *hAT* e *Gypsy* estava induzido (log2FC = 4,02), expressando um transcrito relacionado ao ‘controle da transcrição’ e ‘resposta a hormônio’, caracterizado como um TF responsivo a auxina completo.

Após 16 h, o *nested*-TE (Vu.r_00124958) com indícios de atividade de giberelina mostrou-se reprimido (log2FC = -30). Por sua vez, o TE-DEG de *Copia* (Vu.r_00118343) anotado contra 10 elementos de mesma superfamília apresentou ‘atividade de integração de DNA’ e mostrou-se induzido (log2FC = 4,64). Esse TE-

deg possui uma RNAse_H com domínio completo de transcriptase reversa, apesar de não compreender outros domínios relacionados à transposição.

Expressão diferencial de TEs de *V. unguiculata* (Pingo de Ouro) sob desidratação radicular

Em RD-25min todos os TE-DEGs, completos ou não, de transporte iônico e de cátions foram reprimidos (log2FC entre -1,62 e -4,32) com exceção de um elemento *Unk* (Vu.r_00191670), cuja expressão ocorreu exclusivamente na planta estressada (Material Suplementar 5). Um TE-DEG *Unk* (Vu.r_00025162) anotado em ‘resposta ao estímulo luminoso’ e ‘controle de crescimento’ mostrou-se induzido (log2FC = 1,57), da mesma forma o elemento *Copia* (Vu.r_00162499; log2FC = 2,45) relacionado a ‘estresse osmótico’ e ‘resposta a privação de água’. Os 15 DEGs induzidos de ‘atividade de ligação a TF’ mostraram-se induzidos com log2FC entre 1,57 e 4,05, classificados em superfamílias de Retrotransposons (*Gypsy* e Tad1) e DNA-transposons (*hAT* e MULE), além de *nested*-TEs (dois de MULE com Retroelementos e um com dois elementos *TIR*, *hAT* e *PIF-harbinger*), e *Unk*. Dentre os 15 TFs, seis apresentaram domínio completo (quatro WRKY, um TF histona-like e um bZIP)

Em RD-150min, *nested*-TEs de *LTRs* (Vu.r_00218397) relacionados a ‘transporte de água’ mostraram-se reprimidos (log2FC = -2,41), ao contrário dos *nested*-TEs classe II (*CACTA* e *hAT*) anotados contra termos de ‘controle da giberelina’ e ‘crescimento de raiz’ (Vu.r_00002885 e Vu.r_00215059; log2FC = 1,93 e 4,37, respectivamente). Entre os 19 TE-DEGs induzidos e anotados com ‘atividade de ligação com o TF’, a expressão variou de log2FC 1,70 a 4,04, incluindo transcritos *LTRs, TIRs* e *Unk,* contudo apenas cinco mostraram-se completos (três WRKY, um Myb-CC e um TF histona-like). Adicionalmente, um WRKY foi anotado em associação com um elemento MULE por meio de BLASTx e mostrou-se induzido (log2FC = 1,92) (Material Suplementar 5).

2.2.4 Ancoragem de Elementos Transponíveis Diferencialmente Expressos no Genoma de *V. unguiculata*

Nas Figuras 6, 7 e 8 é possível observar uma distribuição não regular de TEs induzidos e reprimidos ao longo dos pseudocromossomos com base no genoma de feijão-caupi. Além disso, observa-se uma tendência de que genes com o mesmo comportamento (induzidos ou reprimidos) apresentem-se clusterizados, havendo inclusive grandes regiões do genoma sem qualquer expressão em tempos precoces (por exemplo Vu09) ou no tempo mais tardio analisado (como é o caso de Vu03 e Vu06) nos três tratamentos (Figuras 6, 7 e 8). Apesar de algumas sequências estarem ancoradas na região centroméricas e de pseudocentrômeros (chamados também de nós cromossômicos), os DEGs de TEs encontraram-se dispersos nos cromossomos e próximos de regiões gênicas.

Ancoragem de TEs diferencialmente expressos no acesso IT85F-2687 inoculado com CPSMV

Ao ancorar as sequências de CPSMV-1h, foi observado que os cromossomos Vu01, Vu02, Vu05, Vu07, Vu09 apresentaram maior aglomeração de TEs diferencialmente expressos no braço curto do cromossomo, enquanto o Vu11 teve DEGs clusterizados no braço longo e apresentou um cluster denso de sequências induzidas na região distal. As sequências induzidas apresentaram-se mais aglomeradas que as reprimidas, exceto para Vu08, onde foi possível observar no braço curto dois aglomerados densos de sequências reprimidas (Figura 6A). Para CPSMV-16h, aglomerados de sequências reprimidas tornaram-se mais abundantes embora menos densos devido ao baixo número de DEGs (Figura 6B).

Ancoragem de TEs diferencialmente expressos no acesso BR14-Mulato inoculado com CABMV

Em CABMV-1h os cromossomos Vu01, Vu08, Vu09 e Vu10 apresentaram mais sequências no braço curto, enquanto em Vu04 e Vu11 houve maior densidade

no braço longo. O cromossomo vu04 apresentou as sequências menos dispersas, com *clusters* especialmente evidentes entre 0-4 Mb e na região pericentromérica, onde as sequências encontraram-se induzidas. Sequências pericentroméricas também foram observadas em Vu02 (induzidas) e Vu05 (induzidas e reprimidas) (Figura 7A). Para CABMV-16h, o total de sequências moduladas foi muito baixo com algumas mostrando-se clusterizadas em posições terminais (Figura 7B).

Ancoragem de TEs diferencialmente expressos no acesso 'Pingo de Ouro' sob desidratação radicular

Em RD-25min, os cromossomos apresentaram maior densidade de TEs modulados nos braços curtos, exceto para Vu10 e Vu11 onde foram mais dispersos e em Vu05 onde mais sequências são observadas no braço longo. Os cromossomos Vu04 e Vu05 apresentaram grande concentração de sequências centroméricas quando comparados com os demais (Figura 8A). Em RD-150min, a distribuição dos TEs não apresentou grandes mudanças em relação ao tempo inicial, corroborando com o número de transcritos comuns aos dois tempos e indicando que mesmo os transcritos não compartilhados, puderam compreender isoformas de um mesmo gene que apresentaram expressão específica de acordo com o tempo de estresse (Figura 8B).

3. Discussão

A proporção de TEs entre os transcritos diferencialmente expressos nos dois acessos tolerantes a vírus variou entre 7% e 18%, valor próximo ao encontrado nos dados de RNA-Seq de arroz suscetível a RSV (*Rice strip vírus*) [31] (Cho et al., 2015). Contudo, diferente do trabalho com arroz, onde mais de 1.000 TEs foram significativamente modulados e preferencialmente reprimidos, em feijão-caupi cerca de 300 TEs foram significativos, com proporção de TEs induzidos nas folhas quase duas vezes maior que a de TEs reprimidos. Deve-se considerar o fato de que o arroz já foi descrito como uma das plantas com mais alta atividade de TEs [32] (Zhao, 2018b), o que poderia justificar essa diferença.

No acesso sob desidratação, a quantidade de TEs em relação ao total de DEGs na raiz variou entre 15-18% sendo, portanto, muito diferente do observado para tecidos aéreos em espécies de *Arabidopsis* sob condições de desidratação (2-4%) [33] (Göbel et al., 2018). Em parte essa diferença se justifica por uma proporção muito menor de TEs em *Arabidopsis* (10-15%) [34] (Arabidopsis Genome Initiative, 2000) do observado em *V. unguiculata* (50-60%) [27] (Lonardi et al., 2019). Além disso, ressalta-se a diferença de tecido avaliado, dado que para os *LTR*, por exemplo, a expressão ocorre preferencialmente em raízes [10,35] (Domingues et al., 2012; de-Araujo et al., 2010), sendo este o grupo mais representativo entre os elementos diferencialmente expressos aqui identificados.

Entre os termos anotados para os acessos com inóculo viral, destaca-se o ‘metabolismo de carboidratos’, descrito anteriormente como mediador da resistência à infecção por vírus, induzindo genes responsivos e fortalecendo a resposta tolerante da planta [36,37] (Herbers et al., 2001, Scharte et al., 2005). Nesse caso, a resposta é especialmente observada no tempo precoce (1 h), sendo mais rápida no feijão-caupi do que nos estudos citados, onde o acúmulo de açúcares começou apenas após 2 dias de estresse. É preciso observar se outros genes não relacionados a TEs possuem o mesmo comportamento. Contudo, este resultado sugere que os transcritos identificados (6 CPSMV-1h, 4 CABMV-1h, 1 CABMV-16h) podem exercer um papel importante na tolerância da planta, considerando que parte possui domínios completos, compreendendo, portanto, alvos moleculares interessantes, caso sua atividade seja comprovada.

A anotação dos termos GO para o acesso sob desidratação radicular apresentou resultado similar ao estudo de desidratação com soja [38] (Ferreira-Neto et al., 2013), indicando forte relação com a resposta ao estresse, observando-se que os termos mais anotados para soja e os termos encontrados entre TE-DEGs em *V. unguiculata* pertencem às mesmas categorias GO, como ligação com metais (metais no geral, ferro, zinco), ligação com ATP, redução oxidativa e processos metabólicos, entre outros.

As proteínas de ligação (a ácidos nucleicos, proteínas, íons, etc.) são um grupo de sequências bastante presente dentre os TEs identificados para os três acessos, e comumente encontrados em *nested*-TEs de eucariotos [39] (Gao et al., 2012), o que corrobora com um grande número de *nested*-TEs identificados no

presente trabalho. Além disso há estudos que indicam a cooptação de TEs para diversos fins [40] (Bennetzen e Wang, 2014).

Em todos os acessos estudados, atividades como TFs com ligação ao DNA foram identificadas, além de domínios completos de TFs tais como WRKY, MYB e TCP entre outros, sugerindo que – assim como para *A. thaliana* [20, 41] (Tuan-Ngoc et al., 2014; et al., Zhao et al., 2018) e *Boea hygrometrica* da família Gesneriaceae [18] (Zhao et al., 2014) – transposons podem atuar como promotores dos genes aos quais estão relacionados, como por exemplo genes de reposta a estímulo biótico, de resposta a estresse, privação de água e resposta a hormônio, aqui foram identificados.

TEs ativados por estímulo hormonal já foram descritos anteriormente em *A. thaliana* [42] (Ito et al., 2016), trabalho onde foi sugerido que um transposon seja responsável pela resposta tolerante. Os hormônios compõem uma das linhas de defesa das plantas e desencadeiam a modulação de vários genes a partir da transdução de sinais [43] (Nejat et al., 2017), como por exemplo os íons de cálcio, incluindo SPS1 e CBS identificados neste trabalho. Tratam-se de DEGs importantes para a transdução de sinais, sendo responsivos a estresses bióticos e abióticos em plantas superiores [7, 44] (Kim, 2012, Kollist et al., 2018).

Complementarmente, uma grande quantidade de elementos de classificação desconhecida foi anotada contra termos relacionados à resposta a estresse, e induzida. Este grupo de TEs específicos de feijão-caupi não possuem referência e podem representar uma ótima opção para identificação de novos elementos transponíveis, além de geração de marcadores de DNA (polimorfismos), como descrito por Wei et al. (2016) [45].

A distribuição cromossômica dos TEs diferencialmente expressos foi similar à descrita para soja [46] (Nakashima et al., 2018) e tomate [47] (Xu e Du, 2014), na qual elementos modulados estão dispersos nos braços cromossômicos e próximos a genes, indicando transposição recente. Esta teoria corrobora com a hipótese de que transcritos apresentam uma homologia significativa com transposons da mesma superfamília, indicando que compreendem TEs relativamente recentes, justificando o alto grau de similaridade observado.

4. Conclusão

Os elementos *Copia* e *hAT* além de serem abundantes no genoma, se destacaram para todos os estresses, observando-se respostas específicas para cada um. Ambos os transposons estão presentes isoladamente ou em *nested*-TEs.

Os acessos IT85F-2687 e BR14-Mulato possuem perfis de modulação de TEs bastante específicos, não havendo relação direta entre as isoformas observadas, reforçando a teoria de que a expressão de TEs é específica para cada cultivar e tempo após exposição ao estresse (precoce ou tardia), ainda que elas sejam geneticamente próximas.

O acesso Pingo de Ouro sob desidratação radicular modulou um número significativamente maior de TEs. Apesar de compartilhar isoformas entre os dois tempos após o estresse, apresentou modulações contrastantes comparando-se ambos os tempos, também reforçando que a indução ou repressão de elementos depende de uma regulação fina e reconhecimento específico.

Transcritos como de metabolismo de carboidratos, ligação com íons de cálcio, reconhecimento hormonal, estão geralmente induzidos e são profundamente relacionados com a resposta a estresse, portanto candidatos interessantes para experimentos de validação, além dos TEs sem referência prévia são potenciais marcadores para estudos funcionais e geração de marcadores polimórficos.

5. Metodologia

5.1 Material vegetal e estresses (bióticos/abiótico) aplicados

5.1.1 Ensaio de inoculação viral

Para o ensaio com os vírus CPSMV e CABMV amostra recém-propagada do vírus foi inoculada, respectivamente, nos acessos de feijão-caupi IT85F-2687 e BR14-Mulato, considerados resistentes contra o vírus em questão [28,48] (Kido et al., 2011; Jesus-Pires et al., 2019) Os ensaios foram realizados, em casa de vegetação no Instituto Agronômico de Pernambuco (IPA, Recife-PE, Brasil). Inicialmente, as plantas cresceram por 21 dias, sob fotoperíodo natural e temperatura entre 28 e 32°C. Em seguida, cada planta teve uma das folhas do trifólio injuriada e imediatamente inoculada. A injúria mecânica foi realizada com

Carborundum (carbeto de silício) para permitir a inoculação do vírus. O inóculo foi preparado com água destilada e as folhas maceradas, estas provenientes de um acesso sensível contendo vírus. O mesmo número de plantas não injuriadas e não-inoculadas foi coletado para a amostra-controle. Cada experimento (estresse e controle) foi realizado em local isolado e com as mesmas condições ambientais, a fim de evitar a influência de compostos voláteis entre os indivíduos. As folhas das plantas, ainda assintomáticas, bem como as de seus respectivos controles foram coletadas após 60 min e 16 h de inoculação dos vírus, sendo imediatamente congeladas em nitrogênio líquido e armazenadas a -80°C até extração de seu RNA. Para a análise dos dados, foram considerados dois acessos/vírus (IT85F-2687/CPSMV e BR14-Mulato/CABMV), dois tempos de inoculação (60 min e 16 h) e seus respectivos controles (não injuriados e não-inoculados), totalizando oito tratamentos, com três réplicas cada (cinco plantas por réplica) (Material Suplementar 6).

5.1.2 Ensaio de desidratação radicular

O ensaio foi realizado em casa de vegetação na EMBRAPA-Soja (Londrina-PR, Brasil). As sementes do acesso Pingo de Ouro (tolerante à seca) foram tratadas com 0,05% (w/v) Thiram (dissulfeto de tetrametiltiuram) para controle de infecção por patógenos. A seguir, foram colocadas para germinar por dois dias como descrito [49] Kulcheski et al. (2011), sob temperatura de 25°C ± 1°C e 65% ± 5 de umidade relativa. Em seguida, as plântulas foram transferidas para o sistema hidropônico [50] (Rodrigues et al., 2012) em containers de 30 L contendo solução aerada nutritiva em pH 6,6 [51] (Hoagland e Arnon, 1950). As plântulas foram mantidas em casa de vegetação com raízes completamente imersas na solução nutritiva, em temperaturas entre 30 e 35°C, umidade relativa de 60% ± 10 e fotoperíodo (13 h de luz e 11 h de escuro). Após 21 dias, plântulas trifoliadas (estágio de desenvolvimento V2; Fehr et al., 1971 [52]) foram suspensas e tiveram o contato das raízes com a solução totalmente cessado, dando início ao tratamento de desidratação radicular. As raízes foram coletadas após 25 e 150 min de estresse, congeladas imediatamente em nitrogênio líquido e armazenadas a -80°C. Para a análise dos dados, foram considerados dois tempos de desidratação (25 e 150 min) e seus controles

negativos (ausência do estresse) pareados, totalizando quatro tratamentos, com três réplicas biológicas cada (cinco indivíduos por réplica) (Material Suplementar 7).

5.2 Extração de RNA, síntese de cDNA e sequenciamento do RNA-Seq

O RNA total foi obtido a partir das folhas do ensaio de inoculação viral e das raízes do experimento de desidratação radicular. O RNA foi extraído por meio do kit SV Total RNA Isolation System (Promega, US), seguindo as instruções do fabricante. A concentração e a qualidade do RNA total extraído foram avaliadas por meio de gel de agarose (1,5%) e Agilent 2100 Bioanalyzer (Agilent Technologies, EUA). O kit RNAm TruSeq® Stranded LT-Set A (RS-122-2101) (Illumina, San Diego, CA, EUA) foi utilizado na purificação de RNAm e para a construção de bibliotecas de cDNA, de acordo com instruções do fabricante. *Reads paired-end* de 100 nucleotídeos foram geradas pelo sistema Illumina HiSeq 2500, com o uso dos kits: HiSeq® Rapid PE Cluster Kit v2 (PE-402-4002), SBS Kit v2 (200 Cycle; FC-402-4021) e TruSeq® Stranded mRNA LT - Set A (RS-122-2101). O sequenciamento ocorreu no Centro de Genômica Funcional da Universidade de São Paulo (São Paulo, SP, Brasil). Foram sequenciadas 24 bibliotecas do experimento de injúria/inoculação viral (50% CABMV e 50% CPSMV), bem como 12 bibliotecas do experimento de desidratação radicular, incluindo as três réplicas biológicas de cada tratamento, totalizando 36 bibliotecas (Material Suplementar 6).

5.3 Montagem do transcriptoma

As 72 réplicas (36 bibliotecas, com duas réplicas técnicas cada, Material Suplementar 6) foram submetidas ao pipeline RNA-Seq do GenPipes [53] (Bourgey et al., 2018). As métricas de qualidade das *reads* foram acessadas utilizando o RNA-SeQC v1.1.8 [54] (Deluca et al., 2012) e podem ser encontradas no parecer gerado pelo programa (Material Suplementar 2). O pacote Trimmomatic v0.36 foi utilizado para remoção dos adaptadores e de sequências de baixa qualidade [55] (Bolger et al., 2014). Em seguida, o STAR v3.5.3a [56] (Dobin et al., 2013) foi utilizado para fazer o mapeamento das *reads* contra o genoma de referência disponível para *V. unguiculata* v1.1 [27] (Lonardi et al., 2019), disponível no Phytozome v12, mantendo-

se aquelas que não foram mapeadas. O Picard v2.9 foi usado para mascarar as duplicatas sem removê-las, considerando pontuação (score) por qualidade das bases [57] (Van-der-Auwera et al., 2013), enquanto os transcritos foram montados com o pacote Cufflinks v2.2.1. Todos os parâmetros desta e das etapas seguintes estão descritos no Material Suplementar 8.

5.4 Identificação de elementos transponíveis (TEs)

As etapas de identificação dos TEs foram realizadas por meio do RepeatMasker v4.0.7 [58] (Smit et al., 2015) (Figura 2). Inicialmente, dados de eudicotiledôneas do RepBase Update [59] (Bao et al., 2015) (versão 11-24-2018) foram comparados contra os TEs do genoma de referência [27] (Lonardi et al., 2019) a fim de anotar TEs com classificação desconhecida. Em seguida, todos os TEs do genoma foram anotados contra o transcriptoma (figura 9). Foram considerados aqueles que continham pelo menos uma ORF com 80 nt ou mais, relacionadas a uma das superfamílias de TEs, ou que apresentavam uma cobertura e identidade de 80% ou superior [8] (Wicker et al., 2007).

5.5 Expressão diferencial de TEs

A partir do GenPipes, a análise de expressão diferencial para cada um dos tempos dos estresses aplicados (CPSMV, CABMV, RD) foi realizada contra seus respectivos controles usando *Cuffdiff* [60] (Trapnell et al., 2010), seguido da normalização com o *Cuffnorm* [61] (Trapnell et al., 2013). Foram considerados diferencialmente expressos transcritos que apresentaram p-value ≤ 0,05 e log2FC ≥ 1 ou log2FC ≤ -1. A descrição das comparações está disponível no Material Suplementar 1

5.6 Anotação funcional dos transcritos

O Trinotate foi usado para obtenção das anotações funcionais, realizadas a partir de BLAST contra o PFAM e nr do NCBI (Figura 2). As análises de ontologia

gênica foram feitas através do GenPipes com o GOseq [62] (Young et al., 2010) (figura 6).

5.7 Ancoragem dos transcritos no genoma de *V. unguiculata*

Para análise da distribuição cromossômica e da formação de clusters de cada superfamília de TEs no genoma recém-publicado do feijão-caupi, foram acessadas as coordenadas dos transcritos em relação ao genoma de referência [27] (Lonardi et al., 2019). A imagem da ancoragem dos transcritos nos cromossomos foi gerada através do pacote de visualização Circos v0.69 [63] (Krzywinski et al., 2009).

Declarações

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

O presente trabalho foi realizado com o apoio da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil), do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil) e da FACEPE (Fundação de Amparo à Pesquisa do Estado de Pernambuco, Brazil) através de bolsas e auxílio financeiro.

Referências citadas

[1] Gehan MA, Greenham K, Mockler TC, McClung CR. Transcriptional networks-crops, clocks, and abiotic stress. Curr Opin Plant Biol. 2015; doi: 10.1016/j.pbi.2015.01.004

[2] Pandey P, Ramegowda V, Senthil-Kumar M. Shared and unique responses of plants to multiple individual stresses and stress combniations: physiological and molecular mechanism. Front Plant Sci. 2015; doi: 10.3389/fpls.2015.00723

[3] TAKAHASHI, Fuminori; SHINOZAKI, Kazuo. Long-distance signaling in plant stress response. Current opinion in plant biology, v. 47, p. 106-111, 2019.

[4] Garg R, Varshney RK, Jain M. Molecular genetics and genomics of abiotic stress response. Front Plant Sci. 2014; doi: 10.3389/fpls.2014.00398

[5] Finatto T, de Oliveira AC, Chaparro C, da Maia LC, Farias DR, Woyann LG, Mistura CC, Soares-Bresolin AP, Llauro C, Panaud O, Picault N. Abiotic stress and genome dynamics: specific genes and transposable elements response to iron excess in rice. Rice (NY). 2015; doi: 10.1186/s12284-015-0045-6

[6] Gouveia BC, Calil IP, Machado JP, Santos AA, Fontes EP. Immune receptors and co-receptors in antiviral innate immunity in plants. Front Microbiol. 2017; doi: 10.3389/fmicb.2016.02139

[7] Kollist H, Zandalinas SI, Sengupta S, Nuhkat M, Kangasjärvi J, Mittler R. Rapid responses to abiotic stress: priming the landscape for the signal transduction network. Trends Plant Sci. 2019; doi:10.1016/j.tplants.2018.10.003

[8] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007; doi: 10.1038/nrg2165

[9] Vicient CM. Trnscriptional activity of transposable elements in maize. BMC Genomics. 2010; doi: 10.1186/1471-2164-11-601

[10] Domingues DS, Cruz GMQ, Metcalfe CJ, Nogueira FTS, Viventini R, Alves CS, Sluys MAV. Analysis of plant LTR-retrotransposons at fine-scale family level reveals individual molecular patterns. BMC genomics. 2012; doi: 10.1186/1471-2164-13-137

[11] Makarevitch I, Waters Aj, West Pt, Stizer M, Hirsch CN, Ross-Ibarra J, Springer NM. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet. 2015; doi: 10.1371/journal.pgen.1004915

[12] Biémont C. A brief history of the status of transposable elements: From junk DNA to mayor player in evolution. Genetics. 2010; doi: 10.1534/genetics.110.124180

[13] Testori A, Caizzi L, Cutrupi S, Friard O, de Bortoli M, Cora D, Caselle. The role of transposable elements in shaping the combinatorial interaction of transcription factors. BMC genomincs. 2012; doi: 10.1186/1471-2164-13-400

[14] Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017; doi: 10.1038/nrg.2016.139

[15] He ZH, Dong HT, Dong JX, Li DB, Ronald PC. The rice Rim2 transcript accumulates in response to Magnaporthe grisea and its predicted protein product shares similarity with TNP2-like proteins encoded by CACTA transposons. Mol Gen Genet. 2000; 264(1-2):2-10.

[16] Neumann P, Pozárková D, Macas J. Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. Plant Mol Biol. 2003; 53(3):399-410

[17] Gómez-Orte E, Vicient CM, Matínez-Izquierdo JA. Grande retrotransposons contain an accessory gene in the unusually long 3'-internal region that encodes a nuclear protein transcribed from its own promoter. Plant Mol Biol. 2013; doi: 10.1007/s11103-013-0019-2

[18] Zhao Y, Xu T, Shen CY, Xu GH, Chen SH, Song LZ, Li MJ, Wang LL, Zhu Y, Lv WT, Gong ZZ, Liu CM, Deng X. Identification of a retroelement from the resurrection plant Boea hygrometrica that confers osmotic and alkaline tolerance in Arabidopsis thaliana. Plos ONE. 2014; doi: 10.1371/journal.pone.0098098

[19] Negi P, Rai AN, Suprasanna P. Moving through the stressed genome: emerging regulatory roles for transposons in plant stress response. Front Plant Sci. 2016; doi:10.3389/fpls.2016.01448

[20] Wu Q, Smith NA, Zhang D, Zhou C, Wang MB. Root-Specific Expression of a Jacalin Lectin Family Protein Gene Requires a Transposable Element Sequence in the Promoter. Genes (Basel). 2018; doi: 10.3390/genes9110550.

[21] Lisch D, Bennetzen JL. Transposable element origins of epigenetic gene regulation. Curr Opin Plant Biol. 2011; doi: 10.1016/j.pbi.2011.01.003

[22] Silva AC, Santos DC, Teixeira-Junior DL, Silva PB, Santos RC,Siviero A. Cowpea: A Strategic Legume Species for Food Security and Health. InTech. 2018; doi: 10.5772/intechopen.79006

[23]

[24] Boukar O, Fatokun CA, Roberts PA, Abberton M, Huynh BL, Close TJ, Ehlers JD. Cowpea. Handbook of Plant Breeding. 2015; doi: 10.1007/978-1-4939-2797-5_7

[25] Daryanto S, Wang L, Jacinthe PA. Global synthesis of drought effects on food legume Production. PLoS One. 2015; doi: 10.1371/journal.pone.0127401

[26]

[27] Lonardi S, Muñoz-Amatriaín M, Liang Q, Shu S, Wanamaker SI, Lo S, Tanskanen J, Schulman AH, Zhu T, Luo MC, Alhakami H, Ounit R, Hasan AM, Verdier J, Roberts PA, Santos JRP, Ndeve A, Doležel J, Vrána J, Hokin SA, Farmer AD, Cannon SB, Close TJ. The genome of cowpea (*Vigna unguiculata* [L.] Walp.). Plant J. 2019; doi: 10.1111/tpj.14349

[28] Jesus-Pires C, Ferreira-Neto JRC, Bezerra-Neto JP, Kido EA, Pandolfi V, Wanderley-Nogueira AC, Binneck E, Pio-Ribeiro G, Pereira-Andrade GIM Sittolin, Freire-Filho F, Benko-Iseppon AM. Plant Thaumatin-like Proteins: Function, Evolution and Biotechnological Applications. Curr Protein Pept Sci. 2019; doi: 10.2174/1389203720666190318164905.

[29] Simon MV, Benko-Iseppon AM, Resende LV, Winter P, Kahl G. Genetic diversity and phylogenetic relationships in Vigna Savi germplasm revealed by DNA amplification fingerprinting (DAF). Genome. 2007; 50: 538-547.

[30] Spiaggia F, Carvalho R, Benko-Iseppon AM. Preliminary Molecular Characterization of cowpea (*Vigna unguiculata* (L.) Walp.) Accessions by DAF (DNA Amplification Fingerprinting). Gene Conserve. 2009; 8(34): 818-828

[31] Cho WK, Lian S, Kim SM, Seo BY, Jung JK, Kim KH. Time-Course RNA-Seq Analysis Reveals Transcriptional Changes in Rice Plants Triggered by Rice stripe virus Infection. PLoS One. 2015; 10(8):e0136736. doi: 10.1371/journal.pone.0136736

[32] Zhao D, Hamilton JP, Vaillancourt B, Zhang W, Eizenga GC, Cui Y, Jiang J, Buell CR, Jiang N. The unique epigenetic features of Pack-MULEs and their impact on chromosomal base composition and expression spectrum. Nucleic Acids Res. 2018; doi: 10.1093/nar/gky025

[33] Göbel U, Arce AL, He F, Rico A, Schmitz G, de Meaux J. Robustness of Transposable Element Regulation but No Genomic Shock Observed in Interspecific Arabidopsis Hybrids. Genome Biol Evol. 2018; doi: 10.1093/gbe/evy095.

[34] Arabidopsis genome initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000; 408(6814):796-815

[35]

[36] Herbers K, Takahata Y, Melzer M, Mock HP, Hajirezaei M and Sonnewald U. Regulation of carbohydrate partitioning during the interaction of potato virus Y with tobacco. Mol. Plant Path. 2000; 1(1):51-59

[37] Scharte J, Schön H and Weis E. Photosynthesis and carbohydrate metabolism in tobacco leaves during an incompatible interaction with Phytophthora nicotianae. Plant, Cell and Env. 2006; 28:1421-1436

[38] Ferreira Neto JR, Pandolfi V, Guimaraes FC, Benko-Iseppon AM, Romero C, Silva RL, Rodrigues FA, Abdelnoor RV, Nepomuceno AL, Kido EA. Early transcriptional response of soybean contrasting accessions to root dehydration. PLoS One. 2013; doi: 10.1371/journal.pone.0083466

[39] Gao C, Xiao M, Ren X, Hayward A, Yin J, Wu L, Fu D, Li J. Characterization and functional annotation of nested transposable elements in eukaryotic genomes. Genomics. 2012; doi: 10.1016/j.ygeno.2012.07.004

[40] Bennetzen J e Wang H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. Annu Rev Plant Biol. 2014; 65:505-30

[41] Le TN, Schumann U, Smith NA, Tiwari S, Au PC, Zhu QH, Taylor JM, Kazan K, Llewellyn DJ, Zhang R, Dennis ES, Wang MB. DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in Arabidopsis. Genome Biol. 2014; doi: 10.1186/s13059-014-0458-3.

[42] Ito H, Kim JM, Matsunaga W, Saze H, Matsui A, Endo TA, Harukawa Y, Takagi H, Yaegashi H, Masuta Y, Masuda S, Ishida J, Tanaka M, Takahashi S, Morosawa T, Toyoda T, Kakutani T, Kato A, Seki M. A Stress-Activated Transposon in Arabidopsis Induces Transgenerational Abscisic Acid Insensitivity. Sci Rep. 2016; doi: 10.1038/srep23181.

[43] Nejat N e Mantri N. Plant Immune System: Crosstalk Between Responses to Biotic and Abiotic Stresses the Missing Link in Understanding Plant Defence. Curr Issues Mol Biol. 2017; doi: 10.21775/cimb.023.001

[44] Kim KN. Stress responses mediated by the CBL calcium sensors in plants. Plant Biotechnol Rep. 2013; doi: 10.1007/s11816-012-0228-1

[45] Wei B, Liu H, Liu X, Xiao Q, Wang Y, Zhang J, Hu Y, Liu Y, Yu G, Huang Y. Genome-wide characterization of non-reference transposons in crops suggests non-random insertion. BMC Genomics. 2016; doi: 10.1186/s12864-016-2847-3.

[46] Nakashima K, Abe J, Kanazawa A. Chromosomal distribution of soybean retrotransposon SORE-1 suggests its recent preferential insertion into euchromatic regions. Chromosome Res. 2018; doi: 10.1007/s10577-018-9579-y

[47] Xu Y, Du J. Young but not relatively old retrotransposons are preferentially located in gene-rich euchromatic regions in tomato (Solanum lycopersicum) plants. Plant J. 2014; doi: 10.1111/tpj.12656

[48] Houllou-Kido LM, Crovella S, Benko-Iseppon AM. Identification of plant protein kinases in response to abiotic and biotic stresses using SuperSAGE. Current Protein and Peptide Science. 2011; 12(7): 643-656.

[49] Kulcheski FR, de-Oliveira LF, Molina LG, Almerão MP, Rodrigues FA, Marcolino J, Barbosa JF, Stolf-Moreira R, Nepomuceno AL, Marcelino-Guimarães FC et al. Identification of novel soybean microRNAs involved in abiotic and biotic stresses. BMC Genomics. 2011; doi: 10.1186/1471-2164-12-307

[50] Rodrigues FA, Marcolino-Gomes J, de Fátima Corrêa Carvalho J, do Nascimento LC, Neumaier N, Farias JRB, Carazzolle MF, Marcelino FC, Nepomuceno AL. Subtractive libraries for prospecting differentially expressed genes in the soybean under water deficit. Genet Mol Biol. 2012; doi: 10.1590/S1415-47572012000200011

[51] Hoagland DR, Arnon DI, editors. The water culture method for growing plants without soils. Circular 347. Berkeley: California Agricultural Experimental Station; 1950. p.347

[52] Fehr WR, Caviness CF, Burmood DT, Pennington JS. Stage of development descriptions for soybeans, Glycine max (L.) Merrill. Crop Sci. 1971; 11:929-931

[53] Bourgey M, Dali R, Eveleigh R, Chen KC, Letourneau L, Fillon J, Michaud M, Caron M, Sandoval J, Lefebvre F et al. GenPipes: an open-source framework for distributed and scalable genomic analyses. bioRxiv. 2018; doi: 10.1101/459552

[54] Deluca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics. doi: 10.1093/bioinformatics/bts196

[55] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; doi: 10.1093/bioinformatics/btu170

[56] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; doi: 10.1093/bioinformatics/bts635

[57] Van-der-Auwera GA, Carneiro MO, Hartl C, Poplin R, Del-Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; doi: 10.1002/0471250953.bi1110s43

[58] Smit, AFA, Hubley, R, Green, P. RepeatMasker Open-4.0. 2015. http://www.repeatmasker.org. Accessed 20 Dez 2018.

[59] Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015; doi: 10.1186/s13100-015-0041-9

[60] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; doi: 10.1038/nbt.1621

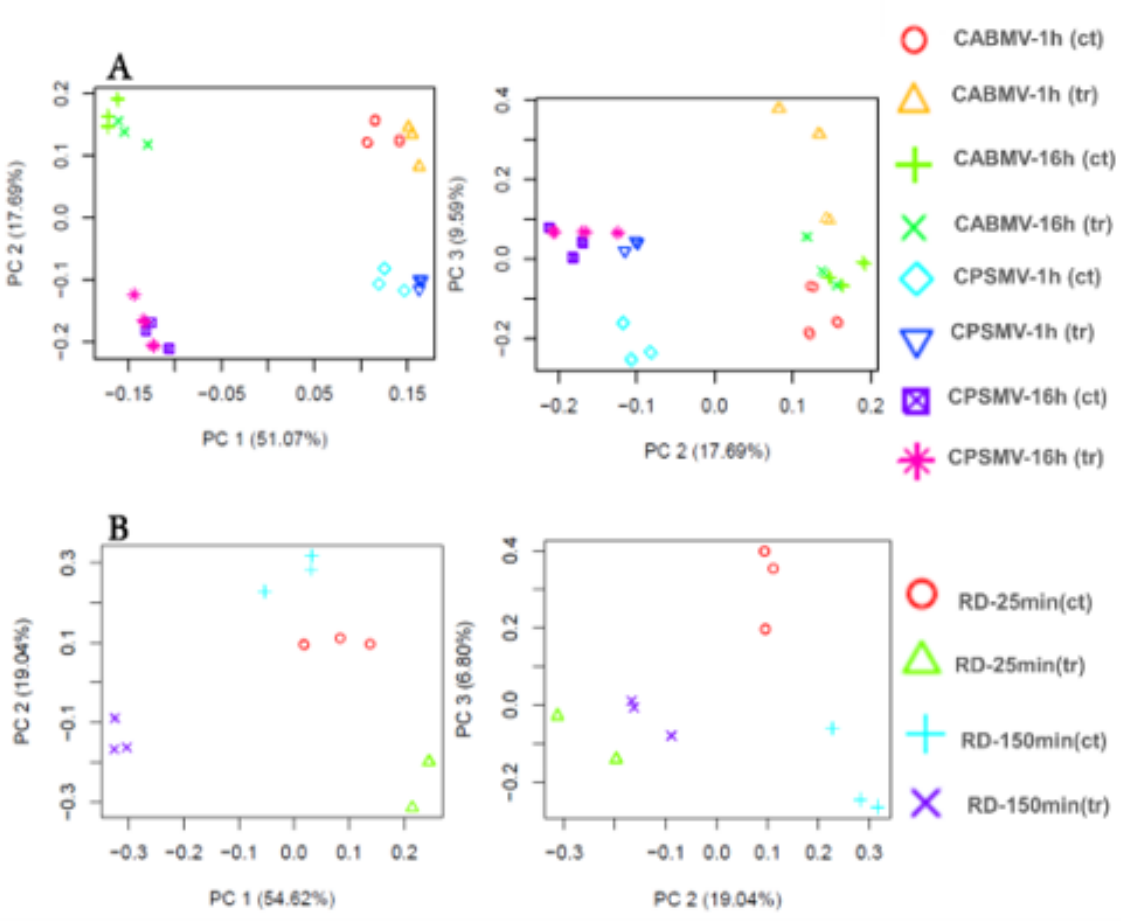
[61] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013; doi: 10.1038/nbt.2450

[62] Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010; doi: 10.1186/gb-2010-11-2-r14

[63] Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; doi: 10.1101/gr.092759.109

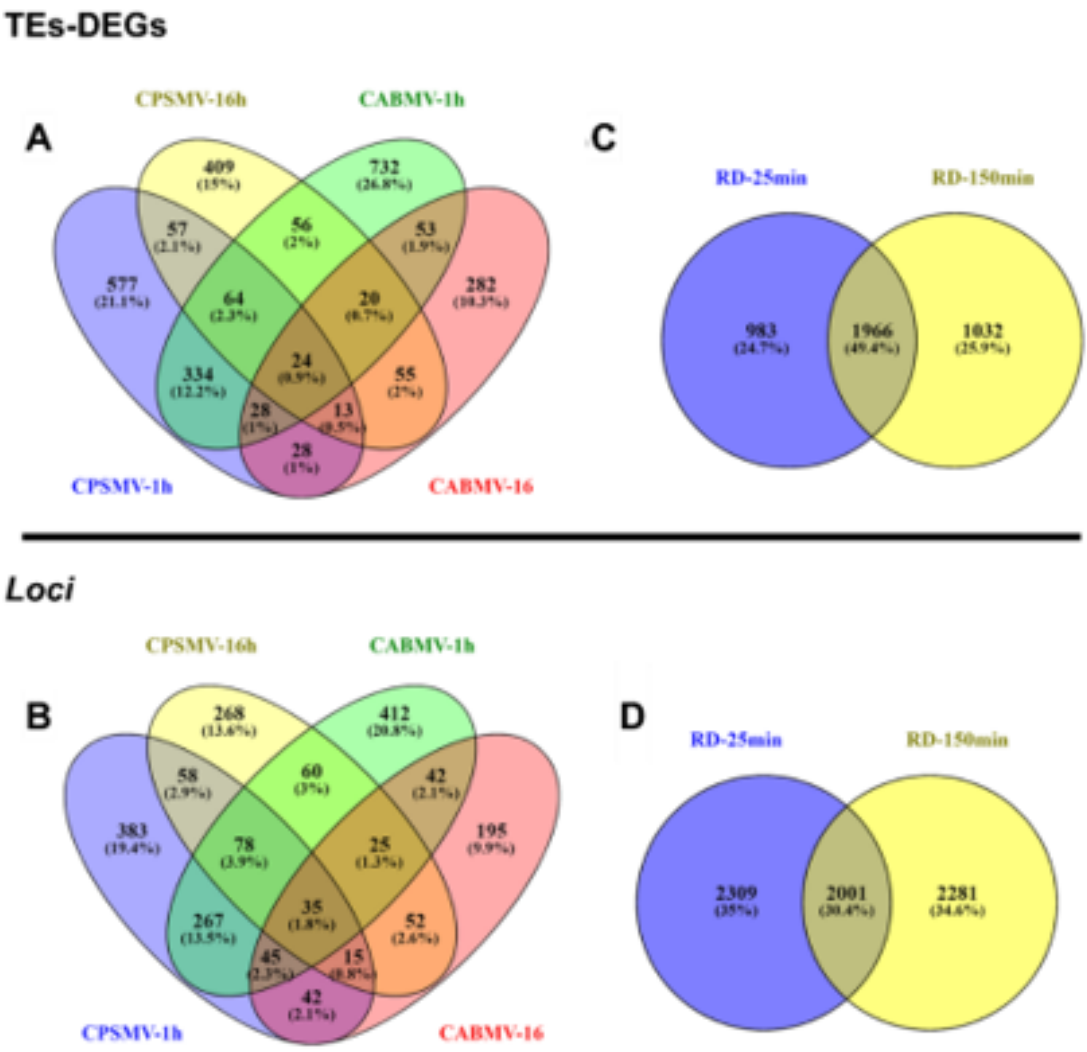
Legendas das Figuras

Figura 1 - Análise de componentes principais (PCA) dos dados do transcriptoma (RNA-Seq) do feijão-caupi (Vigna unguiculata)



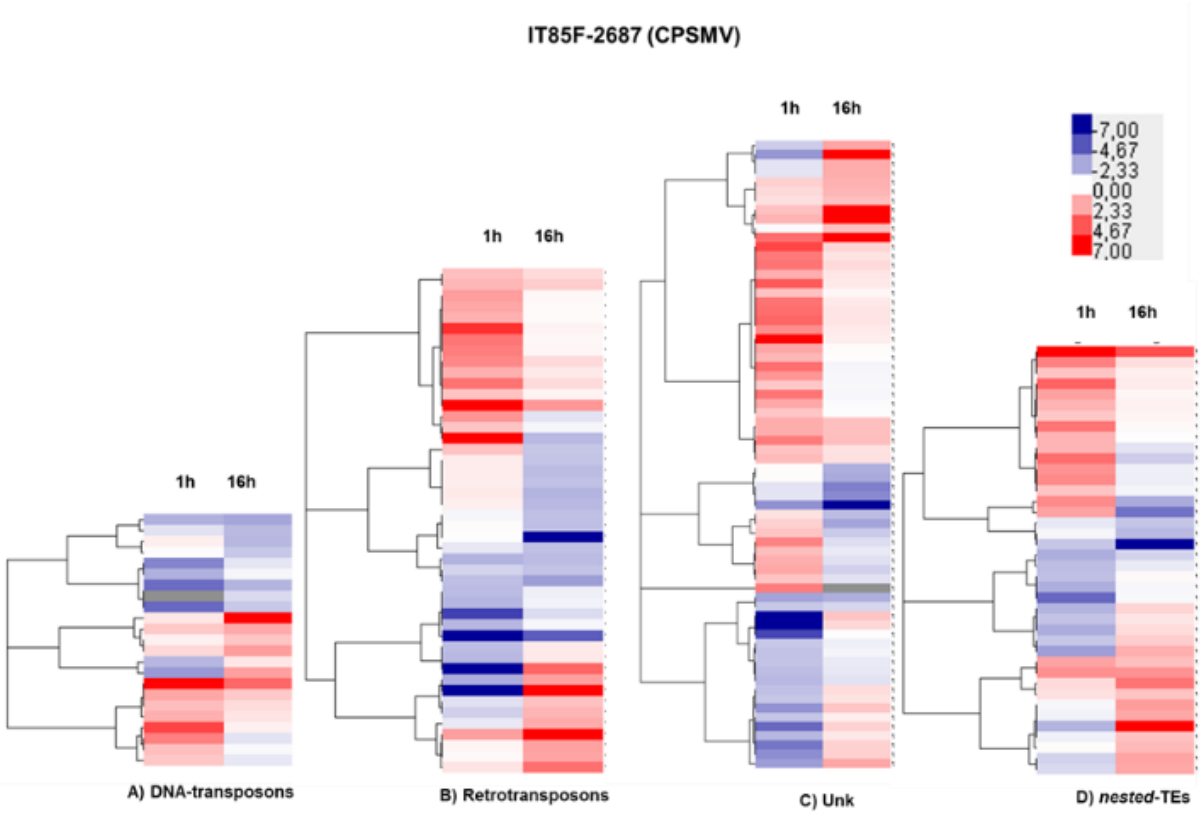
Fonte: A autora (2019). A) A PCA apresenta a relação entre os eixos e dados dos acessos IT85F-2687 inoculado com Cowpea Severe Mosaic Virus (CPSMV) e BR14-Mulato inoculado com Cowpea Aphid-Born Mosaic Virus (CABMV) comparando as triplicatas biológicas de controle (ct) e tratamentos (tr) nos diferentes tempos de estresse. B) A PCA apresenta a correlação entre dados controle (ct) e tratamento do acesso Pingo de Ouro sob desidratação radicular (tr), distinguindo ambos os acessos, bem como os tempos de estresse.

Figura 2 - Total de transcritos de elementos transponíveis diferencialmente expressos (TE-DEGs) no transcriptoma (RNA-Seq) de Vigna unguiculata e total de loci de origem dos TE-DEGs.



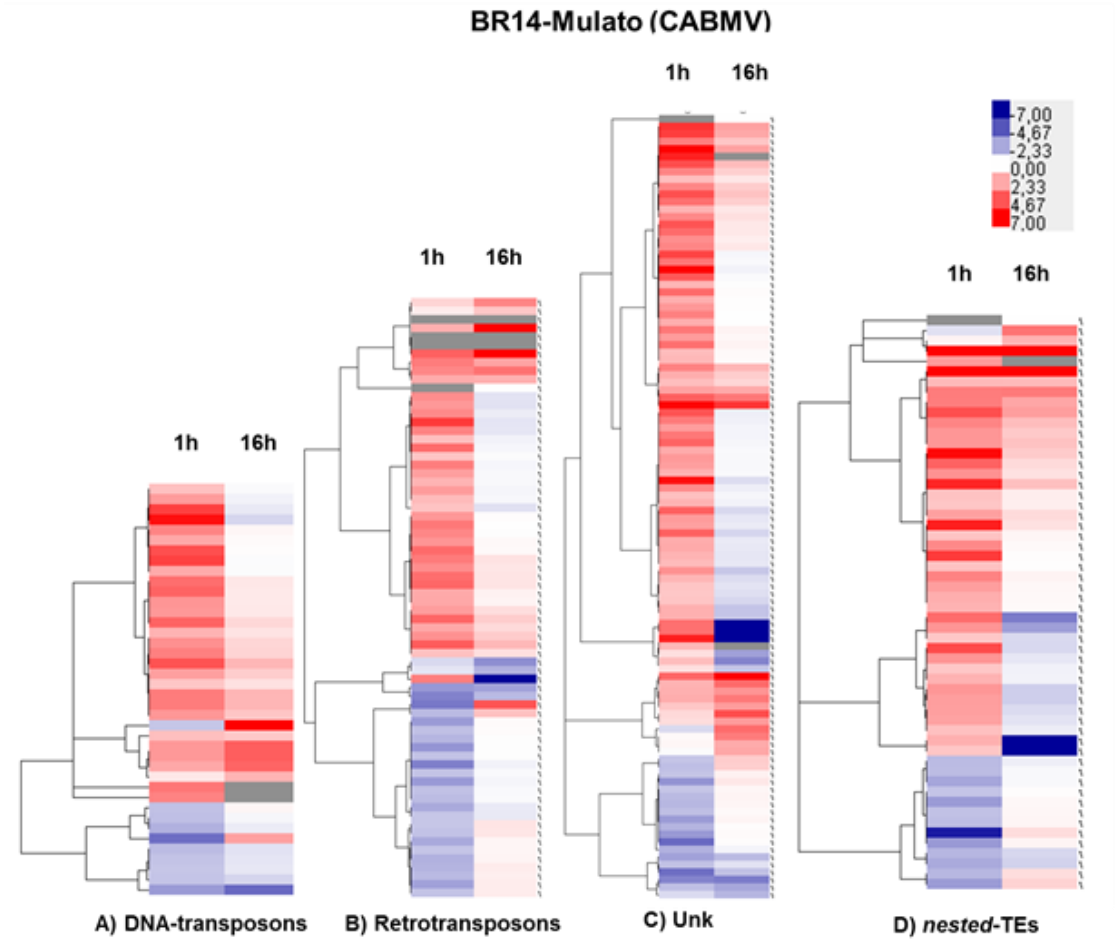
Fonte: A autora (2019). A) O total de TEs-DEGs após 1 h e 16 h de inoculação com vírus para os dois ensaios: acesso IT85F-2687 infectado com *Cowpea Severe Mosaic Virus* (CPSMV) e acesso BR14-Mulato infectado com *Cowpea Aphid-Born Mosaic Virus* (CABMV). B) O total de *loci* de origem para IT85F-2687-CPSMV e BR14-Mulato-CABMV. C) Total de TE-DEGs no acesso Pingo de Ouro submetido à desidratação radicular (RD) durante 25 e 150 min. D) Total de *loci* de origem para Pingo de Ouro submetido à RD em ambos os tempos de estresse.

Figura 3 - Heatmap de genes/isoformas de elementos transponíveis (TEs) diferencialmente expressos no transcriptoma do acesso IT85F-2687 infectado com Cowpea Severe Mosaic Virus (CPSMV).



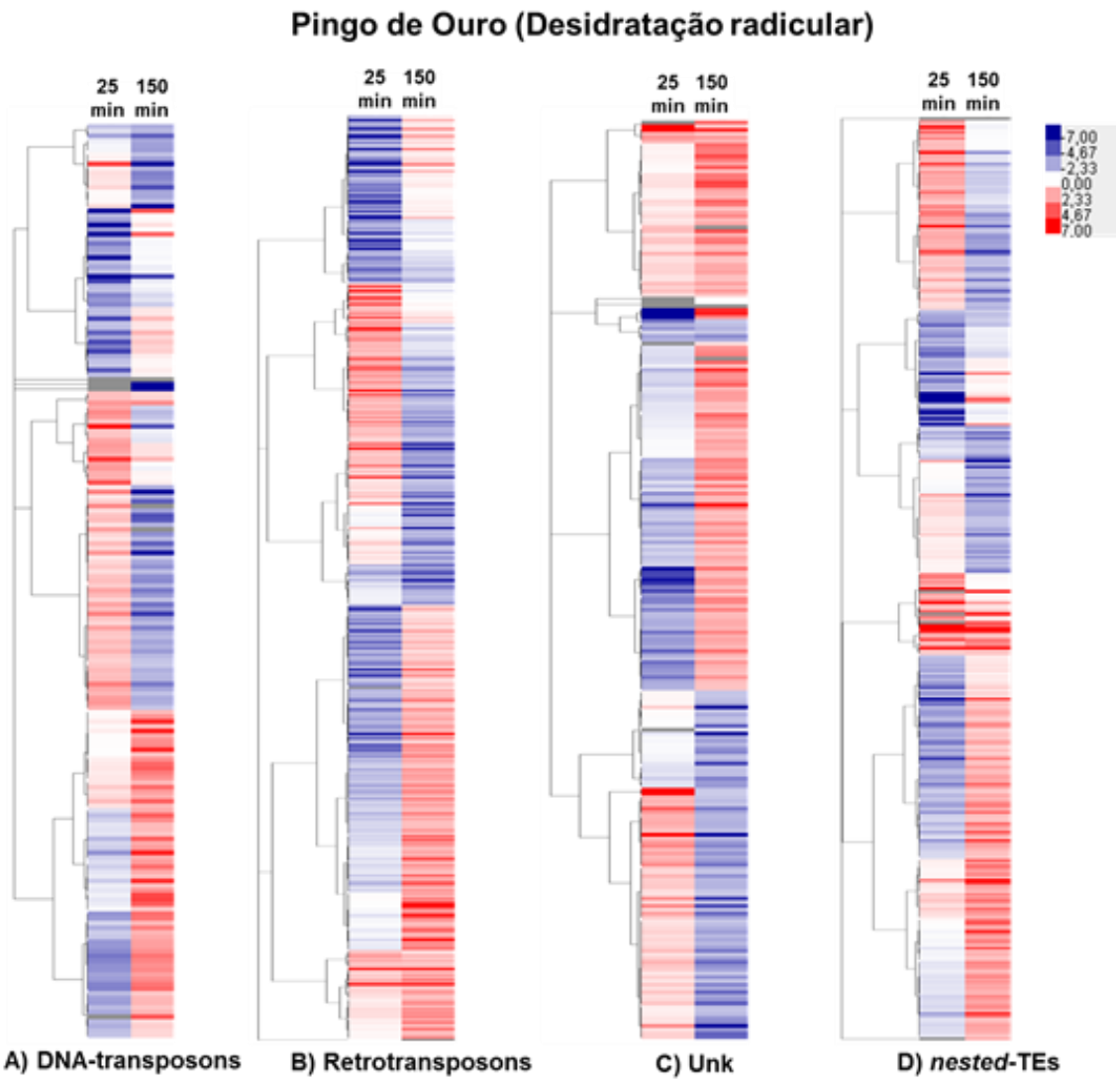
Fonte: A autora (2019). Comparativo do nível de expressão após 1 e 16 h de inoculação do vírus nas folhas. Em vermelho são representadas sequências induzidas, em azul as reprimidas e em cinza sequências que são exclusivamente moduladas no tratamento, sendo ligado ou desligado em relação ao controle . A) Sequências de DNA-transposon; B) Retrotransposons; C) TEs de classificação desconhecida (Unk); D) Nested-TEs

Figura 4 - Heatmap de genes/isoformas diferencialmente expressas de elementos transponíveis (TEs) no transcriptoma do acesso BR14-Mulato infectado com Cowpea Aphid-Born Mosaic Virus (CABMV).



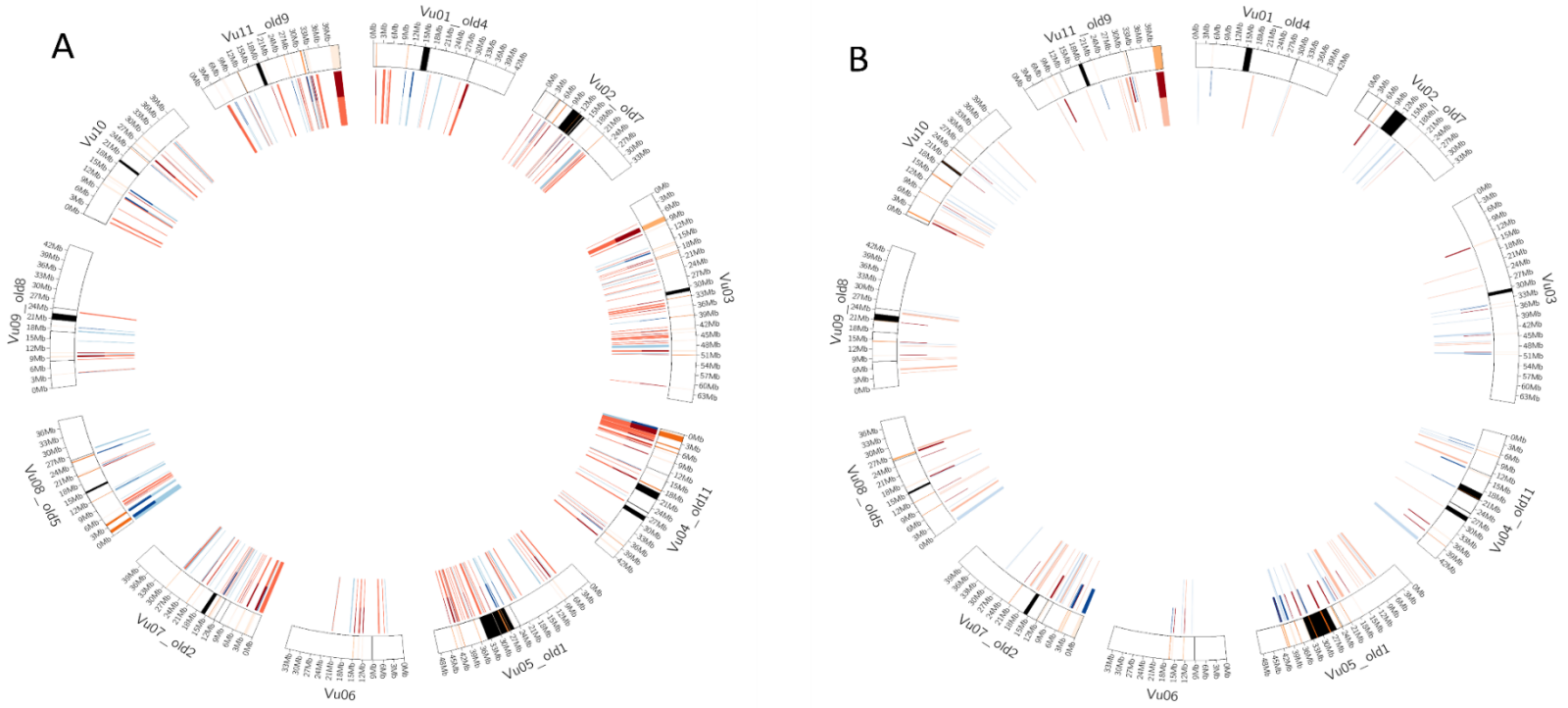
Fonte: A autora (2019). Comparativo do nível de expressão após 1 e 16 h de inoculação do vírus nas folhas. Em vermelho são representadas sequências induzidas, em azul as reprimidas e em cinza sequências que são exclusivamente moduladas no tratamento em relação ao controle. A) Sequências de DNA-transposon; B) Retrotransposons; C) TEs de classificação desconhecida (Unk); D) Nested-TEs.

Figura 5 - : Heatmap de genes/isoformas diferencialmente expressas de elementos transponíveis (TEs) no transcriptoma do acesso 'Pingo de Ouro' submetido à desidratação radicular (RD).



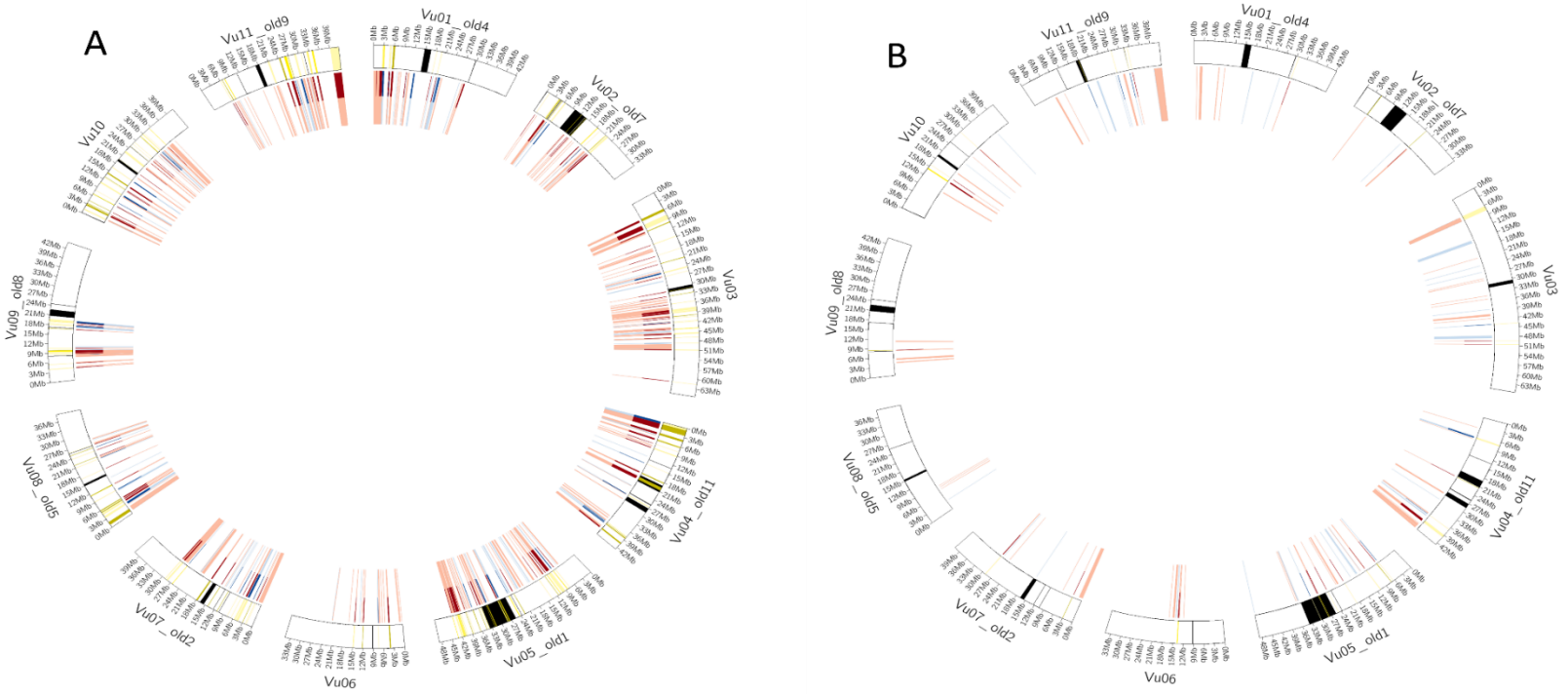
Fonte: A autora (2019). Comparativo do nível de expressão após 25 e 150 min de desidratação radicular. Em vermelho são representadas sequências induzidas, em azul as reprimidas e em cinza sequências que são exclusivamente moduladas no tempo em questão. A) Sequências de DNA-transposon; B) Retrotransposons; C) TEs de classificação desconhecida (Unk); D) Nested-TEs.

Figura 6 - Ancoragem de TEs diferencialmente expressos (DEGs) nas folhas do acesso IT85F-2687 após inoculação com Cowpea Severe Mosaic Virus (CPSMV).



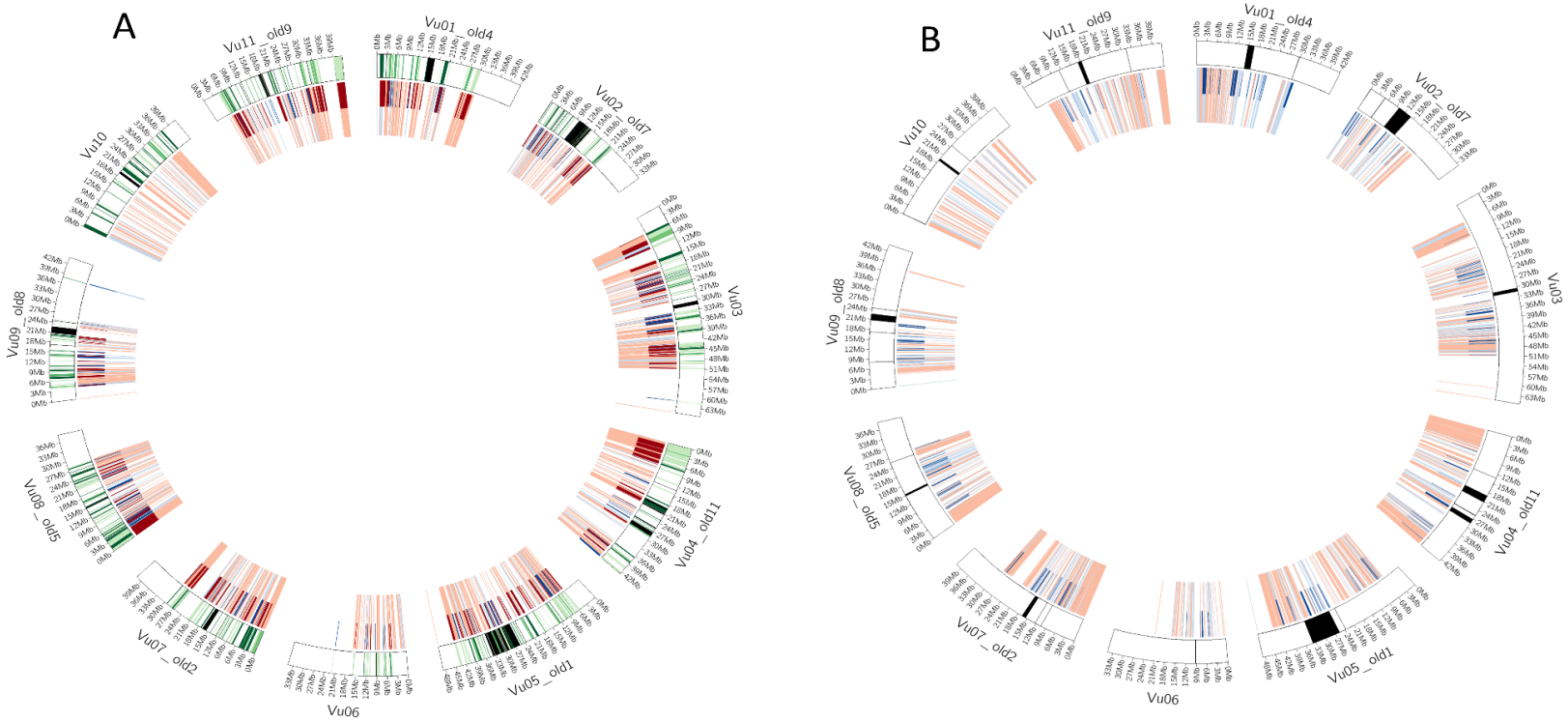
Fonte: A autora (2019). Ancoragem de TEs diferencialmente expressos (DEGs) nas folhas do acesso IT85F-2687 infectado com Cowpea Severe Mosaic Virus (CABMV). (A) acesso IT85F-2687, 1 h após inoculação do vírus. (B) acesso IT85F-2687, 16 h após inoculação do vírus. Nas figuras, mais externamente encontram-se as escalas em Mb, Em seguida os cromossomos com centrômeros indicados em preto. Nos cromossomos encontram-se marcadas as posições de ancoragem dos TEs diferencialmente expressos (em laranja). Na região mais interna, encontra-se a indicação da modulação dos TEs significativamente reprimidos (marcas azuis) e induzidos (marcas vermelhas). Por fim, mais internamente (circunferência interna) encontram-se as marcações de genes associados à periferia dos TEs marcados. Em regiões onde TEs não são observados/ancorados, os genes apresentam-se como uma única linha que ocupa toda a extensão do 3o e 4o anéis.

Figura 7 - Ancoragem de TEs diferencialmente expressos (DEGs) nas folhas do acesso BR14-Mulato infectado com Cowpea Aphid-Born Mosaic Virus (CABMV).



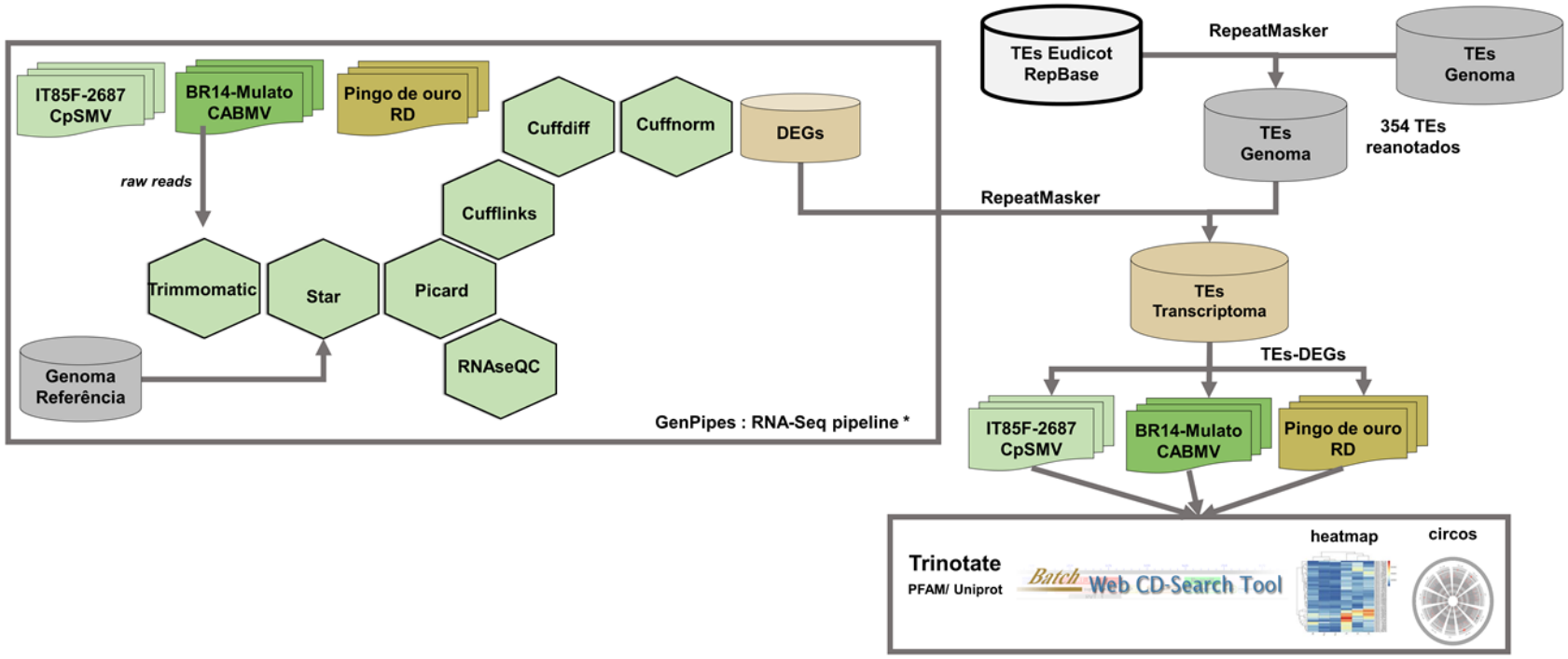
Fonte: A autora (2019). Ancoragem de TEs diferencialmente expressos (DEGs) nas folhas do acesso BR14-Mulato infectado com Cowpea Aphid-Born Mosaic Virus (CABMV). (A) acesso BR14-Mulato 1 h após inoculação do vírus. (B) acesso BR14-Mulato 16 h após inoculação do vírus. Nas figuras, mais externamente encontram-se as escalas em Mb, Em seguida os cromossomos com centrômeros indicados em preto. Nos cromossomos encontram-se marcadas as posições de ancoragem dos TEs diferencialmente expressos (em amarelo). Na região mais interna, encontra-se a indicação da modulação dos TEs significativamente reprimidos (marcas azuis) e induzidos (marcas vermelhas). Por fim, mais internamente (circunferência interna) encontram-se as marcações de genes associados à periferia dos TEs marcados. Em regiões onde TEs não são observados/ancorados, os genes apresentam-se como uma única linha que ocupa toda a extensão do 3º e 4º anéis.

Figura 8 - Ancoragem de TEs diferencialmente expressos (DEGs) no acesso Pingo de Ouro sob desidratação radicular (RD).



Fonte: A autora (2019). Ancoragem de TEs diferencialmente expressos (DEGs) nas folhas do acesso Pingo de Ouro sob desidratação radicular (RD). (A) acesso Pingo de Ouro, 25 min de desidratação. (B) Acesso Pingo de Ouro, 150 min de desidratação. Nas figuras, mais externamente encontram-se as escalas em Mb. Em seguida os cromossomos com centrômeros indicados em preto. Nos cromossomos encontram-se marcadas as posições de ancoragem dos TEs diferencialmente expressos (em verde). Na região mais interna, encontra-se a indicação da modulação dos TEs significativamente reprimidos (marcas azuis) e induzidos (marcas vermelhas). Por fim, mais internamente (circunferência interna) encontram-se as marcações de genes associados à periferia dos TEs marcados. Em regiões onde TEs não são observados ancorados, os genes apresentam-se como uma única linha que ocupa toda a extensão do 3o e 4o anéis.

Figura 9 - Fluxograma das etapas de montagem, anotação e análise funcional do presente estudo.



Fonte: A autora (2019). As reads brutas foram processadas usando o pipeline RNA-Seq do GenPipes contra o genoma de referência disponível no Phytozome. Para tal foram usados o trimmomatic para trimagem, o STAR para alinhamento com o genoma, o picard para mapeamento de *reads* duplicadas, o RNAseQC para acessa as métricas de qualidade e o pacote cufflinks (cufflinks, cuffdiff, cuffnorm) para montagem e calculo da expressão diferencial. Com o transcriptoma montado o banco foi contrastado com o os TEs do genoma. Antes da análise com o RepeatMasker (RM) os TEs do genoma foram reanotados contra o banco de eudicotiledôneas do RepBase. Usando o banco de TEs de *V. unguiculata* atualizado o RM foi usado para acessar TEs expressos. Foram considerados diferencialmente expressos (TE-DEGs) aqueles que apresentavam p-value < 0,05. Em seguida os TE-DEGs foram anotadas e tiveram, os *heatmap* gerados e as sequências ancoradas no genoma de referência para identificação de sua distribuição cromossômica.

5 DISCUSSÃO GERAL

A publicação científica é um canal-chave para a distribuição da informação na ciência (REBHOLZ-SCHUHMANN; OELLRICH; HOEHNDORF, 2012). A mineração de texto, por sua vez, é uma ferramenta facilitadora que busca integrar toda a informação dispersa na literatura. Atualmente, a maioria das ferramentas está disponível apenas online e frequentemente voltada para temas específicos (BACHMAN; GYORI; SORGER, 2018; LI, M. et al., 2019; PRICE; ARKIN, 2017; SUN, WANG; LI, 2017). O LAITOR4HPC e o STRING são as únicas ferramentas com a possibilidade de função programática e paralelização, embora com focos e abordagens diferentes (JENSEN et al., 2009; SNEL et al., 2000; SZKLARCZYK et al., 2017).

Uma das primeiras etapas de mineração de texto (IR) constituía a seleção de literatura relevante (BARBOSA-SILVA et al., 2010, 2011; REBHOLZ-SCHUHMANN; OELLRICH; HOEHNDORF, 2012). Contudo, o LAITOR4HPC, apresentado no presente volume, ultrapassa a barreira das análises com número limitado de artigos, eliminando a etapa de seleção de literatura relevante, permitindo que toda a literatura disponível até o momento seja analisada em função de um tema, proteína ou espécie. Por consequência, torna mais fácil acessar toda a informação existente publicada.

O LAITOR4HPC é uma ferramenta flexível que permite abordagens variadas para dados de qualquer dimensão. Entretanto, a biologia de sistemas pretende subsidiar a compreensão da dinâmica de um organismo, realizando de um estudo multidimensional (KITANO, 2002a). Neste sentido, nenhuma das ferramentas *online* e *offline* é, até o momento, capaz de identificar Elementos Transponíveis (TEs). A etapa de “tagueamento” ou identificação de termos é a etapa mais complexa, devido à falta de padronização na nomenclatura (KRALLINGER; VALENCIA, 2005; MIKA; ROST, 2004). Adicionalmente, apesar do aumento no número de estudos que defendem a abrangente influência dos TEs nos genomas e transcriptomas (CHO, 2018; LE et al., 2014; QIONG et al., 2018), ainda há um grande desafio devido às ambiguidades e falta de padronização em sua nomenclatura (dados não mostrados).

Sem o auxílio da mineração de texto, a bioinformática torna-se o principal meio de estudo destes componentes genéticos, o qual é favorecido pelo uso das tecnologias NGS no sequenciamento e ressequenciamento de organismos de interesse (CHO, 2018), como proposto no presente trabalho para o feijão-caupi (*V. unguiculata*). Embora sejam poucos os estudos de RNA-Seq que focam em TEs, dados transcriptômicos das plantas sob infeção viral e desidratação radicular estão disponíveis para arroz (CHO et al., 2015) e *A. thaliana* (ZHAO, D., 2018), respectivamente. Nestes estudos, o arroz apresenta TEs constituindo entre 13-15% de todos os DEGs significativos, enquanto *A. thaliana* tem apenas 2-4%. Em feijão-caupi, houve uma variação entre 7-18% dependendo do acesso, do estresse ou do tecido analisado. Os resultados dos três diferentes acessos reforçaram o comportamento específico dos elementos transponíveis, tendo pouca ou nenhuma correlação entre as isoformas de cada acesso e tempo de estresse.

Entre os elementos anotados, *Copia*, *hAT* e *nested*-TEs se destacaram como os mais representativos, apesar de terem sido observados transcritos para mais 10 superfamílias. Os dados de TEs diferencialmente expressos, especialmente os transcritos induzidos, foram relacionados à resposta ao estresse, tanto usando termos diretamente ligados à tolerância (como resposta a estímulo biótico e resposta a privação de água) ou à resistência, quanto termos relacionados à transdução de sinais e regulação dos genes. Dentre estes se destacam: sequências de ligação com fatores de transcrição, resposta hormonal, ligação, transporte iônico e metabolismo de carboidratos (KOLLIST et al., 2018; NEJAT; MANTRI, 2017; SCHARTE; SCHÖN; WEIS, 2005).

Diante da modulação observada para os TEs de feijão-caupi, sugere-se que estes possuam papel complementar na resposta aos estresses estudados. Sua expressão é específica e relacionada a respostas diretas e indiretas ao estresse.

6 CONCLUSÕES

A atualização do LAITOR para LAITOR4HPC diminui significativamente o tempo computacional da análise de mineração de texto, mantendo a eficiência de resultados para estudos dos dicionários existentes e bem estabelecidos.

O tempo de análise do LAITOR4HPC varia de acordo com os recursos computacionais disponíveis, sendo crucial uma curadoria manual dos dados ao final das análises e resultados de interesse.

O LAITOR4HPC permite avaliação dos dados por diversas abordagens como avaliação de interações pouco ou muito descritas, enriquecimento ou construção de vias de interação.

Para soja, os dados mostraram haver um maior foco em estudos envolvem as proteínas cloroplastidiais.

A busca pelas interações relacionadas à resposta de estresse biótico em plantas evidenciam que a PR-1 (*Pathogenesis Related – 1*) envolvida na defesa vegetal contra patógenos, é a proteína mais estudada. Sugere-se um esforço maior para o estudo de outras proteínas (incluindo outras classes de proteínas PR) e consequente esclarecimento de outras interações no intuito de expandir a compreensão acerca deste tema.

Os TEs diferencialmente expressos identificados para os acessos IT85F-2687 e BR14-Mulato com inóculo viral possuem expressão estresse, tempo e acesso específicos, com um total de TEs menor após 16 h de estresse.

Em raízes de Pingo de Ouro sob desidratação o total de TEs diferencialmente expresso tende a aumentar na resposta tardia, acompanhado do aumento na proporção de elementos reprimidos.

A quantidade, algumas vezes exorbitante, de transcritos que apresentam similaridade com mais de 10 TEs da mesma superfamília é um indício de atividade relativamente recente, dada a conservação entre sequências.

As anotações funcionais sugerem forte relação com resposta a estresse, incluindo similaridade com sequências de ligação a promotores. Tais evidências, corroboradas por outras publicações, evidenciam a necessidade da construção de um dicionário de TEs para que haja a inclusão destes na abordagem de

mineração de texto, viabilizando mensurar a dimensão da influência destes elementos nos genomas e transcriptomas dos organismos onde são descritos.

REFERÊNCIAS

AKIOBODE, S.; MAREDIA, M. **Global and regional trends in production, trade and consumption of food legume crops.** Michigan: Michigan State University, 2012. Disponível em: https://ageconsearch.umn.edu/record/136293. Acesso em: 25 jun. 2019.

ANDRADE, M. A.; BORK, P. Automated extraction of information in molecular biology. **FEBS Letters**, v. 476, n. 1-2, p.12-17, 2000. Disponível em: https://febs.onlinelibrary.wiley.com/doi/epdf/10.1016/S0014-5793%2800%2901661-6. Acesso em: 12 jul. 2019.

ARABIDOPSIS GENOME INITIATIVE. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, v. 408, p. 796-815, 2000. Disponível em: https://www.nature.com/articles/35048692.pdf. Acesso em: 21 jul. 2019.

BACHMAN, J. A.; GYORI, B. M.; SORGER, P. K. FamPlex: a resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining. **BMC Bioinformatics**, v. 19, n. 248, p. 1-14, 2018. Disponível em: https://doi.org/10.1186/s12859-018-2211-5. Acesso em: 18 jun. 2019.

BARBOSA-SILVA, A.; *et al*. LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships. **BMC Bioinformatics**, v. 11, n. 70, p. 1-10, 2010. Disponível em: https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-11-70.pdf. Acesso em: 20 jun. 2019

BARBOSA-SILVA, A. *et al*. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. **BMC Bioinformatics**, v. 12, n. 435, p. 1-9, 2011. Disponível em: https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-12-435.pdf. Acesso em: 07 jun. 2019

BASTOS, E. A. **A cultura do feijão-caupi no Brasil.** Teresina: Embrapa Meio-Norte, 2016. Disponível em: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1065493/a-cultura-do-feijao-caupi-no-brasil. Acesso em: 04 ago. 2019.

BENCHIMOL, R. L. *et al*. *Pseudocercospora cruenta* na cultivar de feijão-caupi BRS Novaera no estado do Pará. **Biota Amazônia**, Macapá, v. 7, n. 4, p. 60-62, 2017. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/171842/1/P.-cruenta-BRS-Novaera-BIOTA.pdf. Acesso em: 10 jun. 2019

BENKO-ISEPPON, A. M. *et al*. Mendel e suas exceções à luz das ômicas e da biologia de sistemas. *In*: ARAGÃO, F. J. L.; MOREIRA, J. R. **Mendel:** 150 anos depois. Brasília: Embrapa Edições, 2017. p.

BENNETZEN, J. L.Transposable element contributions to plant gene and genome evolution. **Plant Molecular Biology**, v. 42, n. 1, p. 251–269, 2000. Disponível em: https://web.nmsu.edu/~plantgen/supplemental_reading_files/fulltext.pdf. Acesso em: 01 ago. 2019.

BENNETZEN, J. L.; WANG, H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. **Annual Review of Plant Biology**, v. 65, p. 505-530, 2014. Disponível em: https://pubmed.ncbi.nlm.nih.gov/24579996/. Acesso em: 15 ago. 2019.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, 2014. Disponível em: https://doi.org/10.1093/bioinformatics/btu170. Acesso em: 08 jun. 2019.

BORGES, F. *et al*. Transposon-derived small RNAs triggered by miR845 mediate genome dosage response in Arabidopsis. **Nature Genetics**, v. 50, n. 2, p. 186-192, 2018. Disponível em: https://www.nature.com/articles/s41588-017-0032-5. Acesso em: 22 jul. 2019.

BOUKAR, O. *et al.* Cowpea. *In:* DE RON, A. M. (org.) **Grain Legumes**. 10. ed. New York: Springer, 2015. p 219–250.

BOURGEY, M. *et al*. GenPipes: an open-source framework for distributed and scalable genomic analyses. **GigaScience**, v. 8, n. 6, p. 1-11, 2019. Disponível em: https://doi.org/10.1093/gigascience/giz037. Acesso em: 15 jul. 2019.

BUNDOCK, P.; HOOYKAAS, P. An Arabidopsis hAT-like transposase is essential for plant development. **Nature**, v. 436, n. 7048, p. 282-284, 2005. Disponível em: https://www.nature.com/articles/nature03667. Acesso em: 18 jun. 2019.

CAVRAK, V. *et al*. How a retrotransposon exploits the plant's heat stress response for its activation. **PLOS Genetics**, v. 10, n. 1, p. 1-12, jan. 2014. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3907296/. Acesso em: 07 jun. 2019

CHEN, X. *et al*. CGKB: an annotation knowledge base for cowpea (*Vigna unguiculata* L.) methylation filtered genomic genespace sequences. **BMC Bioinformatics**, v. 8, n. 129, p. 1-9, 2007. Disponível em: https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-8-129.pdf. Acesso em: 17 ago. 2019.

CHÉNAIS, B. *et al*.The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. **Gene**, v. 509, n. 1, p. 7-15, nov. 2012. Disponível em: https://doi.org/10.1016/j.gene.2012.07.042. Acesso em: 04 ago. 2019.

CHO, J. Transposon-Derived Non-coding RNAs and Their Function in Plants. **Frontiers in Plant Science**, v. 9, n. 600, p. 1-6, maio 2018. Disponível em: https://www.readcube.com/articles/10.3389/fpls.2018.00600. Acesso em: 16 jul. 2019.

CHO, W. K. *et al*. Time-Course RNA-Seq Analysis Reveals Transcriptional Changes in Rice Plants Triggered by Rice stripe virus Infection. **PloS One**, v. 10, n. 8, p. 1-20, 25 ago. 2015. Disponível em: https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0136736&type=printable. Acesso em: 17 ago. 2019.

CNPAF – EMBRAPA. **Contribuição do caupi (***Vigna unguiculata* (L.) Walp**), na produção e área colhida de feijão no Brasil, de 1985 a 2020.** Santo Antônio de Goiás: Embrapa Arroz e Feijão, 2021. Disponível em: http://www.cnpaf.embrapa.br/socioeconomia/docs/arroz/contribuicaodocaupi.htm. Acesso em: 08 fev. 2022.

COSTA, V. *et al*. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. **European Journal of Human Genetics**, v. 21, n. 2, p.134-142, 2013. Disponível em: https://www.nature.com/articles/ejhg2012129.pdf. Acesso em: 16 jul. 2019.

CÚRCIO, M. J.; DERBYSHIRE, K. M. The outs and Ins of transposition: From Mu to Kangaroo. **Nature Reviews Molecular Cell Biology**, v. 4, n. 11, p. 865-877, nov. 2003. Disponível em: https://www.researchgate.net/publication/8953995_The_Outs_and_Ins_of_Transposition_from_MU_to_Kangaroo. Acesso em: 27 jul. 2019.

DOBIN, A. *et al*. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, v. 29, n. 1, p.15-21, 2013. Disponível em: https://academic.oup.com/bioinformatics/article/29/1/15/272537. Acesso em: 07 jun. 2019.

ELLIOTT, T. A. *et al*. In and out of the rRNA genes: characterization of Pokey elements in the sequence *Daphnia* genome. **Mobile DNA**, v. 4, n. 20, p. 1-13, 2013. Disponível em: https://mobilednajournal.biomedcentral.com/track/pdf/10.1186/1759-8753-4-20.pdf. Acesso em: 18 jun. 2019.

FABBRO, C. D. *et al*. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. **PLoS One**, v. 8, n. 12, p.1-13, dez. 2013. Disponível em: 10.1371/journal.pone.0085024. Acesso em: 21 jun. 2019

FAO. **Suite of Food Security Indicators.** Roma: FAO, 2019. Disponível em: https://www.fao.org/faostat/en/#data/FS. Acesso em: 10 jul 2019.

FATOKUN, C. A.; BOUKAR, O.; MURANAKA, S. Evaluation of cowpea (*Vigna unguiculata* (L.) Walp.) germplasm lines for tolerance to drought. **Plant Genetic Resources**, v. 10, n. 3, p.171–176, dez. 2012. Disponível em: https://doi.org/10.1017/S1479262112000214. Acesso em: 31 jul. 2019.

FERNÁNDEZ, J. M.; HOFFMANN, R.; VALENCIA, A. (2007) iHOP web services. **Nucleic Acids Research**, v. 35, p. 21-26, 2007. Disponível em:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933131/pdf/gkm298.pdf. Acesso em: 23 jul. 2019.

FINATTO, T. *et al*. Abiotic stress and genome dynamics: specific genes and transposable elements response to iron excess in rice. **Rice**, v. 8, n. 13, p. 1-18, 2015. Disponível em: https://thericejournal.springeropen.com/articles/10.1186/s12284-015-0045-6. Acesso em: 04 ago. 2019.

FINOTELLO, F.; DI CAMILLO, B. Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. **Briefings in Functional Genomics**, v. 14, n. 2, p.130-142, 2015. Disponível em: https://doi.org/10.1093/bfgp/elu035. Acesso em:16 jun. 2019.

FOX, H. *et al*.Transcriptome analysis of *Pinus halepensis* under drought stress and during recovery. **Tree Physiology**, v. 38, n. 3, p. 423-441, mar. 2018. Disponível em: https://doi.org/10.1093/treephys/tpx137. Acesso em: 23 jul. 2019.

FREIRE FILHO, F. R. Origem, evolução e domesticação do caupi. *In*: ARAÚJO, J. P. P.; WATT, E. E. (org). **O Caupi No Brasil**. Santo Antônio de Goiás: Embrapa Arroz e Feijão, 1988. p 26-46.

FREIRE FILHO, F. R. *et al*. **Feijão-caupi no Brasil:** produção, melhoramento genético, avanços e desafios. 1 ed. Teresina: Embrapa Meio-Norte, 2011. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/84470/1/feijao-caupi.pdf. Acesso em: 09 jul. 2019.

FREITAS, R. M. O. *et al*. Physiology responses of cowpea under water stress and rewatering in no-tillage and conventional tillage systems. **Revista Caatinga**, Mossoró, v. 30, n. 3, p. 559-567, 2016. Disponível em: http://dx.doi.org/10.1590/1983-21252017v30n303rc. Acesso em: 14 ago. 2019.

FRITCHE NETO, R.; BORÉM, A. **Plant breeding for biotic stress resistance.** 1 ed. New York: Springer Science & Business Media, 2012.

FUJITA, A. K. *et al*. Integrating Pathways of Parkinson's Disease in a Molecular Interaction Map. **Mol Neurobiol**, v. 48, n. 1, p. 88-102, 2014. Disponível em: https://link.springer.com/content/pdf/10.1007/s12035-013-8489-4.pdf. Acesso em: 07 jun. 2019.

GEMAYEL, R. *et al*. Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. **Genes**, v. 3, n. 3, p. 461-480, 2012. Disponível em: https://doi.org/10.3390/genes3030461. Acesso em: 05 jun. 2019.

GHOSH, S. *et al*. Software for systems biology: from tools to integrated platforms. **Nature Reviews Genetics**, v. 12, n. 12, p. 821-832, 2011. Disponível em: 05 jun. 2019.

GHOSH, S.; CHAN, C-K. K. Analysis of RNA-Seq Data Using TopHat and Cufflinks. **Methods in Molecular Biology**, v. 1374, p. 339-361, 2016. Disponível em: https://link.springer.com/protocol/10.1007%2F978-1-4939-3167-5_18. Acesso em: 18 jul. 2019.

GÖBEL, U. *et al*. Robustness of Transposable Element Regulation but No Genomic Shock Observed in Interspecific *Arabidopsis* Hybrids. **Genome Biology and Evolution**, v. 10, n. 6, p. 1403-1415, 2018. Disponível em: https://doi.org/10.1093/gbe/evy095. Acesso em: 17 ago. 2019.

GRABHERR, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature Biotechnology**, v. 29, n. 7, p. 644-652, 2011. Disponível em: https://www.nature.com/articles/nbt.1883. Acesso em: 15 jun. 2019.

GRANDBASTIEN, M-A. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. **Biochimica et Biophysica Acta**, v. 1849, n. 4, p. 403-416, abr. 2015. Disponível em: https://doi.org/10.1016/j.bbagrm.2014.07.017. Acesso em: 24 jul. 2019.

GRECO, I. *et al*. Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. **Journal of Translational Medicine**, v. 10, n. 217, p. 1-10, 2012. Disponível em: https://doi.org/10.1186/1479-5876-10-217. Acesso em: 23 jul. 2019.

GUÉRILLOT, R. *et al*. The diversity of prokaryotic DDE transposase of Mutator superfamily, insertion specificity and association with conjugation machineries. **Genome Biology Evolution**, v. 6, n. 2, p. 260-272, fev. 2014. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3942029/pdf/evu010.pdf. Acesso em: 04 ago. 2019.

GUIMARÃES, C. M.; STONE, L. F.; BRUNINI, O. Adaptação do feijoeiro (*Phaseolus vulgaris* L.) à seca. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Campina Grande, v. 10, n. 1, p. 70-75, 2006. Disponível em: https://doi.org/10.1590/S1415-43662006000100011. Acesso em: 18 jun. 2019.

HAAS, B. J. *et al*. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. **Nature Protocols**, v. 8, n. 8, p. 1494-1512, 2013. Disponível em: https://www.nature.com/articles/nprot.2013.084. Acesso em: 12 ago. 2019.

HANADA, K. *et al*. The functional role of *pack-MULEs* in rice inferred from purifying selection and expression profile. **The Plant Cell**, v. 21, n. 1, p. 25-38, jan. 2009. Disponível em: https://doi.org/10.1105/tpc.108.063206. Acesso em: 05 ag. 2019.

HÉNAFF, E. *et al*. Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of Brassica species. **The Plant Journal**, v. 77, n. 6, p. 852-862, 2014. Disponível em: https://doi.org/10.1111/tpj.12434. Acesso em: 20 jul. 2019.

HRDLICKOVA, R.; TOLOUE, M.; TIAN, B. RNA-Seq methods for transcriptome analysis. **Wiley Interdiscip Rev RNA**, v. 8, n. 1, p. 1757-7012, jan. 2017. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5717752/pdf/nihms923593.pdf. Acesso em: 27 jun. 2019.

HÜMBNER, S.; KOROL, A. B.; SCHMID, K. J. RNA-Seq analysis identifies genes associated with differential reproductive success under drought-stress in accessions of wild barley *Hordeum spontaneum.* **BMC Plant Biology**, v. 15, n. 134, p. 1-14, 2015. Disponível em: https://doi.org/10.1186/s12870-015-0528-z. Acesso em: 06 jun. 2019.

IVASHUTA, S. *et al.* Genotype-dependent transcriptional activation of novel repetitive elements during cold acclimation of alfalfa (*Medicago sativa*). **The Plant Journal**, v. 31, n. 5, p. 615-627, 2002. Disponível em: https://doi.org/10.1046/j.1365-313X.2002.01383.x. Acesso em: 31 jul. 2019.

JAIN, M. Next-generation sequencing technologies for gene expression profiling in plants. **Briefings in Functional Genomics**, v. 11, n. 1, p. 63-70, jan. 2012. Disponível em: https://doi.org/10.1093/bfgp/elr038. Acesso em: 17 ago. 2019.

JENSEN, L. J. *et al*. STRING 8--a global view on proteins and their functional interactions in 630 organisms. **Nucleic Acids Research**, v.37, n. 1, p. 412-416, 2009. Disponível em: https://doi.org/10.1093/nar/gkn760. Acesso em: 26 jul. 2019.

JESUS-PIRES, C. et al. Plant Thaumatin-like Proteins: Function, Evolution and Biotechnological Applications. **Current Protein and Peptide Science**, v. 21, n. 1, p. 36-51, mar. 2019. Disponível em: https://www.researchgate.net/publication/33186763. Acesso em: 21 jul. 2019.

KAKUMANU, A. *et al*. Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-Seq. **Plant Physiology**, v. 160, n. 2, p. 846-867, out. 2012. Disponível em: https://doi.org/10.1104/pp.112.200444. Acesso em: 04 ago. 2019.

KIDO, E. A. *et al*. Identification of plant protein kinases in response to abiotic and biotic stresses using SuperSAGE. **Current Protein & Peptide Science**, v. 12, n. 7, p. 643-656, 2011. Disponível em: http://www.eurekaselect.com/article/33834. Acesso em: 25 jun. 2019.

KITANO, H. Computational systems biology. **Nature**, v. 420, n. 6912, p. 206-210, 2002a. Disponível em: https://www.nature.com/articles/nature01254. Acesso em: 03 jul. 2019.

KITANO, H. Systems biology: a brief overview. **Science**, v. 295, n. 5560, p. 1662-1664, 1 mar. 2002b. Disponível em: https://www.researchgate.net/publication/11489394. Acesso em: 21 jul. 2019.

KOLLIST, H. et al. Rapid responses to abiotic stress: priming the landscape for the signal transduction network. **Trends in Plant Science**, v. 24, n. 1, p. 25-37, jan. 2019. Disponível em: https://doi.org/10.1016/j.tplants.2018.10.003. Acesso em: 17 ago. 2019.

KRALLINGER, M.; VALENCIA, A. Text-mining and information-retrieval services for molecular biology. **Genome Biology**, v. 6, n. 7, p. 224-232, 2005. Disponível em: https://doi.org/10.1186/gb-2005-6-7-224. Acesso em: 18 jul. 2019.

KUNDU, A. *et al*. High throughput sequencing reveals modulation of microRNAs in *Vigna mungo* upon Mungbean Yellow Mosaic India Virus inoculation highlighting stress regulation. **Plant Science**, v. 257, p. 96-105, abr. 2017. Disponível em: https://pubmed.ncbi.nlm.nih.gov/28224923/. Acesso em: 20 jun. 2019.

LANGMEAD, B. *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biology**, v. 10, n. 3, p. 25-34, 2009. Disponível em: https://doi.org/10.1186/gb-2009-10-3-r25. Acesso em: 04 ago. 2019.

LE, T. N. *et al*. DNA demethylases target promoter transposable elements to positively regulate stress responsive genes in Arabidopsis. **Genome Biology**, v. 15, n. 9, p. 1-18, 2014. Disponível em: https://doi.org/10.1186/s13059-014-0458-3. Acesso em: 15 ago. 2019.

LI, M. *et al.* PPICurator: A Tool for Extracting Comprehensive Protein–Protein Interaction Information. Proteomics, v. 19, n. 4, p. 18-29, fev. 2019. Disponível em: https://doi.org/10.1002/pmic.201800291. Acesso em: 01 ago. 2019.

LI, Y.; LI, C.; JIN, Y. Domestication of transposable elements into MicroRNA genes in plants. **PLoS One**, v. 6, n. 5, p. 1-13, maio 2011. Disponível em: https://doi.org/10.1371/journal.pone.0019212. Acesso em: 17 ago. 2019.

LISCH, D.; BENNETZEN, J. L. Transposable element origins of epigenetic gene regulation. **Current Opinion in Plant Biology**, v. 14, n. 2, p. 156-161, 2011. Disponível em: https://pubmed.ncbi.nlm.nih.gov/21444239/. Acesso em: 08 jun. 2019.

LIU, S. *et al*. Mu transposon insertion sites and meiotic recombination events colocalize with epigenetic marks for open chromatin across the maize genome. **PLoS Genetics**, v. 5, n. 11, p. 1-13, nov. 2009. Disponível em: https://doi.org/10.1371/journal.pgen.1000733. Acesso em: 03 jul. 2019

LO, S. *et al*. Identification of QTL controlling domestication-related traits in cowpea (*Vigna unguiculata* L. Walp). **Scientific Reports**, v. 8, n. 6261, p. 1-9, 2018. Disponível em: https://doi.org/10.1038/s41598-018-24349-4. Acesso em: 16 jul. 2019.

LONARDI, S. *et al*. The genome of cowpea (*Vigna unguiculata* [L.] Walp.). **The Plant Journal**, v. 98, p. 767-782, 2019. Disponível em: https://doi.org/10.1111/tpj.14349. Acesso em: 25 jul. 2019.

LOPES, F. R. *et al*. Transcriptional activity, chromosomal distribution and expression effects of transposable elements in *Coffea* genomes. **PLoS One**, v. 8, n. 11, p. 1-16, nov. 2013. Disponível em: https://doi.org/10.1371/journal.pone.0078931. Acesso em: 06 jun. 2019.

MAKAREVITCH, I. *et al*. Transposable elements contribute to activation of maize genes in response to abiotic stress. **PLoS Genetics,** v. 11, n. 1, p. 1-12, jan. 2015. Disponível em: https://doi.org/10.1371/journal.pgen.1004915. Acesso em: 20 jun. 2019.

MARTIN, J. A.; WANG, Z. Next-generation transcriptome assembly. **Nature Reviews Genetics**, v. 12, n. 10, p. 671-682, 2011. Disponível em: https://www.nature.com/articles/nrg3068. Acesso em: 25 jul. 2019.

MCCLINTOCK, B. The origin and behavior of mutable loci in maize. **PNAS USA**, v. 36, n. 6, p. 344–355, 1950. Disponível em: https://doi.org/10.1073/pnas.36.6.344. Acesso em: 17 ago. 2019.

MCCLINTOCK, B. Induction of Instability at Selected Loci in Maize. **Genetics**, v. 38, n. 6, p. 579-599, 1953. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1209627/pdf/579.pdf. Acesso em: 01 ago. 2019.

MCCLINTOCK, B. The significance of responses of the genome to challenge (Nobel lecture). **Science**, v. 226, n. 4676, p. 792-801, 1984. Disponível em: https://www.science.org/doi/10.1126/science.15739260. Acesso em: 01 ago. 2019.

MCCUE, A. D.; SLOTKIN, R. K. Transposable element small RNAs as regulators of gene expression. **Trends Genetics**, v. 28, n. 12, p. 616-623, 2012. Disponível em: https://doi.org/10.1016/j.tig.2012.09.001. Acesso em: 10 jun. 2019.

MENDES, R. M. S. *et al*. Relações fonte-dreno em feijão-de-corda submetido à deficiência hídrica. **Ciência Agronômica**, Fortaleza, v. 38, n. 1, p. 95-103, 2007. Disponível em: http://ccarevista.ufc.br/seer/index.php/ccarevista/article/view/158/152. Acesso em: 14 jul. 2019.

MIKA, S.; ROST, B. NLProt: extracting protein names and sequences from papers. **Nucleic Acids Research**, v. 32, p. 634-637, 2004. Disponível em: https://doi.org/10.1093/nar/gkh427. Acesso em: 30 jun. 2019.

MUEHLBAUER, G. J. *et al*. A hAT superfamily transposase recruited by the cereal grass genome. **Molecular Genetics and Genomics**, v. 275, n. 6, p. 553-563, 2006. Disponível em: https://www.researchgate.net/publication/7306460_A_hAT_superfamily_transposase_recruited_by_the_cereal_grass_genome. Acesso em: 19 jun. 2019.

NASCIMENTO, S. P. *et al*. Tolerância ao déficit hídrico em genótipos de feijão-caupi. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Campina Grande, v. 15, n. 8, p. 853-860, 2011. Disponível em: https://www.scielo.br/j/rbeaa/a/XmDHyjThcWxXV4MqjNVX6Dh/?format=pdf&lang=pt. Acesso em: 04 ago. 2019.

NEJAT, N.; MANTRI, N. Plant Immune System: Crosstalk Between Responses to Biotic and Abiotic Stresses the Missing Link in Understanding Plant Defence. **Current Issues in Molecular Biology**, v. 23, p. 1-16, 2017. Disponível em: https://doi.org/10.21775/cimb.023.001. Acesso em: 27 jul. 2019.

NEUMANN, P.; POZÁRKOVÁ, D.; MACAS, J. Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. **Plant Molecular Biology**, v. 53, n. 3, p. 399-410, 2003. Disponível em: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.551.7542&rep=rep1&type=pdf. Acesso em: 12 jul. 2019.

NG, S.-Y. *et al*. Long noncoding RNAs in development and disease of the central nervous system. **Trends Genetics**, v. 29, n. 8, p. 461-468, 2013. Disponível em: https://doi.org/10.1016/j.tig.2013.03.002. Acesso em: 27 jul. 2019.

OMOIGUI, L. O. *et al*. Identification of new sources of resistance to *Striga gesnerioides* in cowpea *Vigna unguiculata* accessions. **Genetic Resources and Crop Evolution**, v. 64, n. 5, p. 901-911, 2016. Disponível em: https://www.researchgate.net/publication/301903210. Acesso em: 05 ago. 2019.

PASSOS, A. R. *et al*. Divergência genética em feijão-caupi. **Bragantia**, Campinas, v. 66, n. 4, p. 579-586, 2007. Disponível em: https://doi.org/10.1590/S0006-87052007000400007. Acesso em: 14 jul. 2019.

PAVLOPOULOS, G. A. *et al*. Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. **Gigascience**, v. 4, n. 1, p. 38-66, dez. 2015. Disponível em: https://doi.org/10.1186/s13742-015-0077-2. Acesso em: 30 jun. 2019.

PETRYSZAK, R. *et al*. Expression Atlas update - an integrated database of gene and protein expression in humans, animals and plants. **Nucleic Acids Research**, v. 44, n. 1, p.746-752, 2016. Disponível em: https://doi.org/10.1093/nar/gkv1045. Acesso em: 18 jul. 2019.

PIÉGU, B. *et al*. A survey of transposable elements classification system – A call for a fundamental update to meet the challenge of their diversity and complexity. **Molecular Phylogenetics and Evolution**, v. 86, p. 90-109, 2015. Disponível em: https://serval.unil.ch/resource/serval:BIB_71563EC2D515.P001/REF.pdf. Acesso em: 21 jun. 2019.

PIRIYAPONGSA, J.; JORDAN, I. K. Dual coding of siRNAs and miRNAs by plant transposable elements. **RNA**, v. 14, n. 5, p. 814-821, 2008. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2327354/. Acesso em: 02 ago. 2019.

PLATT, R.; VANDEWEGE, M.; RAY, D. Mammalian transposable elements and their impacts on genome evolution. **Chromosome Research**, v. 26, n. 1-2, p. 25-43, 2018. Disponível em: https://doi.org/10.1007/s10577-017-9570-z. Acesso em: 10 jun. 2019.

PRICE, W. N.; ARKIN, A. P. PaperBLAST: Text Mining Papers for Information about Homologs. **mySistems**, v. 2, n. 4, p. 1-10, jul./ago. 2017. Disponível em: https://doi.org/10.1128/mSystems.00039-17. Acesso em: 10 jul. 2019.

RAIZADA, M. N.; BREWER, K. V.; WALBOT, V. A maize MuDR transposon promoter shows limited autoregulation. **Molecular Genetics Genomics**, v. 265, n. 1, p. 82-94, mar. 2001. Disponível em: https://doi.org/10.1007/s004380000393. Acesso em: 16 jun. 2019.

REBHOLZ-SCHUHMANN, D.; OELLRICH, A.; HOEHNDORF, R. Text-mining solutions for biomedical research: enabling integrative biology. **Nature Reviews Genetics**, v. 13, n. 12, p. 829-839, 2012. Disponível em: https://www.nature.com/articles/nrg3337. Acesso em: 26 jul. 2019.

SAYERS, E. W. *et al.* Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 47, n. 1, p. 23-28, 2019. Disponível em: https://doi.org/10.1093/nar/gky1069. Acesso em: 17 ago. 2019.

SCHARTE, J.; SCHÖN, H.; WEIS, E. Photosynthesis and carbohydrate metabolism in tobacco leaves during an incompatible interaction with *Phytophthora nicotianae*. **Plant, Cell and Environment**, v. 28, p. 1421-1435, 2005. Disponível em: https://doi.org/10.1111/j.1365-3040.2005.01380.x. Acesso em: 03 jul. 2019.

SCHURCH, N. J. *et al*. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? **RNA**, v. 22, n. 6, p. 839-851, 2016. Disponível em: http://www.rnajournal.org/cgi/doi/10.1261/rna.053959.115. Acesso em: 10 jun. 2019.

SINGH, B. B. (2006) Cowpea breeding at IITA: Highlights of advances impacts. *In:* CONGRESSO NACIONAL DE FEIJÃO-CAUPI, 1., 2006, Teresina**. Anais: Reunião nacional de feijão-caupi - Tecnologias para o agronegócio.** Teresina: Embrapa Meio-Norte, 2006, p. 121.

SLOTKIN, R. K.; MARTIENSSEN, R. Transposable elements and the epigenetic regulation of the genome. **Nature Reviews Genetics**, v. 8, n. 4, p. 272-85, abr. 2007. Disponível em: http://www.somosbacteriasyvirus.com/genome2.pdf. Acesso em: 12 ago. 2019.

SNEL, B. *et al*. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. **Nucleic Acids Research**, v. 28, n. 18, p. 3442-3444, 2000. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC110752/. Acesso em: 20 ago. 2019.

STELPFLUG, S. C. *et al.* An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development. **The Plant Genome**, v. 9, n. 1, p. 1-16, mar. 2016. Disponível em: https://doi.org/10.3835/plantgenome2015.04.0025. Acesso em: 23 jul. 2019.

SUN, D.; WANG, M.; LI, A. MPTM: A tool for mining protein post-translational modifications from literature. **Journal of Bioinformatics and Computacional Biology**, v. 15, n. 5., p. 1740-1745, 2017. Disponível em: https://doi.org/10.1142/S0219720017400054. Acesso em: 25 jul. 2019.

SUONIEMI, A.; NARVANTO, A.; SCHULMAN, A. H. The BARE-1 retrotransposon is transcribed in barley from an *LTR* promoter active in transient assays. **Plant Molecular Biology**, v. 31, n. 2, p. 295-306, 1996. Disponível em: https://www.researchgate.net/publication/14445265. Acesso em: 30 jun. 2019.

SZKLARCZYK, D. *et al*. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. **Nucleic Acids Research**, v. 45, p. 362-368, 2017. Disponível em: https://doi.org/10.1093/nar/gkw937. Acesso em: 17 ago. 2019.

THOTTAPPILLY, G.; ROSSEL, H. W. Virus diseases of cowpea in tropical Africa. **Tropical Pest Management**, v. 38, n. 4, p. 337-348, 1992. Disponível em: https://doi.org/10.1080/09670879209371724. Acesso em: 21 jun. 2019.

TOVKACH, A. *et al.* Transposon-mediated alteration of TaMATE1B expression in wheat confers constitutive citrate efflux from root apices. **Plant Physiology**, v. 161, n. 2, p. 880-892, fev. 2013. Disponível em: https://doi.org/10.1104/pp.112.207142. Acesso em: 22 jul. 2019.

TRAPNELL, C.; PACHTER, L.; SALZBERG, S. L. TopHat: discovering splice junctions with RNA-Seq. **Bioinformatics**, v. 25, n. 9, p. 1105-1111, 2009. Disponível em: https://doi.org/10.1093/bioinformatics/btp120. Acesso em: 10 jun. 2019.

TRAPNELL, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**, v. 7, n. 3, p. 562-578, 2012. Disponível em: https://authors.library.caltech.edu/74752/3/nihms-366741.pdf. Acesso em: 27 jun. 2019.

VANHOLME, R. *et al.* A systems biology view of responses to lignin biosynthesis perturbations in Arabidopsis. **The Plant Cell**, v. 24, n. 9, p. 3506-3529, set. 2012. Disponível em: https://doi.org/10.1105/tpc.112.102574. Acesso em: 12 ago. 2019

VERDIER, J. *et al*. A regulatory network-based approach dissects late maturation processes related to the acquisition of desiccation tolerance and longevity of *Medicago truncatula* seeds. **Plant Physiology**, v. 163, n. 2, p. 757-774, out. 2013. Disponível em: https://doi.org/10.1104/pp.113.222380. Acesso em: 15 jul. 2019.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57-63, jan. 2009. Disponível em: https://www.nature.com/articles/nrg2484. Acesso em: 12 ago. 2019.

WICKER, T. *et al*. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973-982, dez. 2007. Disponível em: https://www.researchgate.net/publication/5861686. Acesso em: 15 jul. 2019.

WU, Q. *et al*. Root-Specific Expression of a Jacalin Lectin Family Protein Gene Requires a Transposable Element Sequence in the Promoter. **Genes**, v. 9, n. 11, p. 550-564, nov. 2018. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6266147/. Acesso em: 26 jul. 2019.

YAO, S. *et al*. The *Vigna unguiculata* Gene Expression Atlas (VuGEA) from de novo assembly and quantification of RNA-seq data provides insights into seed maturation mechanisms. **The Plant Journal**, v. 88, n. 2, p. 318-327, 2016. Disponível em: https://doi.org/10.1111/tpj.13279. Acesso em: 08 jun. 2019.

YE, C.; JI, G.; LIANG, C. detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes. **Scientific Reports**, v. 6, n. 19688, p. 1-11, 2016. Disponível em: https://doi.org/10.1038/srep19688. Acesso em: 04 ago. 2019.

ZEH, D. W.; ZEH, J. A.; ISHIDA, Y. Transposable elements and an epigenetic basis for punctuated equilibria. **BioEssays**, v. 31, n. 7, p. 715-726, 2009. Disponível em: https://doi.org/10.1002/bies.200900026. Acesso em: 15 ago. 2019.

ZENONI, S. *et al*. Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. **Plant Physiology**, v. 152, n. 4, p. 1787-1795, abr. 2010. Disponível em: https://doi.org/10.1104/pp.109.149716. Acesso em: 20 jul. 2019.

ZHAO, D. *et al*. The unique epigenetic features of Pack-MULEs and their impact on chromosomal base composition and expression spectrum. **Nucleic Acids Research**, v. 46, n. 5, p. 2380-2397, 2018. Disponível em: https://doi.org/10.1093/nar/gky091. Acesso em: 15 ago. 2019.

ZHAO, Y. *et al.* Identification of a retroelement from the resurrection plant Boea hygrometrica that confers osmotic and alkaline tolerance in *Arabidopsis thaliana*. **PLoS One**, v. 9, n. 5, p. 1-12, maio 2014. Disponível em: https://doi.org/10.1371/journal.pone.0098098. Acesso em: 19 jul. 2019.

ZHU, F. *et al*. Biomedical text mining and its applications in cancer research. **Journal of Biomedical Informatics**, v. 46, n. 2, p. 200-211, 2013. Disponível em: https://www.researchgate.net/publication/233536867_Biomedical_text_mining_and_its_applications_in_cancer_research. Acesso em: 01 ago. 2019.

ANEXO A - MATERIAL SUPLEMENTAR: ARTIGO 1

Supplementary material 1: PMID list. PMID list of all abstracts analyzed with LAITOR4HPC for all case study.

P.S.: Acessar o anexo enviado por e-mail.

Supplementary material 2: Searching bioentities co-occurrences in all available abstracts for one specie (First case study report). The information here embrace the total proteins, co-occurrences and terms mapped by LAITOR4HPC in Glicine max. In order to accomplish that, the whole MEDLINE corpus was analyzed under the taxid filter (Glycine max – Taxonomy ID: 3847) .The analysed corpus is representative of MEDLINE database until June 2017, with 1134 XML files containing ~30,000 abstracts each. (A) The ammount of co-occurrences for each type of interaction is describe, with INT_1 being the more and INT_4 the less likelly to be an effective interaction. (B) For each category (protein, co-occurrence, term) there are : (1) Total of unique (non-redundant) elements; (2) the total of mapped elements; (3) The absolute number that each elemet was mapped (repetitions).

P.S.: Acessar o anexo enviado por e-mail.

Supplementary material 3: Key words list used for the second case study.

	Keywords
1	pr AND plant
2	ltp AND plant
3	amp AND plant
4	virus AND plant
5	nbs-lrr AND plant
6	hevein AND plant
7	fungus AND plant
8	snakin AND plant
9	thionin AND plant
10	knottin AND plant
11	defense AND plant
12	defensin AND plant
13	cyclotide AND plant
14	pathogen AND plant
15	resistance AND plant
16	thaumatin-like protein AND plant
17	protein interaction AND plant disease

18 protein interaction AND plant infection
19 protein interaction AND plant biotic stress
20 protein interaction AND plant pathogen interaction

Supplementary material 4: Using keywords to look for all described interactions on one subject, all plants summary. (Second case study report). The information here embrace the total proteins, co-occurrences and terms mapped by LAITOR4HPC 12 plant species under the biotic stress related abstracts set . In order to accomplish that, many keywords were used to search for the subject and the same data set filteres for differente specieas by LAITOR4HPC taxid filter opition. The analysed corpus is representative of MEDLINE database until June 2017. (A) The ammount of co-occurrences for each type of interaction is describe considering all plants together. INT_1 is the more and INT_4 the less likelly to be an effective interaction. (B) For each category (protein, co-occurrence, term) there are : (1) Total of unique (non-redundant) elements; (2) the total of mapped elements; (3) The absolute number that each elemet was mapped (repetitions).

P.S.: Acessar o anexo enviado por e-mail.

Supplementary material 5: Using keywords to look for all described interactions on one subject. (Second case study report). The information here embrace the total proteins, co-occurrences and terms mapped by LAITOR4HPC in Arabdopsis thaiana. In order to accomplish that, the biotic stress related abstract set was analyzed under the taxid filter (Arabidopsis thaliana – Taxonomy ID: 3702). The analysed corpus is representative of MEDLINE database until June 2017. (A) The ammount of co-occurrences for each type of interaction is describe, with INT_1 being the more and INT_4 the less likelly to be an effective interaction. (B) For each category (protein, co-occurrence, term) there are : (1) Total of unique (non-redundant) elements; (2) the total of mapped elements; (3) The absolute number that each elemet was mapped (repetitions).

P.S.: Acessar o anexo enviado por e-mail.

Supplementary material 6: Building pathways. (Third case study report). A list with the type 1 interactions of 31 PMIDs mapped for Arabidopisis thaliana based on "plant AND defensin" keyword and used for pathway construction on CellDesigner is presented. It is possible to observe both bioentities (species 1 and species 2), the interaction term (correlation) followed by the PMID related to the mapped information and the information about cured data (Automatic and Manual annotation). Automatic annotation is related to all that was mapped by LAITOR4HPC correctly. Manual annotation is either confirming the automatic annotation or adding non-mapped co-ocurrences. In some cases the mapping was correct, but not related to the subject of interest.

Interaction type	Species 1	Correlation	Species 2	PMID	Automatic annotation	Manual annotation
INT_1	PDF1	induced|	pathogenesis-related protein 1	28605157	not mapped	not mapped
INT_1	EIN3	suppression|activation|	PDF1.2	28168848	Yes	Yes
INT_1	EIL1	suppression|activation|	PDF1.2	28168848	Yes	Yes
INT_1	ORA59	activator|	PDF1.2	28168848	Yes	Yes
INT_1	EIN3	mediated|	ORA59	28168848	Yes	Yes
INT_1	ANAC032	represses|	MYC2	27632992	Yes	Yes
INT_1	ANAC032	activates|repressing|	NIMIN1	27632992	Yes	Yes
INT_1	ANAC032	enhances|binding|	MYC2	27632992	Yes	Yes
INT_1	ANAC032	enhances|binding|	NIMIN1	27632992	Yes	Yes
INT_1	MYC2	down-regulating|	PDF1	26962208	Yes	Yes
INT_1	JAZ1	down-regulating|	MYC2	26962208	Yes	Yes

INT_1	JAZ1	suppression|	MYC2	26962208	Yes	Yes
INT_1	JAZ1	induce|	MYC2	26962208	Yes	Yes
INT_1	ERF9	down-regulated|promotion|	PDF1.2	24157210	Yes	No
INT_1	ERF9	binding|promoters|	PDF1.2	24157210	Yes	No
INT_1	ERF9	suppression|	PDF1.2	24157210	Yes	No
INT_1	ERF1	up-regulated|	PDF1.2	22371506	Yes	no
INT_1	PDF1.2	up-regulated|	coi1	22371506	no	Yes
INT_1	PR-1	induction|	transcription factor	21414016	not related	not related
INT_1	ORA59	activation|	PDF1.2	21246258	Yes	Yes
INT_1	ORA59	acts|regulator|	PDF1.2	21246258	Yes	Yes
INT_1	ORA59	promoter|interacting|	PDF1.2	21246258	Yes	Yes
INT_1	ORA59	bound|trans-activated|	PDF1.2	21246258	Yes	Yes
INT_1	ORA59	bound|	PDF1.2	21246258	Yes	Yes
INT_1	ORA59	form|	PDF1.2	21246258	Yes	Yes
INT_1	WRKY50	suppressor|	ssi2	21030507	no	Yes
INT_1	WRKY51	suppressor|	ssi2	21030507	no	Yes
INT_1	EIN2	induced|	atmyb44	20826953	Yes	Yes
INT_1	CPK3	phosphorylating|	HsfB2a	20798597	Yes	Yes
INT_1	HsfB1	regulation|	Pdf1.2	19529832	Yes	no
INT_1	HsfB2b	regulation|	Pdf1.2	19529832	Yes	no
INT_1	PDF1.2	down-regulated|	PR-1	19121119	not mapped	not mapped
INT_1	NAI1	activated|	PYK10	18248598	not related	not related
INT_1	FLC	up-regulated|	HDA6	18212027	no	Yes

INT_1	PDF1.2	regulates|	transcription factor	17616737	no	Yes
INT_1	PDF1.2	activates|	PROPEP1	17566109	Yes	no
INT_1	PEPR1	activates|mediated|	PROPEP1	17566109	Yes	no
INT_1	PDF1.2	mediated|	PEPR1	17566109	Yes	no
INT_1	PROPEP 1	enhances|	PROPEP3	17566109	Yes	no
INT_1	PROPEP 2	enhances|	PROPEP3	17566109	Yes	no
INT_1	PDF1.2	interacts|suppresses|	glutaredoxin	17397508	Yes	Yes
INT_1	ARF2	inhibition|	PDF1	16822015	Yes	Yes
INT_1	AtPep1	activates|	PDF1.2	16785434	Yes	no
INT_1	cpr5	activate|	npr1	16648642	not mapped	not mapped
INT_1	EDS5	activity|	ssi2	15828688	no	Yes
INT_1	PDF1.2	activate|	ssi2	15828688	no	Yes
INT_1	PDF1.2	activate|	PR1	15828688	no	Yes
INT_1	PDF1.2	requirement|control|induction|	RCE1	15125769	no	Yes
INT_1	PDF1.2	induction|	ssi2	14615603	Yes	no
INT_1	PDF1.2	required|	ssi2	14615603	Yes	no
INT_1	BOS1	interaction|	coi1	14555693	no	Yes
INT_1	ein2	enhance|	ssi1	12848424	no	Yes
INT_1	jar1	enhance|	ssi1	12848424	no	Yes
INT_1	PDF1.2	stimulated|	PR1	11971137	no	Yes

INT_1	GA1	enhancing|	GA4	11457902	not related	not related
INT_1	GA3	enhancing|	GA4	11457902	not related	not related
INT_1	NPR1	required|	PR-1	11439131	not related	not related
INT_1	COI1	required|	cev1	11340179	no	Yes
INT_1	ETR1	required|	cev1	11340179	no	Yes
INT_1	NPR1	suppressor|requirement|	ssi1	9927638	no	Yes
INT_1	PDF1.2	suppressed|	ssi1	9927638	no	Yes

Supplementary material 7: SbmlPathway file containing the built pathway in a CellDesign compatible format.

P.S.: Acessar o anexo enviado por e-mail.

ANEXO B - MATERIAL SUPLEMENTAR: ARTIGO 2

Material suplementar 1: Relatório RNAseq-QC. Tabelas geradas automaticamente pelo RNAseq-QC. O documento contém o total de reads sequenciadas e sobreviventes, dados de análise exploratória entre outras métricas de qualidade da montagem. Para acessar, após extrair o arquivo compactado, acesse o arquivo no formato html chamado "index" que reunirá todos os arquivos em um formato de relatório organizado e fácil de ser compreendido.

P.S.: Acessar o anexo enviado por e-mail.

Material suplementar 2: Relatório GenPipes RNA-Seq. Tabelas geradas automaticamente pelo *pipeline* GenPipes. O documento contém diversas métricas de qualidade como regiões alinhadas com éxons e íntrons, identificação de rRNA, entre outras. Para acessar, após extrair o arquivo compactado, acesse o arquivo no formato html chamado "index" que reunirá todos os arquivos em um formato de relatório organizado e fácil de ser compreendido.

P.S.: Acessar o anexo enviado por e-mail.

Material suplementar 3: Lista como os 354 transposons do genoma que foram reclassificados a partir do banco de Eudicotiledoneas do RepBase. Foram considerados apenas elementos com pelo menos 80% de identidade e de cobertura.

TEs_genome_ID	Previous classification	Eudicot-based-Classification
1021595,187764	Unknown	C:hAT-4_GM|Vu11
1021595,378429	Unknown	C:Gypsy-60_GR-I|Vu09
1021595,466803	Unknown	C:Gypsy-4_CP-I|Vu02
1021595,524351	Unknown	C:Gypsy-119_GM-I|Vu09_3
1021595,584363	Unknown	C:Gypsy-4_CP-I|Vu02
1021595,741340	Unknown	C:Gypsy-60_GR-I|Vu09
1021595,118739	Unknown	C:TGM1_GM|Vu04
1021595,322270	Unknown	C:hAT-3_GM|Vu02_6

1021595,382000	Unknown	+:EnSpm-3_GM|Vu04
1021595,432501	Unknown	C:hAT-3_GM|Vu02_6
1021595,467432	Unknown	C:L1-13_GM|Vu11
1021595,124723	Unknown	C:Copia-21_GR-I|Vu09
1021595,428496	Unknown	+:Gypsy-119_GM-I|Vu01_3
1021595,445088	Unknown	C:Gypsy-60_GR-I|Vu09
1021595,447273	Unknown	C:ENSPM2_MT|Vu10
1021595,479780	Unknown	C:Copia-21_GR-I|Vu06
1021595,482859	Unknown	C:Copia-21_GR-I|Vu10
1021595,566592	Unknown	C:ENSPM2_MT|Vu05
1021595,788210	Unknown	+:hAT-2_GM|Vu04_10
1021595,962130	Unknown	+:hAT-2_GM|Vu05_2
1021595,246011	Unknown	C:MUDRB_PT|Vu02
1021595,422924	Unknown	C:MUDRB_PT|Vu02
1021595,269411	Unknown	C:MUDRB_PT|Vu02
1021595,706700	Unknown	+:Copia-72_VV-LTR|Vu07
1021595,106072	Unknown	C:hAT-2_GM|Vu11_4
1021595,107598	Unknown	C:hAT-3_GM|Vu06_2
1021595,108398	Unknown	C:Gypsy-35_GAr-I|Vu09
1021595,108714	Unknown	C:Sat-2_STu|Vu11_1
1021595,108723	Unknown	C:L1-1_STu|Vu02_1
1021595,113060	Unknown	C:GmGYPSY10_I|Vu07_2
1021595,114215	Unknown	C:Gypsy-12_RC-I|Vu09
1021595,114400	Unknown	+:EnSpm-3_GM|Vu03
1021595,114916	Unknown	C:Sat-2_STu|Vu11_1
1021595,115300	Unknown	+:Copia-21_GR-I|Vu02
1021595,115720	Unknown	C:MuDr3_MT|Vu04_1
1021595,115723	Unknown	+:MuDr3_MT|Vu01_1
1021595,115775	Unknown	C:hAT-2_GM|Vu01_5
1021595,124332	Unknown	+:HAT-13_Mad|Vu03
1021595,124464	Unknown	+:MuDr3_MT|Vu03_1
1021595,124783	Unknown	+:hAT-6_TC|Vu10
1021595,124916	Unknown	+:Copia-63_GM-I|Vu04
1021595,125660	Unknown	C:Copia-21_GR-I|Vu06
1021595,128508	Unknown	C:L1-13_GM|Vu11
1021595,128543	Unknown	+:Gypsy-76_NS-I|Vu05
1021595,129677	Unknown	+:Copia-63_GM-I|Vu04
1021595,129874	Unknown	C:TGM5_GM|Vu10
1021595,130032	Unknown	+:MuDr4_MT|Vu09
1021595,131077	Unknown	C:hAT-2_GM|Vu09_7
1021595,131090	Unknown	+:hAT-2_GM|Vu09_8
1021595,133607	Unknown	+:L1-1_STu|Vu03_59
1021595,137855	Unknown	C:hAT-6_TC|Vu08_2
1021595,138145	Unknown	C:hAT-2_GM|Vu08_3

1021595,138301	Unknown	C:hAT-3_GM|Vu06_6
1021595,138465	Unknown	C:MuDr4_MT|Vu04
1021595,138640	Unknown	C:L1-1_STu|Vu11_11
1021595,138771	Unknown	+:Copia29-VV_I|Vu09
1021595,138952	Unknown	C:Gypsy-21_ST-I|Vu09
1021595,138981	Unknown	C:TONT2-I_PV|Vu09_10
1021595,139150	Unknown	+:MuDr4_MT|Vu08
1021595,140097	Unknown	C:TONT2-I_PV|Vu09_10
1021595,146177	Unknown	C:hAT-2_GM|Vu11_4
1021595,146424	Unknown	C:Copia-10_Cia-LTR|Vu09
1021595,149882	Unknown	C:Gypsy-10_MN-LTR|Vu07
1021595,149896	Unknown	+:hAT-6_TC|Vu07_4
1021595,150160	Unknown	+:MuDr4_MT|Vu03
1021595,150258	Unknown	+:Gypsy-4_GAr-I|Vu10
1021595,153160	Unknown	C:hAT-6_TC|Vu02
1021595,156726	Unknown	C:MtPH-M-3-Ia|Vu09
1021595,158190	Unknown	C:TGM1_GM|Vu11
1021595,158450	Unknown	C:hAT-3_GM|Vu06_6
1021595,159680	Unknown	C:hAT-2_GM|Vu05_7
1021595,159771	Unknown	+:hAT-4_GM|Vu05
1021595,162045	Unknown	+:MUDRAV2_MT|Vu07
1021595,162228	Unknown	+:GmCOPIA10_I|Vu02
1021595,162676	Unknown	+:MuDr4_MT|Vu03
1021595,164751	Unknown	+:INMU1|Vu05
1021595,168721	Unknown	C:Gypsy-35_GAr-I|Vu09
1021595,169844	Unknown	C:hAT-2_GM|Vu07_2
1021595,170708	Unknown	C:MUDRB_PT|Vu02
1021595,172511	Unknown	C:hAT-3_GM|Vu06_6
1021595,173364	Unknown	C:hAT-2_ALy|Vu01
1021595,173815	Unknown	C:hAT-2_GM|Vu06_2
1021595,177211	Unknown	+:MuDr3_MT|Vu01
1021595,177560	Unknown	C:Copia-72_VV-LTR|Vu09
1021595,177873	Unknown	+:hAT-2_GM|Vu09_2
1021595,179630	Unknown	+:Copia-21_GR-I|Vu01
1021595,181400	Unknown	C:ENSPM2_MT|Vu10
1021595,185635	Unknown	C:Gypsy-21_ST-I|Vu09
1021595,186270	Unknown	C:MUDRB_PT|Vu03
1021595,187181	Unknown	+:hAT-2_GM|Vu09
1021595,188241	Unknown	C:hAT-2_GM|Vu06_4
1021595,190456	Unknown	C:hAT-3_GM|Vu06_2
1021595,192634	Unknown	C:MtPH-M-3-Ia|Vu09
1021595,193060	Unknown	+:hAT-2_GM|Vu09_2
1021595,193935	Unknown	C:Gypsy-35_GAr-I|Vu09
1021595,199300	Unknown	C:L1-13_GM|Vu11

1021595,200859	Unknown	+:hAT-2_GM|Vu04_3
1021595,204466	Unknown	C:Gypsy-10_Mad-I|Vu03
1021595,204732	Unknown	C:hAT-2_GM|Vu07
1021595,206824	Unknown	+:Copia-62_GM-I|Vu02_6
1021595,207847	Unknown	C:Copia-72_VV-LTR|Vu03_1
1021595,211044	Unknown	+:Gypsy-21_ST-I|Vu09
1021595,211340	Unknown	+:MuDr4_MT|Vu03_1
1021595,211943	Unknown	C:Gypsy-4_GAr-I|Vu09
1021595,218050	Unknown	+:EnSpm-1_JC|Vu07
1021595,218506	Unknown	C:TGM1_GM|Vu11
1021595,220816	Unknown	+:Copia-62_GM-I|Vu02_6
1021595,221765	Unknown	C:Copia-21_GR-I|Vu04
1021595,222684	Unknown	C:EnSpm-1_GM|Vu10
1021595,229314	Unknown	+:hAT-2_GM|Vu11_2
1021595,229675	Unknown	+:EnSpm-1_JC|Vu07
1021595,229888	Unknown	+:Copia-32_ECa-LTR|Vu02_1
1021595,232967	Unknown	+:hAT-12_FV|Vu06
1021595,235198	Unknown	C:Copia-10_Cia-LTR|Vu09
1021595,236615	Unknown	C:TGM5_GM|Vu10
1021595,237084	Unknown	+:hAT-N2_FV|Vu11
1021595,238462	Unknown	+:hAT-2_GM|Vu10_3
1021595,238975	Unknown	C:L1-1_STu|Vu11_53
1021595,240310	Unknown	+:Gypsy-34_GR-I|Vu07
1021595,241601	Unknown	C:hAT-2_GM|Vu06_2
1021595,243805	Unknown	C:Gypsy-21_ST-I|Vu09
1021595,244360	Unknown	+:MuDr4_MT|Vu08
1021595,244684	Unknown	C:hAT-2_GM|Vu03_15
1021595,245141	Unknown	+:EnSpm-1_JC|Vu07
1021595,245850	Unknown	+:hAT-2_GM|Vu11_3
1021595,245900	Unknown	+:hAT-2_GM|Vu03_6
1021595,247770	Unknown	C:hAT-6_TC|Vu11
1021595,248087	Unknown	C:hAT-2_GM|Vu11_4
1021595,250859	Unknown	C:L1-13_GM|Vu11
1021595,254737	Unknown	+:EnSpm-3_GM|Vu03
1021595,256230	Unknown	C:Gypsy-12_RC-I|Vu09
1021595,256326	Unknown	C:HAT1_MT|Vu11
1021595,257404	Unknown	+:Copia-33_GAr-LTR|Vu05
1021595,259462	Unknown	C:MUDRB_PT|Vu03
1021595,260762	Unknown	C:RAM9B_I|Vu10
1021595,260890	Unknown	C:Copia-72_VV-LTR|Vu10
1021595,260959	Unknown	C:MuDr3_MT|Vu04_1
1021595,261938	Unknown	+:Gypsy-8_GAr-I|Vu05
1021595,265427	Unknown	C:Gypsy-118_GM-I|Vu04
1021595,265647	Unknown	+:MUDRAV2_MT|Vu07

1021595,267640	Unknown	+:L1-1_STu|Vu05_224
1021595,271046	Unknown	+:Copia-21_GR-I|Vu11_1
1021595,271715	Unknown	+:Copia29-VV_I|Vu09
1021595,281003	Unknown	+:EnSpm-1_JC|Vu07
1021595,285393	Unknown	C:Gypsy-21_ST-I|Vu04
1021595,289839	Unknown	C:TGM1_GM|Vu11
1021595,292220	Unknown	+:Copia-63_GM-I|Vu08
1021595,292694	Unknown	C:L1-13_GM|Vu11
1021595,299460	Unknown	C:Gypsy-35_GAr-I|Vu01
1021595,300347	Unknown	+:MUDRAV2_MT|Vu10
1021595,301519	Unknown	+:Copia-63_GM-I|Vu08
1021595,303837	Unknown	C:Gypsy-10_MN-LTR|Vu07
1021595,304442	Unknown	C:Copia-21_GR-I|Vu11
1021595,304945	Unknown	C:hAT-2_GM|Vu08_2
1021595,307207	Unknown	C:Gypsy-118_GM-I|Vu04
1021595,309074	Unknown	C:Gypsy-119_GM-I|Vu09_3
1021595,309140	Unknown	C:Gypsy-4_CP-I|Vu02
1021595,311231	Unknown	+:Gypsy-21_ST-I|Vu03
1021595,311237	Unknown	C:MuDr3_MT|Vu04_1
1021595,313973	Unknown	C:hAT-3_GM|Vu07_2
1021595,314631	Unknown	+:Copia-63_GM-I|Vu04
1021595,315700	Unknown	+:Gypsy-34_GR-I|Vu07
1021595,319659	Unknown	C:Copia-21_GR-I|Vu10_2
1021595,322455	Unknown	C:L1-13_GM|Vu02
1021595,322711	Unknown	+:MuDr3_MT|Vu02_1
1021595,323099	Unknown	C:Gypsy-12_RC-I|Vu09
1021595,323460	Unknown	+:Copia-32_ECa-LTR|Vu02_1
1021595,323685	Unknown	C:hAT-2_GM|Vu08_3
1021595,323914	Unknown	C:hAT-2_GM|Vu10_2
1021595,324466	Unknown	C:TGM1_GM|Vu04
1021595,324685	Unknown	+:MuDr4_MT|Vu03_1
1021595,327073	Unknown	+:MuDr3_MT|Vu02_1
1021595,330253	Unknown	C:hAT-2_GM|Vu11_4
1021595,331872	Unknown	+:MuDr3_MT|Vu01_1
1021595,337155	Unknown	+:EnSpm-1_JC|Vu07
1021595,338770	Unknown	C:L1-13_GM|Vu11
1021595,342640	Unknown	+:L1-1_STu|Vu03_47
1021595,343881	Unknown	+:Gypsy-34_GR-I|Vu07
1021595,343884	Unknown	C:Gypsy-10_MN-LTR|Vu07
1021595,345760	Unknown	C:TGM1_GM|Vu04
1021595,347918	Unknown	C:Copia-21_GR-I|Vu09
1021595,348842	Unknown	C:hAT-2_GM|Vu09_6
1021595,349602	Unknown	C:Gypsy-118_GM-I|Vu04
1021595,352838	Unknown	+:EnSpm-1_TC|Vu08_1

1021595,353189	Unknown	+:Gypsy-34_Mad-I|Vu02
1021595,355707	Unknown	C:Gypsy-119_GM-I|Vu09_3
1021595,356406	Unknown	C:hAT-2_GM|Vu09_1
1021595,356940	Unknown	C:MtPH-M-3-Ia|Vu09
1021595,358562	Unknown	+:hAT-2_GM|Vu07_5
1021595,359180	Unknown	C:L1-13_GM|Vu01
1021595,359535	Unknown	C:hAT-2_GM|Vu07_2
1021595,364480	Unknown	C:Gypsy-6_GAr-I|Vu04
1021595,365439	Unknown	C:hAT-2_GM|Vu10_1
1021595,367376	Unknown	C:hAT-3_GM|Vu03_12
1021595,373218	Unknown	C:Copia-21_GR-I|Vu11
1021595,373483	Unknown	C:TGM5_GM|Vu10
1021595,375425	Unknown	C:RAM9B_I|Vu09
1021595,379310	Unknown	C:GmGYPSY10_I|Vu07_2
1021595,379450	Unknown	C:hAT-2_GM|Vu09_4
1021595,380965	Unknown	C:MuDr4_MT|Vu04
1021595,382764	Unknown	+:MuDr4_MT|Vu03_1
1021595,383051	Unknown	+:Copia-21_GR-I|Vu11
1021595,385095	Unknown	C:Gypsy-35_GAr-I|Vu01
1021595,385518	Unknown	+:MUDRAV2_MT|Vu07
1021595,390086	Unknown	C:Sat-1_CPa|Vu07_6
1021595,390417	Unknown	C:MuDr2_MT|Vu08
1021595,394174	Unknown	C:Gypsy-118_GM-I|Vu04
1021595,396800	Unknown	C:Gypsy-45_BRa-I|Vu10
1021595,398170	Unknown	+:Gypsy-21_ST-I|Vu08
1021595,400481	Unknown	+:Gypsy-76_NS-I|Vu05
1021595,401075	Unknown	C:Gypsy-118_GM-I|Vu04
1021595,401774	Unknown	+:MuDr4_MT|Vu09
1021595,403732	Unknown	C:hAT-2_GM|Vu03_13
1021595,404384	Unknown	C:hAT-2_GM|Vu03_2
1021595,406023	Unknown	+:Copia-63_GM-I|Vu04
1021595,407118	Unknown	+:Copia-21_GR-I|Vu08
1021595,411176	Unknown	C:hAT-2_ALy|Vu01
1021595,411881	Unknown	+:MuDr3_MT|Vu02_1
1021595,413163	Unknown	C:hAT-2_GM|Vu01_4
1021595,418229	Unknown	+:hAT-2_GM|Vu10_3
1021595,422326	Unknown	+:hAT-2_GM|Vu05_10
1021595,422980	Unknown	+:hAT-3_GM|Vu03_10
1021595,425467	Unknown	+:Gypsy-18_JC-LTR|Vu01
1021595,430854	Unknown	C:TGM5_GM|Vu10
1021595,432275	Unknown	C:hAT-2_GM|Vu10
1021595,435273	Unknown	C:hAT-2_GM|Vu03_13
1021595,437589	Unknown	C:hAT-2_GM|Vu07
1021595,440210	Unknown	+:L1-1_STu|Vu05_208

1021595,440640	Unknown	+:INMU1|Vu05
1021595,441462	Unknown	C:hAT-2_GM|Vu07
1021595,446811	Unknown	C:Gypsy-21_ST-I|Vu04
1021595,452791	Unknown	+:hAT-2_GM|Vu01_6
1021595,456840	Unknown	C:hAT-2_GM|Vu06_5
1021595,457811	Unknown	C:hAT-3_GM|Vu06_6
1021595,459095	Unknown	+:Gypsy-21_ST-I|Vu09
1021595,459521	Unknown	+:hAT-2_GM|Vu09_5
1021595,461651	Unknown	C:hAT-2_GM|Vu05_7
1021595,466224	Unknown	+:Gypsy-18_GR-I|Vu05
1021595,466317	Unknown	C:hAT-2_GM|Vu08_3
1021595,466565	Unknown	C:TGM5_GM|Vu10
1021595,468155	Unknown	C:hAT-2_GM|Vu08_2
1021595,471210	Unknown	C:TGM1_GM|Vu04
1021595,475458	Unknown	C:Gypsy-10_MN-LTR|Vu07
1021595,475960	Unknown	+:Gypsy-18_JC-LTR|Vu01
1021595,477733	Unknown	+:hAT-2_GM|Vu09_4
1021595,485206	Unknown	C:GYPOT1_I|Vu02
1021595,485913	Unknown	C:hAT-2_GM|Vu02_2
1021595,488421	Unknown	C:MtPH-M-3-Ia|Vu09
1021595,495153	Unknown	C:hAT-3_GM|Vu03_12
1021595,496185	Unknown	+:Gypsy-18_GR-I|Vu05
1021595,496604	Unknown	C:hAT-2_GM|Vu08_3
1021595,496831	Unknown	C:Gypsy-12_RC-I|Vu09
1021595,497135	Unknown	C:hAT-2_GM|Vu11_3
1021595,498817	Unknown	+:hAT-2_GM|Vu09
1021595,506630	Unknown	+:GYPOT1_I|Vu01
1021595,509226	Unknown	+:MUDRAV2_MT|Vu07
1021595,512335	Unknown	+:hAT-2_GM|Vu04_3
1021595,513663	Unknown	+:MuDr3_MT|Vu01
1021595,513919	Unknown	+:Gypsy-34_Mad-I|Vu02
1021595,514067	Unknown	C:Copia-21_GR-I|Vu11
1021595,514505	Unknown	+:Gypsy-18_JC-LTR|Vu01
1021595,514910	Unknown	C:Copia-21_GR-I|Vu06
1021595,518660	Unknown	+:Gypsy-70_MN-I|Vu05
1021595,518692	Unknown	+:EnSpm-1_GM|Vu05
1021595,518784	Unknown	+:Gypsy-119_GM-I|Vu10
1021595,519740	Unknown	+:MuDr3_MT|Vu01
1021595,525678	Unknown	+:Gypsy-21_ST-I|Vu03
1021595,528642	Unknown	C:Copia-21_GR-I|Vu10_2
1021595,529360	Unknown	C:hAT-4_GM|Vu03
1021595,532121	Unknown	C:hAT-2_GM|Vu10_2
1021595,534453	Unknown	C:Gypsy-60_GR-I|Vu09
1021595,534551	Unknown	+:Gypsy-18_JC-LTR|Vu01

1021595,536797	Unknown	C:hAT-2_GM|Vu03_8
1021595,537204	Unknown	C:hAT-2_GM|Vu02_2
1021595,537289	Unknown	+:MuDr4_MT|Vu09
1021595,537650	Unknown	+:hAT-6_TC|Vu10
1021595,538979	Unknown	+:EnSpm-1_JC|Vu07
1021595,539589	Unknown	+:hAT-N2_FV|Vu11
1021595,540095	Unknown	+:MuDr4_MT|Vu03_1
1021595,540598	Unknown	C:Gypsy-119_GM-I|Vu09_3
1021595,543120	Unknown	C:Gypsy-21_ST-I|Vu04
1021595,549060	Unknown	C:Gypsy-119_GM-I|Vu10
1021595,558585	Unknown	C:Gypsy-21_ST-I|Vu09
1021595,561253	Unknown	C:TGM1_GM|Vu11
1021595,563850	Unknown	C:hAT-2_GM|Vu08_1
1021595,564035	Unknown	C:hAT-2_GM|Vu08_2
1021595,565026	Unknown	+:Gypsy-21_ST-I|Vu03
1021595,565421	Unknown	C:hAT-3_GM|Vu07_2
1021595,565517	Unknown	C:Gypsy-118_GM-I|Vu04
1021595,567966	Unknown	+:hAT-2_GM|Vu07_8
1021595,568745	Unknown	C:MuDr3_MT|Vu04_1
1021595,570570	Unknown	C:Copia-21_GR-I|Vu09
1021595,574232	Unknown	C:Copia-21_GR-I|Vu11
1021595,576632	Unknown	+:hAT-6_TC|Vu05
1021595,576765	Unknown	C:Gypsy-118_GM-I|Vu04
1021595,576770	Unknown	+:Gypsy-6_GAr-I|Vu03
1021595,576876	Unknown	+:EnSpm-3_GM|Vu03
1021595,578607	Unknown	C:MuDr3_MT|Vu11
1021595,579450	Unknown	+:Copia-21_GR-I|Vu03_1
1021595,582330	Unknown	+:RAM9B_I|Vu10
1021595,584953	Unknown	+:MuDr4_MT|Vu03
1021595,587728	Unknown	C:hAT-2_GM|Vu07_1
1021595,587940	Unknown	+:EnSpm-3_GM|Vu04
1021595,588420	Unknown	C:hAT-2_GM|Vu02_2
1021595,588690	Unknown	C:Gypsy-35_GAr-I|Vu01
1021595,589572	Unknown	C:Sat-1_CPa|Vu07_6
1021595,591470	Unknown	C:L1-13_GM|Vu02
1021595,591868	Unknown	C:MuDr4_MT|Vu04
1021595,592147	Unknown	C:L1-13_GM|Vu02
1021595,592529	Unknown	C:MUDRB_PT|Vu03
1021595,593763	Unknown	C:hAT-3_GM|Vu06_2
1021595,594599	Unknown	+:hAT-2_GM|Vu08_1
1021595,595030	Unknown	+:MuDr3_MT|Vu01_1
1021595,595753	Unknown	C:MuDR-8_GM|Vu01
1021595,596302	Unknown	+:ENSPM2_MT|Vu10
1021595,596994	Unknown	+:hAT-2_GM|Vu09_5

1021595,597773	Unknown	C:GmGYPSY10_I|Vu07_2
1021595,598380	Unknown	+:hAT-2_GM|Vu05_7
1021595,600210	Unknown	+:MuDr4_MT|Vu09
1021595,602272	Unknown	+:MuDr3_MT|Vu01
1021595,612838	Unknown	+:hAT-2_GM|Vu08_2
1021595,615190	Unknown	C:Gypsy-4_CP-I|Vu02
1021595,616759	Unknown	+:EnSpm-3_GM|Vu04
1021595,616768	Unknown	+:Copia-21_GR-I|Vu06
1021595,618257	Unknown	C:hAT-2_GM|Vu03_14
1021595,618400	Unknown	C:Copia-72_VV-LTR|Vu04_1
1021595,620439	Unknown	C:L1-13_GM|Vu02
1021595,620655	Unknown	C:L1-13_GM|Vu02
1021595,621291	Unknown	+:MuDr4_MT|Vu08
1021595,622152	Unknown	+:INMU1|Vu05
1021595,623187	Unknown	+:Gypsy-34_GR-I|Vu07
1021595,627621	Unknown	+:MuDr3_MT|Vu01_1
1021595,628223	Unknown	C:hAT-3_GM|Vu06_6
1021595,629871	Unknown	+:hAT-2_GM|Vu01_4
1021595,632350	Unknown	C:MuDr4_MT|Vu04
1021595,634738	Unknown	C:Sat-1_CPa|Vu07_6
1021595,639627	Unknown	+:L1-1_STu|Vu05_49
1021595,643151	Unknown	+:Gypsy-13_GAr-I|Vu11
1021595,673200	Unknown	C:TONT2-I_PV|Vu09_10
1021595,734020	Unknown	C:hAT-3_GM|Vu07_2
1021595,734880	Unknown	+:EnSpm-3_GM|Vu04
1021595,766580	Unknown	+:EnSpm-1_JC|Vu07
1021595,831130	Unknown	C:L1-13_GM|Vu01
1021595,854880	Unknown	C:hAT-2_GM|Vu09_6
1021595,855640	Unknown	C:GmGYPSY11_I|Vu09_1
1021595,891320	Unknown	C:hAT-6_TC|Vu08
1021595,895500	Unknown	C:HAT1_MT|Vu04
1021595,905400	Unknown	C:hAT-2_ALy|Vu01
1021595,940320	Unknown	C:hAT-2_GM|Vu05_7
1021595,949730	Unknown	+:EnSpm-1_GM|Vu05
1021595,961690	Unknown	+:hAT-4_GM|Vu05
1021595,970210	Unknown	C:Sat-2_STu|Vu11
1021595,985240	Unknown	C:Copia-21_GR-I|Vu06_2
1021595,986600	Unknown	+:Copia-13_Cia-I|Vu01

Material suplementar 4: Transcritos diferencialmente expressos anotados contra elementos transponíveis (TEs). Transcritos são indicados pelo respectivo ID e estão seguidos do referente *lucus* no genoma. A expressão de cada isoforma, em formato log2FC (Fold Change), destaca isoformas induzidas (vermelho) e reprimidas (azul). Tabelas disponíveis para os dois tempos de cada experimento (IT85F-2687 inoculado com Cowpea Severe Mosaic Virus *(CPSMV*), BR14-Mulato inoculado com Cowpea Aphid-Born Mosaic Virus (CABMV) e Pingo de Ouro sob desidratação radicular).

Material suplementar 5: Anotação funcional dos elementos transponíveis (TEs) diferencialmente expressos. TEs são indicados pelo respectivo ID e com a informação sobre a classificação da superfamíliaindicada no artigo, expressão em formato log2 fold change (FC), a lista com o ID de cada TE que apresentou match com o transcrito com no mínimo 80% de identidade e cobertura. A anotação dos termos GO geradas com o trinotate a partir de dados do PFAN e Uniprot estão descritas nas últimas colunas da tabela, bem como o resultado do BLAST contra o uniprot, este apenas usado para anotar sequências não anotadas análise de ontologia. Tabelas disponíveis para os dois tempos de cada experimento (IT85F-2687 inoculado com Cowpea Severe Mosaic Virus *(CPSMV*), BR14-Mulato inoculado com Cowpea Aphid-Born Mosaic Virus (CABMV) e Pingo de Ouro sob desidratação radicular).

Material suplementar 6: A descrição de todo o material vegetal (1 a 12) está separados por triplicatas biológicas e técnincas, cada um dos três tratamentos pode ser distinguido por cor. CpSMV (vírus) em verde, CABMV (vírus) em amarelo e Desidratação radicular em azul. As diferentes tonalidades distinguem controle e tratamento. Cada tratamento possui dois tempos de duração (1h e 16h / 25 min e 150 min). Estão destacados também o horário de início do estresse, bem como horário de coleta do referido tecido.

Planta	C-value (genoma)	Hora do experimento		Stress	Condição	Duração	biblioteca	Replica Biológica	Replica Tec	Abreviação amostra	Run Type	Tecido	Estágio da planta
		Início	Coleta										
V. unguiculata (IT85F-2687)	0.6	16h	17h	CpSMV	Control	1 h	**37**	1.1	**37aMVct60mR1**	**37aMVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									37aMVct60mR2		PAIRED		
									37bMVct60mR1	**37bMVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									37bMVct60mR2		PAIRED		
V. unguiculata (IT85F-2687)	0.6	16h	17h	CpSMV	Control	1 h	**38**	1.2	**38aMVct60mR1**	**38aMVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									38aMVct60mR2		PAIRED		
									38bMVct60mR1	**38bMVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									38bMVvt60mR2		PAIRED		
V. unguiculata (IT85F-2687)	0.6	16h	17h	CpSMV	Control	1 h	**39**	1.3	**39aMVct60mR1**	**39aMVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									39aMVct60mR2		PAIRED		
									39bMVct60mR1	**39bMVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									39bMVct60mR2		PAIRED		
V. unguiculata (IT85F-2687)	0.6	16h	17h	CpSMV	virus	1 h	**40**	2.1	**40aMVvi60mR1**	**40aMVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									40aMVvi60mR2		PAIRED		

V. unguiculata (IT85F-2687)	0.6	16h	17h	CpSMV	virus	1 h	**41**	2.2	**40bMVvi60mR1**	**40bMVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									40bMVvi60mR2		PAIRED		
									41aMV-vi60mR1	**41aMVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									41aMVvi60mR2		PAIRED		
									41bMVvi60mR1	**41bMVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									41bMVvi60mR2		PAIRED		
V. unguiculata (IT85F-2687)	0.6	16h	17h	CpSMV	virus	1 h	**42**	2.3	**42aMVvi60mR1**	**42aMVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									42aMVvi60mR2		PAIRED		
									42bMVvi60mR1	**42bMVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									42bMVvi60mR2		PAIRED		
V. unguiculata (IT85F-2687)	0.6	16h	8h	CpSMV	Control	16 h	**43**	3.1	**43aMVct16hR1**	**43aMVct16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									43aMVct16hR2		PAIRED		
									43bMVct16hR1	**43bMVct16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									43bMVct16hR2		PAIRED		
V. unguiculata (IT85F-2687)	0.6	16h	8h	CpSMV	Control	16 h	**44**	3.2	**44aMVct16hR1**	**44aMVct16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									44aMVct16hR2		PAIRED		
									44bMVct16hR1	**44bMVct16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									44bMVct16hR2		PAIRED		
V. unguiculata (IT85F-2687)	0.6	16h	8h	CpSMV	Control	16 h	**45**	3.3	**45aMVct16hR1**	**45aMVct16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									45aMVct16hR2		PAIRED		
									45bMVct16hR1	**45bMVct16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									45bMVct16hR2		PAIRED		
V. unguiculata (IT85F-2687)	0.6	16h	8h	CpSMV	virus	16 h	**46**	4.1	**46aMVvi16hR1**	**46aMVvi6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									46aMVvi16hR2		PAIRED		

V. unguiculata (IT85F-2687)	0.6	16h	8h	CpSMV	virus	16 h	**47**	4.2	**46bMVvi16hR1**	**46bMVvi6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									46bMVvi16hR2		PAIRED		
									47aMVvi16hR1	**47aMVvi6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									47aMVvi16hR2		PAIRED		
									47bMVvi16hR1	**47bMVvi6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									47bMVvi16hR2		PAIRED		
V. unguiculata (IT85F-2687)	0.6	16h	8h	CpSMV	virus	16 h	**48**	4.3	**48aMVvi16hR1**	**48aMVvi6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									48aMVvi16hR2		PAIRED		
									48bMVvi16hR1	**48bMVvi6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									48bMVvi16hR2		PAIRED		
V. unguiculata (BR14-Mulato)	0.6	16h	17h	CABMV	Control	1 h	**25**	5.1	**25aPVct60mR1**	**25aPVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									25aPVct60mR2		PAIRED		
									25bPVct60mR1	**25bPVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									25bPVct60mR2		PAIRED		
V. unguiculata (BR14-Mulato)	0.6	16h	17h	CABMV	Control	1 h	**26**	5.2	**26aPVct60mR1**	**26aPVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									26aPVct60mR2		PAIRED		
									26bPVct60mR1	**26bPVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									26bPVct60mR2		PAIRED		
V. unguiculata (BR14-Mulato)	0.6	16h	17h	CABMV	Control	1 h	**27**	5.3	**27aPVct60mR1**	**27aPVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									27aPVct60mR2		PAIRED		
									27bPVct60mR1	**27bPVct60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									27bPVct60mR2		PAIRED		
V. unguiculata (BR14-Mulato)	0.6	16h	17h	CABMV	virus	1 h	**28**	6.1	**28aPVvi60mR1**	**28aPVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)

									28aPVvi60mR2		PAIRED		
									28bPVvi60mR1	**28bPVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									28bPVvi60mR2		PAIRED		
		16h	17h						**29aPVvi60mR1**	**29aPVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (BR14-Mulato)	0.6			CABMV	virus	1 h	**29**	6.2	**29aPVvi60mR2**		PAIRED		
									29bPVvi60mR1	**29bPVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									29bPVvi60mR2		PAIRED		
		16h	17h						**30aPVvi60mR1**	**30aPVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (BR14-Mulato)	0.6			CABMV	virus	1 h	**30**	6.3	**30aPVvi60mR2**		PAIRED		
									30bPVvi60mR1	**30bPVvi60m**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									30bPVvi60mR2		PAIRED		
		16h	8h						**31aPVct16hR1**	**31aPVct6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (BR14-Mulato)	0.6			CABMV	Control	16 h	**31**	7.1	**31aPVct16hR2**		PAIRED		
									31bPVct16hR1	**31bPVct6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									31bPVct16hR2		PAIRED		
		16h	8h						**32aPVct16hR1**	**32aPVct6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (BR14-Mulato)	0.6			CABMV	Control	16 h	**32**	7.2	**32aPVct16hR2**		PAIRED		
									32bPVct16hR1	**32bPVct6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									32bPVct16hR2		PAIRED		
		16h	8h						**33aPVct16hR1**	**33aPVct6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (BR14-Mulato)	0.6			CABMV	Control	16 h	**33**	7.3	**33aPVct16hR2**		PAIRED		
									33bPVct16hR1	**33bPVct6h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									33bPVct16hR2		PAIRED		
V. unguiculata (BR14-Mulato)	0.6	16h	8h	CABMV	virus	16 h	**34**	8.1	**34aPVvi16hR1**	**34aPVvi16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)

									34aPVvi16hR2		PAIRED		
									34bPVvi16hR1	**34bPVvi16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									34bPVvi16hR2		PAIRED		
		16h	8h						**35aPVvi16hR1**	**35aPVvi16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (BR14-Mulato)	0.6			CABMV	virus	16 h	**35**	8.2	**35aPVvi16hR2**		PAIRED		
									35bPVvi16hR1	**35bPVvi16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									35bPVvi16hR2		PAIRED		
		16h	8h						**36aPVvi16hR1**	**36aPVvi16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (BR14-Mulato)	0.6			CABMV	virus	16 h	**36**	8.3	**36aPVvi16hR2**		PAIRED		
									36bPVvi16hR1	**36bPVvi16h**	PAIRED	Folha	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									36bPVvi16hR2		PAIRED		
		-	-						**A2aPOct025mR1**	**A2aPOct025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (Pingo de Ouro)	0.6			Desidratação radicular	controle	25m	**2**	9.1	**A2aPOct025mR2**		PAIRED		
									B2bPOct025mR1	**B2bPOct025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									B2bPOct025mR2		PAIRED		
		-	-						**C1aPOct025mR1**	**C1aPOct025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (Pingo de Ouro)	0.6			Desidratação radicular	controle	25m	**1**	9.2	**C1aPOct025mR2**		PAIRED		
									D1bPOct025mR1	**D1bPOct025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									D1bPOct025mR2		PAIRED		
		-	-						**E11aPOct025mR1**	**E11aPOct025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (Pingo de Ouro)	0.6			Desidratação radicular	controle	25m	**11**	9.3	**E11aPOct025mR2**		PAIRED		
									F11bPOct025mR1	**F11bPOct025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									F11bPOct025mR2		PAIRED		
V. unguiculata	0.6	-	-	Desidratação	Desidratação	25m	**8**	10.1	**G8aPOhy025mR1**	**G8aPOhy025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com

(Pingo de Ouro)			radicular									folíolos completamente abertos)
								G8aPOhy025mR2		PAIRED		
								H8bPOhy025mR1	**H8bPOhy025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
								H8bPOhy025mR2		PAIRED		
		- -						**I12aPOhy025mR1**	**I12aPOhy025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (Pingo de Ouro)	0.6		Desidratação radicular	Desidratação	25m	**12**	10.2	**I12aPOhy025mR2**		PAIRED		
								J12bPOhy025mR1	**J12bPOhy025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
								J12bPOhy025mR2		PAIRED		
		- -						**K9aPOhy025mR1**	**K9aPOhy025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (Pingo de Ouro)	0.6		Desidratação radicular	Desidratação	25m	**9**	10.3	**K9aPOhy025mR2**		PAIRED		
								L9bPOhy025mR1	**L9bPOhy025m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
								L9bPOhy025mR2		PAIRED		
		- -						**M5aPOct150mR1**	**M5aPOct150m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (Pingo de Ouro)	0.6		Desidratação radicular	controle	150m	**5**	11.1	**M5aPOct150mR2**		PAIRED		
								N5bPOct150mR1	**N5bPOct150m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
								N5bPOct150mR2		PAIRED		
		- -						**O3aPOct150mR1**	**O3aPOct150m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (Pingo de Ouro)	0.6		Desidratação radicular	controle	150m	**3**	11.2	**O3aPOct150mR2**		PAIRED		
								P3bPOct150mR1	**P3bPOct150m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
								P3bPOct150mR2		PAIRED		
		- -						**Q4aPOct150mR1**	**Q4aPOct150m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (Pingo de Ouro)	0.6		Desidratação radicular	controle	150m	**4**	11.3	**Q4aPOct150mR2**		PAIRED		
								R4bPOct150mR1	**R4bPOct150m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
								R4bPOct150mR2		PAIRED		

		-	-					**S10aPOhy150mR1**	**S10aPOhy150m**	PAIRED	Raiz	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
V. unguiculata (Pingo de Ouro)	0.6			Desidratação radicular	Desidratação	150m	**10**	12.1	**S10aPOhy150mR2**		PAIRED	
									T10bPOhy150mR1	**T10bPOhy150m**	PAIRED	Raiz
									T10bPOhy150mR2		PAIRED	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									U7aPOhy150mR1	**U7aPOhy150m**	PAIRED	Raiz
V. unguiculata (Pingo de Ouro)	0.6			Desidratação radicular	Desidratação	150m	**7**	12.2	**U7aPOhy150mR2**		PAIRED	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									V7bPOhy150mR1	**V7bPOhy150m**	PAIRED	Raiz
									V7bPOhy150mR2		PAIRED	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									X6aPOhy150mR1	**X6aPOhy150m**	PAIRED	Raiz
V. unguiculata (Pingo de Ouro)	0.6			Desidratação radicular	Desidratação	150m	**6**	12.3	**X6aPOhy150mR2**		PAIRED	V2 (terceiro nó do ramo principal com folíolos completamente abertos)
									Z6bPOhy150mR1	**Z6bPOhy150m**	PAIRED	Raiz
									Z6bPOhy150mR2		PAIRED	V2 (terceiro nó do ramo principal com folíolos completamente abertos)

Material Suplementar 7: Lista de parâmetros usados nos programas para montagem e análise do transcriptoma de *Vigna unguiculata* em relação ao seu genoma de referência (v1.1) disponível no Phytozome V12.

Program	Version	Parameters
RNA-SeQC	v1.1.8	--ttype 2
Trimmomatic	v0.36	--phred33 trailing=30 minlen=32
STAR	v3.5.3a	--outSAMunmapped within --outSAMattrRGline --outSAMtype BAM unsorted
Picard	v2.9	remove_sequencing_duplicates=False duplicate_scoring_startegy=sum_of_base_qualities
Cufflinks	v2.2.1	Cufflinks -max_bundle_frag=1000000 Cuffmerge Cuffdiff -u Cuffnorm
RepeatMasker	v4.0.7	-s -no_is -lib /path/to/RepBase/edcotrep.ref

ANEXO C - NORMAS DA REVISTA – BMC BIOINFORMATICS (SOFTWARE ARTICLE)

Criteria

Software articles should describe a tool likely to be of broad utility that represents a significant advance over previously published software (usually demonstrated by direct comparison with available related software).

Availability of software to reviewers and other researchers

The software application/tool described in the manuscript must be available for testing by reviewers in a way that preserves their anonymity. If published, software applications/tools must be freely available to any researcher wishing to use them for non-commercial purposes, without restrictions such as the need for a material transfer agreement. Because weblinks frequently become broken, *BMC Bioinformatics* strongly recommends that all software applications/tools are included with the submitted manuscript as additional files to ensure that the software will continue to be available.

BMC Bioinformatics recommends, but does not require, that the source code of the software should be made available under a suitable open-source license that will entitle other researchers to further develop and extend the software if they wish to do so. Typically, an archive of the source code of the current version of the software should be included with the submitted manuscript as a supplementary file. Since it is likely that the software will continue to be developed following publication, the manuscript should also include a link to the home page for the software project. For open source projects, we recommend that authors host their project with a recognized open-source repository such as bioinformatics.org or sourceforge.net

Should a description of a website be submitted as a software article or a database article?

Descriptions of websites and web-based tools should be submitted as software articles if the intention is that the software that drives the website will be

made available to other researchers to extend and use on other websites. On the other hand, if a website's functionality is closely tied to a specific database then the article should instead be submitted as a database article.

Preparing your manuscript

The information below details the section headings that you should include in your manuscript and what information should be within each section.

Please note that your manuscript must include a 'Declarations' section including all of the subheadings (please see below for more information).

Title page

The title page should:

present a title that includes, if appropriate, the study design e.g.:

"A versus B in the treatment of C: a randomized controlled trial", "X is a risk factor for Y: a case control study", "What is the impact of factor X on subject Y: A systematic review" or for non-clinical or non-research studies: a description of what the article reports list the full names, institutional addresses and email addresses for all authors if a collaboration group should be listed as an author, please list the Group name as an author. If you would like the names of the individual members of the Group to be searchable through their individual PubMed records, please include this information in the “Acknowledgements” section in accordance with the instructions below indicate the corresponding author.

Abstract

The Abstract should not exceed 350 words. Please minimize the use of abbreviations and do not cite references in the abstract. The abstract must include the following separate sections:

Background: the context and purpose of the study

Results: the main findings

Conclusions: a brief summary and potential implications

Keywords

Three to ten keywords representing the main content of the article.

Background

The Background section should explain the relevant context and the specific issue that the software described is intended to address.

Implementation

This should include a description of the overall architecture of the software implementation, along with details of any critical issues and how they were addressed.

Results

This should include the findings of the study including, if appropriate, results of statistical analysis which must be included either in the text or as tables and figures. This section may be combined with the Discussion section for Software articles.

Discussion (if appropriate)

The user interface should be described and a discussion of the intended uses of the software, and the benefits that are envisioned, should be included, together with data on how its performance and functionality compare with, and improve, on functionally similar existing software. A case study of the use of the software may be presented. The planned future development of new features, if any, should be mentioned.

Conclusions

This should state clearly the main conclusions and provide an explanation of the importance and relevance of the case, data, opinion, database or software reported.

Availability and requirements

Lists the following:

Project name: e.g. My bioinformatics project

Project home page: e.g. http://sourceforge.net/projects/mged

Operating system(s): e.g. Platform independent

Programming language: e.g. Java

Other requirements: e.g. Java 1.3.1 or higher, Tomcat 4.0 or higher

License: e.g. GNU GPL, FreeBSD etc.

Any restrictions to use by non-academics: e.g. licence needed

List of abbreviations

If abbreviations are used in the text they should be defined in the text at first use, and a list of abbreviations should be provided.

Declarations

All manuscripts must contain the following sections under the heading 'Declarations':

Ethics approval and consent to participate

Consent for publication

Availability of data and material

Competing interests

Funding

Authors' contributions

Acknowledgements

Authors' information (optional)

Please see below for details on the information to be included in these sections.

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

Ethics approval and consent to participate

Manuscripts reporting studies involving human participants, human data or human tissue must:

include a statement on ethics approval and consent (even where the need for approval was waived)

include the name of the ethics committee that approved the study and the committee’s reference number if appropriate

Studies involving animals must include a statement on ethics approval.

See our editorial policies for more information.

If your manuscript does not report on or involve the use of any animal or human data or tissue, this section is not applicable to your submission. Please state “Not applicable” in this section.

Consent for publication

If your manuscript contains any individual person’s data in any form (including details, images or videos), consent for publication must be obtained from that person, or in the case of children, their parent or legal guardian. All presentations of case reports must have consent for publication.

You can use your institutional consent form or our consent form if you prefer. You should not send the form to us on submission, but we may request to see a copy at any stage (including after publication).

See our editorial policies for more information on consent for publication.

If your manuscript does not contain data from any individual person, please state “Not applicable” in this section.

Availability of data and materials

All manuscripts must include an ‘Availability of data and materials’ statement. Data availability statements should include information on where data supporting the results reported in the article can be found including, where applicable, hyperlinks to publicly archived datasets analysed or generated during the study. By data we mean the minimal dataset that would be necessary to interpret, replicate and build upon the findings reported in the article. We recognise it is not always possible to share research data publicly, for instance when individual privacy could be compromised, and in such instances data availability should still be stated in the manuscript along with any conditions for access.

Data availability statements can take one of the following forms (or a combination of more than one if required for multiple datasets):

The datasets generated and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS]

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

All data generated or analysed during this study are included in this published article [and its supplementary information files].

The datasets generated and/or analysed during the current study are not publicly available due [REASON WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request.

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

The data that support the findings of this study are available from [third party name] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of [third party name].

Not applicable. If your manuscript does not contain any data, please state 'Not applicable' in this section.

More examples of template data availability statements, which include examples of openly available and restricted access datasets, are available here.

BioMed Central also requires that authors cite any publicly available data on which the conclusions of the paper rely in the manuscript. Data citations should include a persistent identifier (such as a DOI) and should ideally be included in the reference list. Citations of datasets, when they appear in the reference list, should include the minimum information recommended by DataCite and follow journal style. Dataset identifiers including DOIs should be expressed as full URLs. For example:

Hao Z, AghaKouchak A, Nakhjiri N, Farahmand A. Global integrated drought monitoring and prediction system (GIDMaPS) data sets. figshare. 2014. http://dx.doi.org/10.6084/m9.figshare.853801

With the corresponding text in the Availability of data and materials statement:

The datasets generated during and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS].[Reference number]

Competing interests

All financial and non-financial competing interests must be declared in this section.

See our editorial policies for a full explanation of competing interests. If you are unsure whether you or any of your co-authors have a competing interest please contact the editorial office.

Please use the author's initials to refer to each author's competing interests in this section.

If you do not have any competing interests, please state "The authors declare that they have no competing interests" in this section.

Funding

All sources of funding for the research reported should be declared. The role of the funding body in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript should be declared.

Authors' contributions

The individual contributions of authors to the manuscript should be specified in this section. Guidance and criteria for authorship can be found here.

Please use initials to refer to each author's contribution in this section, for example: "FC analyzed and interpreted the patient data regarding the hematological disease and the transplant. RH performed the histological examination of the kidney, and was a major contributor in writing the manuscript. All authors read and approved the final manuscript."

Acknowledgements

Please acknowledge anyone who contributed towards the article who does not meet the criteria for authorship including anyone who provided professional writing services or materials.

Authors should obtain permission to acknowledge from all those mentioned in the Acknowledgements section.

See our editorial policies for a full explanation of acknowledgements and authorship criteria.

If you do not have anyone to acknowledge, please write "Not applicable" in this section.

For Group Authorship (for manuscripts involving a collaboration group): if you would like the names of the individual members of a collaboration Group to be searchable through their individual PubMed records, please ensure that the title of the collaboration Group is included on the title page and in the submission system and also include collaborating author names as the last paragraph of the “Acknowledgements” section. Please add authors in the format First Name, Middle initial(s) (optional), Last Name. You can add institution or country information for each author if you wish, but this should be consistent across all authors.

Please note that individual names may not be present in the PubMed record at the time a published article is initially included in PubMed as it takes PubMed additional time to code this information.

Authors' information

This section is optional.

You may choose to use this section to include any relevant information about the author(s) that may aid the reader's interpretation of the article, and understand the standpoint of the author(s). This may include details about the authors' qualifications, current positions they hold at institutions or societies, or any other relevant background information. Please refer to authors using their initials. Note this section should not be used to describe any competing interests.

Endnotes

Endnotes should be designated within the text using a superscript lowercase letter and all notes (along with their corresponding letter) should be included in the Endnotes section. Please format this section in a paragraph rather than a list.

References

All references, including URLs, must be numbered consecutively, in square brackets, in the order in which they are cited in the text, followed by any in tables or legends. The reference numbers must be finalized and the reference list fully formatted before submission.

Examples of the BioMed Central reference style are shown below. Please ensure that the reference style is followed precisely.

See our editorial policies for author guidance on good citation practice.

Web links and URLs: All web links and URLs, including links to the authors' own websites, should be given a reference number and included in the reference list rather than within the text of the manuscript. They should be provided in full, including both the title of the site and the URL, as well as the date the site was accessed, in the following format: The Mouse Tumor Biology Database.

http://tumor.informatics.jax.org/mtbwi/index.do. Accessed 20 May 2013. If an author or group of authors can clearly be associated with a web link (e.g. for blogs) they should be included in the reference.

Example reference style:

Article within a journal

Smith JJ. The world of science. Am J Sci. 1999;36:234-5.

Article within a journal (no page numbers)

Rohrmann S, Overvad K, Bueno-de-Mesquita HB, Jakobsen MU, Egeberg R, Tjønneland A, et al. Meat consumption and mortality - results from the European Prospective Investigation into Cancer and Nutrition. BMC Med. 2013;11:63.

Article within a journal by DOI

Slifka MK, Whitton JL. Clinical implications of dysregulated cytokine production. Dig J Mol Med. 2000; doi:10.1007/s801090000086.

Article within a journal supplement

Frumin AM, Nussbaum J, Esposito M. Functional asplenia: demonstration of splenic activity by bone marrow scan. Blood 1979;59 Suppl 1:26-32.

Book chapter, or an article within a book

Wyllie AH, Kerr JFR, Currie AR. Cell death: the significance of apoptosis. In: Bourne GH, Danielli JF, Jeon KW, editors. International review of cytology. London: Academic; 1980. p. 251-306.

OnlineFirst chapter in a series (without a volume designation but with a DOI)

Saito Y, Hyuga H. Rate equation approaches to amplification of enantiomeric excess and chiral symmetry breaking. Top Curr Chem. 2007. doi:10.1007/128_2006_108.

Complete book, authored

Blenkinsopp A, Paxton P. Symptoms in the pharmacy: a guide to the management of common illness. 3rd ed. Oxford: Blackwell Science; 1998.

Online document

Doe J. Title of subordinate document. In: The dictionary of substances and their effects. Royal Society of Chemistry. 1999. http://www.rsc.org/dose/title of subordinate document. Accessed 15 Jan 1999.

Online database

Healthwise Knowledgebase. US Pharmacopeia, Rockville. 1998. http://www.healthwise.org. Accessed 21 Sept 1998.

Supplementary material/private homepage

Doe J. Title of supplementary material. 2000. http://www.privatehomepage.com. Accessed 22 Feb 2000.

University site

Doe, J: Title of preprint. http://www.uni-heidelberg.de/mydata.html (1999). Accessed 25 Dec 1999.

FTP site

Doe, J: Trivial HTTP, RFC2169. ftp://ftp.isi.edu/in-notes/rfc2169.txt (1999). Accessed 12 Nov 1999.

Organization site

ISSN International Centre: The ISSN register. http://www.issn.org (2006). Accessed 20 Feb 2007.

Dataset with persistent identifier

Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, et al. Genome data from sweet and grain sorghum (Sorghum bicolor). GigaScience Database. 2011. http://dx.doi.org/10.5524/100012.

ANEXO D - NORMAS DA REVISTA – BMC BIOINFORMATICS (RESEARCH ARTICLE)

Criteria

Research articles should report on original primary research, but may report on systematic reviews of published research provided they adhere to the appropriate reporting guidelines which are detailed in our editorial policies. Please note that non-commissioned pooled analyses of selected published research will not be considered.

BMC Bioinformatics strongly encourages that all datasets on which the conclusions of the paper rely should be available to readers. We encourage authors to ensure that their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible. Please see Springer Nature's information on recommended repositories. Where a widely established research community expectation for data archiving in public repositories exists, submission to a community-endorsed, public repository is mandatory. A list of data where deposition is required, with the appropriate repositories, can be found on the Editorial Policies Page. **Preparing your manuscript**

The information below details the section headings that you should include in your manuscript and what information should be within each section.

Please note that your manuscript must include a 'Declarations' section including all of the subheadings (please see below for more information).

Title page

The title page should:

present a title that includes, if appropriate, the study design

list the full names, institutional addresses and email addresses for all authors

if a collaboration group should be listed as an author, please list the Group name as an author. If you would like the names of the individual members of the Group to be searchable through their individual PubMed records, please include this information in the "Acknowledgements" section in accordance with the instructions below

indicate the corresponding author

Abstract

The Abstract should not exceed 350 words. Please minimize the use of abbreviations and do not cite references in the abstract. The abstract must include the following separate sections:

Background: the context and purpose of the study

Results: the main findings

Conclusions: a brief summary and potential implications

Keywords

Three to ten keywords representing the main content of the article.

Background

The Background section should explain the background to the study, its aims, a summary of the existing literature and why this study was necessary.

Results

This should include the findings of the study including, if appropriate, results of statistical analysis which must be included either in the text or as tables and figures.

Discussion

For research articles this section should discuss the implications of the findings in context of existing research and highlight limitations of the study. For study protocols and methodology manuscripts this section should include a discussion of any practical or operational issues involved in performing the study and any issues not covered in other sections.

Conclusions

This should state clearly the main conclusions and provide an explanation of the importance and relevance of the study to the field.

Methods

The methods section should include:

the aim, design and setting of the study the characteristics of participants or description of materials a clear description of all processes, interventions and comparisons. Generic names should generally be used. When proprietary brands are used in research, include the brand names in parentheses the type of statistical analysis used, including a power calculation if appropriate

List of abbreviations

If abbreviations are used in the text they should be defined in the text at first use, and a list of abbreviations can be provided.

Declarations

All manuscripts must contain the following sections under the heading 'Declarations':

Ethics approval and consent to participate

Consent for publication

Availability of data and material

Competing interests

Funding

Authors' contributions

Acknowledgements

Authors' information (optional)

Please see below for details on the information to be included in these sections.

If any of the sections are not relevant to your manuscript, please include the heading and write 'Not applicable' for that section.

Ethics approval and consent to participate

Manuscripts reporting studies involving human participants, human data or human tissue must:

include a statement on ethics approval and consent (even where the need for approval was waived)

include the name of the ethics committee that approved the study and the committee’s reference number if appropriate

Studies involving animals must include a statement on ethics approval.

See our editorial policies for more information.

If your manuscript does not report on or involve the use of any animal or human data or tissue, please state “Not applicable” in this section.

Consent for publication

If your manuscript contains any individual person’s data in any form (including any individual details, images or videos), consent for publication must be obtained from that person, or in the case of children, their parent or legal guardian. All presentations of case reports must have consent for publication.

You can use your institutional consent form or our consent form if you prefer. You should not send the form to us on submission, but we may request to see a copy at any stage (including after publication).

See our editorial policies for more information on consent for publication.

If your manuscript does not contain data from any individual person, please state “Not applicable” in this section.

Availability of data and materials

All manuscripts must include an ‘Availability of data and materials’ statement. Data availability statements should include information on where data supporting the results reported in the article can be found including, where applicable, hyperlinks to publicly archived datasets analysed or generated during the study. By data we mean the minimal dataset that would be necessary to interpret, replicate and build upon the findings reported in the article. We recognise it is not always possible to share research data publicly, for instance when individual privacy could be compromised, and in such instances data availability should still be stated in the manuscript along with any conditions for access.

Data availability statements can take one of the following forms (or a combination of more than one if required for multiple datasets):

The datasets generated and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS]

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

All data generated or analysed during this study are included in this published article [and its supplementary information files].

The datasets generated and/or analysed during the current study are not publicly available due [REASON WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request.

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

The data that support the findings of this study are available from [third party name] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of [third party name].

Not applicable. If your manuscript does not contain any data, please state 'Not applicable' in this section.

More examples of template data availability statements, which include examples of openly available and restricted access datasets, are available here.

BioMed Central also requires that authors cite any publicly available data on which the conclusions of the paper rely in the manuscript. Data citations should include a persistent identifier (such as a DOI) and should ideally be included in the reference list. Citations of datasets, when they appear in the reference list, should include the minimum information recommended by DataCite and follow journal style. Dataset identifiers including DOIs should be expressed as full URLs. For example:

Hao Z, AghaKouchak A, Nakhjiri N, Farahmand A. Global integrated drought monitoring and prediction system (GIDMaPS) data sets. figshare. 2014. http://dx.doi.org/10.6084/m9.figshare.853801

With the corresponding text in the Availability of data and materials statement:

The datasets generated during and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS].[Reference number]

Competing interests

All financial and non-financial competing interests must be declared in this section.

See our editorial policies for a full explanation of competing interests. If you are unsure whether you or any of your co-authors have a competing interest please contact the editorial office.

Please use the authors initials to refer to each authors' competing interests int his section.

If you do not have any competing interests, please state "The authors declare that they have no competing interests" in this section.

Funding

All sources of funding for the research reported should be declared. The role of the funding body in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript should be declared.

Authors' contributions

The individual contributions of authors to the manuscript should be specified in this section. Guidance and criteria for authorship can be found in our editorial policies.

Please use initials to refer to each author's contribution in this section, for example: "FC analyzed and interpreted the patient data regarding the hematological disease and the transplant. RH performed the histological examination of the kidney, and was a major contributor in writing the manuscript. All authors read and approved the final manuscript."

Acknowledgements

Please acknowledge anyone who contributed towards the article who does not meet the criteria for authorship including anyone who provided professional writing services or materials.

Authors should obtain permission to acknowledge from all those mentioned in the Acknowledgements section.

See our editorial policies for a full explanation of acknowledgements and authorship criteria.

If you do not have anyone to acknowledge, please write "Not applicable" in this section.

Group authorship (for manuscripts involving a collaboration group): if you would like the names of the individual members of a collaboration Group to be searchable through their individual PubMed records, please ensure that the title of the collaboration Group is included on the title page and in the submission system and also include collaborating author names as the last paragraph of the “Acknowledgements” section. Please add authors in the format First Name, Middle initial(s) (optional), Last Name. You can add institution or country information for each author if you wish, but this should be consistent across all authors.

Please note that individual names may not be present in the PubMed record at the time a published article is initially included in PubMed as it takes PubMed additional time to code this information.

Authors' information

This section is optional.

You may choose to use this section to include any relevant information about the author(s) that may aid the reader's interpretation of the article, and understand the standpoint of the author(s). This may include details about the authors' qualifications, current positions they hold at institutions or societies, or any other relevant background information. Please refer to authors using their initials. Note this section should not be used to describe any competing interests.

Endnotes

Endnotes should be designated within the text using a superscript lowercase letter and all notes (along with their corresponding letter) should be included in the Endnotes section. Please format this section in a paragraph rather than a list.

References

All references, including URLs, must be numbered consecutively, in square brackets, in the order in which they are cited in the text, followed by any in tables or legends. The reference numbers must be finalized and the reference list fully formatted before submission.

Examples of the BioMed Central reference style are shown below. Please ensure that the reference style is followed precisely.

See our editorial policies for author guidance on good citation practice.

Web links and URLs: All web links and URLs, including links to the authors' own websites, should be given a reference number and included in the reference list rather than within the text of the manuscript. They should be provided in full, including both the title of the site and the URL, as well as the date the site was accessed, in the following format: The Mouse Tumor Biology Database. http://tumor.informatics.jax.org/mtbwi/index.do. Accessed 20 May 2013. If an author or group of authors can clearly be associated with a web link (e.g. for blogs) they should be included in the reference.

Example reference style:

Article within a journal

Smith JJ. The world of science. Am J Sci. 1999;36:234-5.

Article within a journal (no page numbers)

Rohrmann S, Overvad K, Bueno-de-Mesquita HB, Jakobsen MU, Egeberg R, Tjønneland A, et al. Meat consumption and mortality - results from the European Prospective Investigation into Cancer and Nutrition. BMC Med. 2013;11:63.

Article within a journal by DOI

Slifka MK, Whitton JL. Clinical implications of dysregulated cytokine production. Dig J Mol Med. 2000; doi:10.1007/s801090000086.

Article within a journal supplement

Frumin AM, Nussbaum J, Esposito M. Functional asplenia: demonstration of splenic activity by bone marrow scan. Blood 1979;59 Suppl 1:26-32.

Book chapter, or an article within a book

Wyllie AH, Kerr JFR, Currie AR. Cell death: the significance of apoptosis. In: Bourne GH, Danielli JF, Jeon KW, editors. International review of cytology. London: Academic; 1980. p. 251-306.

OnlineFirst chapter in a series (without a volume designation but with a DOI)

Saito Y, Hyuga H. Rate equation approaches to amplification of enantiomeric excess and chiral symmetry breaking. Top Curr Chem. 2007. doi:10.1007/128_2006_108.

Complete book, authored

Blenkinsopp A, Paxton P. Symptoms in the pharmacy: a guide to the management of common illness. 3rd ed. Oxford: Blackwell Science; 1998.

Online document

Doe J. Title of subordinate document. In: The dictionary of substances and their effects. Royal Society of Chemistry. 1999. http://www.rsc.org/dose/title of subordinate document. Accessed 15 Jan 1999.

Online database

Healthwise Knowledgebase. US Pharmacopeia, Rockville. 1998. http://www.healthwise.org. Accessed 21 Sept 1998.

Supplementary material/private homepage

Doe J. Title of supplementary material. 2000. http://www.privatehomepage.com. Accessed 22 Feb 2000.

University site

Doe, J: Title of preprint. http://www.uni-heidelberg.de/mydata.html (1999). Accessed 25 Dec 1999.

FTP site

Doe, J: Trivial HTTP, RFC2169. ftp://ftp.isi.edu/in-notes/rfc2169.txt (1999). Accessed 12 Nov 1999.

Organization site

ISSN International Centre: The ISSN register. http://www.issn.org (2006). Accessed 20 Feb 2007.

Dataset with persistent identifier

Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, et al. Genome data from sweet and grain sorghum (Sorghum bicolor). GigaScience Database. 2011. http://dx.doi.org/10.5524/100012.

APÊNDICE A - LAITOR4HPC: A TEXT MINING PIPELINE BASED ON HPC FOR BUILDING INTERACTION NETWORKS

Piereck *et al. BMC Bioinformatics* (2020) 21:365
https://doi.org/10.1186/s12859-020-03620-4

BMC Bioinformatics

SOFTWARE Open Access

LAITOR4HPC: A text mining pipeline based on HPC for building interaction networks

Bruna Piereck[1], Marx Oliveira-Lima[1], Ana Maria Benko-Iseppon[1*], Sarah Diehl[2], Reinhard Schneider[2], Ana Christina Brasileiro-Vidal[1] and Adriano Barbosa-Silva[2,3*]

* Correspondence: ana.iseppon@gmail.com; adriano.barbosa@qmul.ac.uk
[1]Genetics Department, Laboratório de Genética e Biologia Vegetal, Universidade Federal de Pernambuco, Recife, Pernambuco, Brazil
[2]University of Luxembourg, Luxembourg Centre for Systems Biomedicine, Bioinformatics Core, Esch-sur-Alzette, Luxembourg
Full list of author information is available at the end of the article

Abstract

Background: The amount of published full-text articles has increased dramatically. Text mining tools configure an essential approach to building biological networks, updating databases and providing annotation for new pathways. PESCADOR is an online web server based on LAITOR and NLProt text mining tools, which retrieves protein-protein co-occurrences in a tabular-based format, adding a network schema. Here we present an HPC-oriented version of PESCADOR's native text mining tool, renamed to LAITOR4HPC, aiming to access an unlimited abstract amount in a short time to enrich available networks, build new ones and possibly highlight whether fields of research have been exhaustively studied.

Results: By taking advantage of parallel computing HPC infrastructure, the full collection of MEDLINE abstracts available until June 2017 was analyzed in a shorter period (6 days) when compared to the original online implementation (with an estimated 2 years to run the same data). Additionally, three case studies were presented to illustrate LAITOR4HPC usage possibilities. The first case study targeted soybean and was used to retrieve an overview of published co-occurrences in a single organism, retrieving 15,788 proteins in 7894 co-occurrences. In the second case study, a target gene family was searched in many organisms, by analyzing 15 species under biotic stress. Most co-occurrences regarded *Arabidopsis thaliana* and *Zea mays*. The third case study concerned the construction and enrichment of an available pathway. Choosing *A. thaliana* for further analysis, the defensin pathway was enriched, showing additional signaling and regulation molecules, and how they respond to each other in the modulation of this complex plant defense response.

Conclusions: LAITOR4HPC can be used for an efficient text mining based construction of biological networks derived from big data sources, such as MEDLINE abstracts. Time consumption and data input limitations will depend on the available resources at the HPC facility. LAITOR4HPC enables enough flexibility for different approaches and data amounts targeted to an organism, a subject, or a specific pathway. Additionally, it can deliver comprehensive results where interactions are classified into four types, according to their reliability.

Keywords: Bioinformatics, PHP, Text mining, Soybean, *Arabidopsis thaliana*, Systems Biology

Background

In the past, scientific information used to be shared via letters between peers. This evolved to printed journals and magazines, and, during the early days of computation, diskettes became a popular way to exchange articles before the advent of the World Wide Web. Today, in the digital era, information has become more accessible, but it has also generated a new venture [1]. Likewise, keeping updated with the "state-of-the-art" and relating all the information available on most fields of study, if not all of them, have turned into an emerging challenge since the 21st century information boom in scientific publishing. According to NCBI resource coordinators (2018) [2], the number of full-text articles have been increasing at a rate of 11.35% a year!

To understand biology in all its complexity, it is necessary to comprehend the structure and dynamics of organisms from cellular to organismal levels. Thus, the focus must change from one element (e.g., protein, gene, phenotype) to a multidimensional point of view. Systems Biology approaches aim to access multi-OMICS data in a variety of experimental conditions and time series to exhaustively generate networks, which may offer an organism's response pathways overview under different situations [3, 4].

Research outcomes and relevant Systems Biology studies data are mostly reported in scientific journals [5]. The need for a more efficient way to explore the plethora of information buried in the various literature silos, has motivated the application of information retrieval and extraction techniques in biology. The area of text or literature mining has emerged, and it is expanding to fill the gap between published and useful information from scientific journals. Given the increase in articles' availability and heterogeneity, text mining tools can boost the construction of new networks using pre-existing information, not to mention revealing insufficiently studied interactions of interest [6, 7]. Text mining can identify and extract biological entities co-occurrences in different levels, such as cellular, tissue and organism-specific contexts, allowing their integration in more informative networks [5, 8, 9].

Text mining tools follow three fundamental processes described by Krallinger and Valencia [10]: (i) information retrieval (finding relevant literature to be analyzed), (ii) biological entities (bioentities) identification (e.g., protein, gene, taxon tagging) and (iii) biological interaction terms to relate/associate the tagged entities. PESCADOR [9] is a web server based on LAITOR [8] and NLProt [11] text mining tools. It uses a list of articles identifiers (PubMed IDs – PMIDS) as a query to search and retrieve relevant abstracts. Furthermore, PESCADOR tags bioentities or biointeractions terms mentioned in the text collection (*corpus*) and identifies biological concepts and their co-occurrences along with bioentities. These co-occurrences are classified into four types according to their reliability, ranging from 1 (more likely to correspond to effective interactions) to 4 (less likely to correspond to effective interactions). Consequently, to build reliable pathways, manual curation is advised [8]. Type classification criteria are: (1) bioentity names co-occur in the same sentence with biointeraction term(s) between them; (2) bioentity names in the same sentence with biointeraction term in any position; (3) bioentity names in the same sentence, permissive identification of biointeraction terms; and (4) all biological entities of the abstract are retained (co-occurrence in the same sentence is not mandatory). Thus, co-occurrences of biological concepts are taken into consideration and reported for co-occurrences of types 1–4. Due to their complexity, the recognition of bioentities is usually the most time-consuming step. Consequently, making use of text mining approaches in big data has been a hard task.

Here, we propose a parallel, fast and unlimited text mining approach by adding customized programming functions suitable for HPC (high-performance computing). Text mining tools have been a valuable approach to support systems biology, not only for updating databases, but also for providing *ab initio* annotation of new pathways, by using automated text processing [12, 13]. To our knowledge, only STRING [14–17] has a programmatic version, but with a different approach than the one proposed here. STRING looks for co-occurrences based on a protein query with two text mining steps added after the update. On the other hand, LAITOR4HPC enables access to all entities ever described in a given species, as well as flexible keyword searches by naming a condition, a subject, or a specific protein, among other possibilities.

Three approaches were addressed to exemplify LAITOR4HPC's use cases. Firstly, all available abstracts of a selected species were analyzed, generating a report containing the absolute number of proteins, co-occurrences and interaction terms. The result provided enough data for stratification from most to least studied subjects. Secondly, considering a subject associated with biotic stress, 15 plants were analyzed to access the different levels of knowledge on a specific field, an approach that may enrich and generate pathways. Finally, a conceptual plant defensin (PDF) pathway is presented for *Arabidopsis thaliana*. PDFs are cysteine-rich, structurally conserved antimicrobial peptides, responsive to biotic stress, including bacteria [18], fungi [19] and insects [20, 21]. Besides the fact that PDF has previously been studied in many plants [21], most of the information about its regulation is scattered in the literature. Here we show the potential of LAITOR4HPC to gather comprehensive information on biological co-occurrences, allowing a conceptual and dynamic view of pathways.

HPC parallelization and execution

All analyses were performed on the Gaia Cluster at Luxembourg University, High-Performance Computing Department. System configuration and cluster organization can be accessed online [22].

In the LAITOR4HPC version, abstracts must be provided as NCBI-PubMed XML format, and the files can be downloaded from PubMed server by doing a search using keywords, or accessed on MEDLINE FTP servers. A Python 2.7 script was written to parse the XML tree structure to recover the PMID, title and abstract of each record. The referred script was already updated to Python 3.0. The script provides an output, which is used as NLProt input [11], and the NLProt output is then used as LAITOR4HPC input.

To run the Python parser, we used the interactive (head) node, which is composed of Bull B500, 2 * Intel Xeon L5640 @ 2,26 GHz, 12 cores and 2880 Gb of RAM. Meanwhile, the following steps were run under the request of running nodes as described online. The NLProt step (bioentity tagging) analysis was launched as four distinct jobs, with 15 cores each (60 cores in total) and LAITOR4HPC was run as a single job, using a total of 20 cores.

Parallelization

GNU Parallel software [23] was used to parallelize the analyses, with the flag “-j N”, where N represents the number of cores to be used, and each core is running the *i-th* input file at a time. To this end, a file containing the list of paths for all the input files

was generated and shared across the cores to be used. Nevertheless, any tool with a similar function must work for the purpose of the LAITOR4HPC tool.

Implementation

The time analyses and sources described concerned the first case study, since it was the most computationally intensive and time-consuming job. To run the time analyses, we have used all papers available until June 2017 from our selected corpus (i.e., MEDL INE), retrieved as previously described. A list with all PMIDs from our corpus is available in Supplementary Material 1. Then, the XML files for the corpus were parsed, and the parsed output was used as NLProt software input to highlight all bioentities (i.e., genes, proteins, taxon names, tissues and cell types). We used NLProt 1.0.2, made available by Rostlab [24]. The final step was to run LAITOR4HPC. LAITOR was initially developed using PHP [25] and its database was designed using MySQL database management system [8, 26]. LAITOR4HPC implementation is intended to be a standalone application, differently from LAITOR version which is integrated to PESCADOR, since jobs originated from web servers usually are executed in a dedicated (or virtual) machine, rather than in an HPC environment. Nevertheless, some of the newly implemented features can also run in a single core, such as the in-memory database query and the name tagging recovery.

A new optional step is to run the summary generator script. This script was written in Python3, with two running modes: (I) Basic: Generates N + 1 summary file, where N represents each LAITOR output in one folder and the extra file with an overall summary of all concatenated data. This is useful when a taxon is analyzed with different keywords searches, or when a big dataset is split for faster running; (II) Spread: Can check several folders to join all results of the basic summaries in a unique summary report. This is useful in cases where the same dataset is analyzed against many species.

All summaries inform how many proteins, co-occurrences and terms were targeted in the analyses. It is important to mention that the basic summary returns a text file for each LAITOR4HPC output, with the extension ".summary". This file contains an 'extra section' describing all terms and how many times they were related to a given co-occurrence. Additionally, if only one file is available, the basic summary step will retrieve two very similar files. The additional file (+1) of the basic summary is named "A.join.dataset.summary" and does not contain the 'extra section'.

To distribute LAITOR as a parallel process, it was necessary to make sure that the processes running on different nodes could query the bioentities and biointeraction dictionaries seamlessly. However, MySQL requires its installation in every node for it to be used, which is possible, but against the user practices in most HPC systems, including ours. Therefore, we chose to switch the original disk-stored LAITOR databases (MySQL) by an in-memory database system. For that purpose, we used SQLite (version 3.0): a self-contained, highly reliable, embedded, full-featured, public-domain SQL database engine [27]. Consequently, we needed to adapt the queries from the former system to the latter (Fig. 1).

Three case studies were performed, aiming to encompass the different LAITOR4HPC applications. The first case study aimed to search all bioentities co-occurrences in all available abstracts for a given species (*Glycine max*); the second, to use keywords to

```
###OLD MySQL QUERIES
(A) #Connect to database
    $conn=mysql_connect($server,$user,$pass);
    $sele=mysql_select_db($db);

(B) #Preparing and executing query
    $query=mysql_query("select name_txt from" $table_genes.
                       "where tax_id=\"".$tax_id."\"");
    if(mysql_num_rown($query)>0){

(C) #Fetching the results
         $result=mysql_fetch_array($query);
         return($result['name_txt']);
    }
    else{
         return(FALSE);
    }

###NEW SQLite QUERIES
(D) #Creating in-memory SQL database
    $pdo=new PDO(sqlite::memory);
    $pdo->setAttribute(PDO::ATTR_ERRMODE, PDO::ERRMODE_EXCEPTION);
    $pdo->exec('ATTACH "./laitor_nocase.db" as laitor_db');

(E) #Preparing the SQL query
    $sql= $pdo->prepare("select * from" .$table_gene.
                        "where tax_id=\"".$tax_id."\"");
    $sql->setFetchMode(PDO:FETCH_ASSOC);

(F) #Executing the SQL query
    $sql->execute();


(G) #Fetching the results
    $result = $sql->fechAll();
```

Fig. 1 LAITOR4HPC database management system updates. The principles are the same as the previously online version. However, MySQL connects to a server where the database is stored in the disk a, whereas SQLite loads the database file in the RAM of the node executing the query d. The remaining processes are similar when using both technologies: b, e, f preparing and executing the query; and c, g retrieving the results

look for all described interactions on one subject (biotic stress) in 15 different plant species; and the third, to build a pathway based on the information retrieved by the keywords "Plant AND Defensin" in *A. thaliana*.

For the first case study, the taxonomy identifier (tax-ID) filter option of LAITOR4HPC was used to check all soybean (*Glycine max* – Taxonomy ID: 3847) interactions described in 1134 XML files (approximately 30,000 abstracts each), comprising every MEDLINE paper available until June 2017. In this case study, no restriction on the subject was made. Therefore, all possible co-occurrences ever published about soybean could be retrieved. The basic summary report was used to access the 10 most studied proteins, co-occurrences and related terms.

In a second approach, a collection comprising a set of biotic stress-related keywords retrieved from MEDLINE (NCBI) was submitted to the pipeline 15 times, one for each plant species. This case study aimed to uncover interactions that are being over studied and some that are probably being ignored for some species. The basic summary report has evidenced the most studied proteins, co-occurrences and related terms, as well as

the neglected ones for each species. Additionally, the 'spread report mode' has allowed the context analysis of each co-occurrence, independent of species specification.

For the third case study, the defensin-associated pathway regulation was built for *A. thaliana*. The parsed XML file related to the keywords "Plant AND Defensin" was selected. Furthermore, only interactions tagged as type 1 (proteins in the same sentence with the biointeraction term between them) were chosen to be used on CellDesigner [28] for construction of the pathway model. All the retrieved abstracts related to this step were manually curated, to verify possible false positives and interactions that may have not been tagged, thus allowing the expansion of the pathway beyond the automatic annotation (those interactions retrieved exclusively by the pipeline).

CellDesigner was used to make a conceptual visualization of the pathway by connecting the biointeraction terms with tagged proteins and reporting events of activation, regulation and inhibition. For a better visualization, different bioentities (genes, proteins and simple molecules) were represented by different shapes and colors.

Scalability test

The scalability test was performed using three XML files containing 1000 abstracts each, from three different species: *Caenorhabditis elegans*, *Homo sapiens* and *Arabidopsis thaliana* in a computer composed by an Intel Xeon(R) E-2124G CPU @ 3.40 GHz × 4 cores and 32 Gb of RAM following the same pattern proposed here to LAITOR4HPC.

First, we ran the Python parser script for each species using the GNU parallel. Then, the NLProt step was performed in two different stages: parallelized and non-parallelized, to evaluate the running time, per core usage and the number of tagged proteins. In the first case, the analysis was performed by running the three files using three, two and one core, sequentially. In the second run, the files were evaluated separately by using one core, but also a single file containing all the 3000 abstracts.

Finally, the LAITOR4HPC was carried out separately in one core to tag the interactions in each file and the running time, since the necessity of a specific tax-ID precludes the parallelization in this specific study case Table 1.

Results

First case study (soybean) and implementation

The first step, the Python parser script, was run on the head node against the 1134 XML files (approximately 31 M abstracts) in nearly 5 min. The second and third steps, comprised by NLProt and LAITOR, took 6 days in total to analyze all files filtering for soybean tax-ID. This represents an average processing rate of 0.017 s per abstract,

Table 1 Files, cores and their respective steps processed in each stage of the scalability test

No. files	No. Cores	Step
3	3	Parsing
3	3	NLProt
3	2	NLProt
3	1	NLProt
1	1	LAITOR4HPC (3x)

which is a speed-up of approximately 117 times in comparison to the original implementation (in which the NLProt tagging alone took around 2 s to complete) [9]. The running time should vary depending on node configuration and cores available on the HPC, but it is faster than using a single core approach.

Figure 2 represents the general pipeline obtained for the preparation of the MEDL INE abstracts as an input for the LAITOR4HPC text mining process. After downloading the full MEDLINE collection, a dataset of 1134 XML files was obtained, each containing approximately 30,000 PMIDS (Fig. 2a). These files were transferred to the HPC environment via SCP (Secure Copy Protocol) over an SSH (Secure Shell) protocol (Fig. 2b). The Python parser converted these records into readable NLProt MEDLINE input files (Fig. 2c). After that, the NLProt job was launched (Fig. 2d), where four nodes and a total of 60 computing cores were used to run i NLProt processes (where: $\{i \in \mathbb{Z} \mid \{0 < i < 1305\}\}$) to tag the bioentity names within those 1304 input files (Fig. 2e). Upon conclusion, those 1304 NLProt output files were made available in the head node (Fig. 2f), ready for the LAITOR4HPC step.

The LAITOR4HPC job execution uses the DB file and the NLProt output files as inputs (Fig. 2g). The jobs were launched from the head node, to be executed by one node with 20 cores. Each *i-th* process was directed to a corresponding computing core together with the DB file, the LAITOR4HPC script and the NLProt output. Every

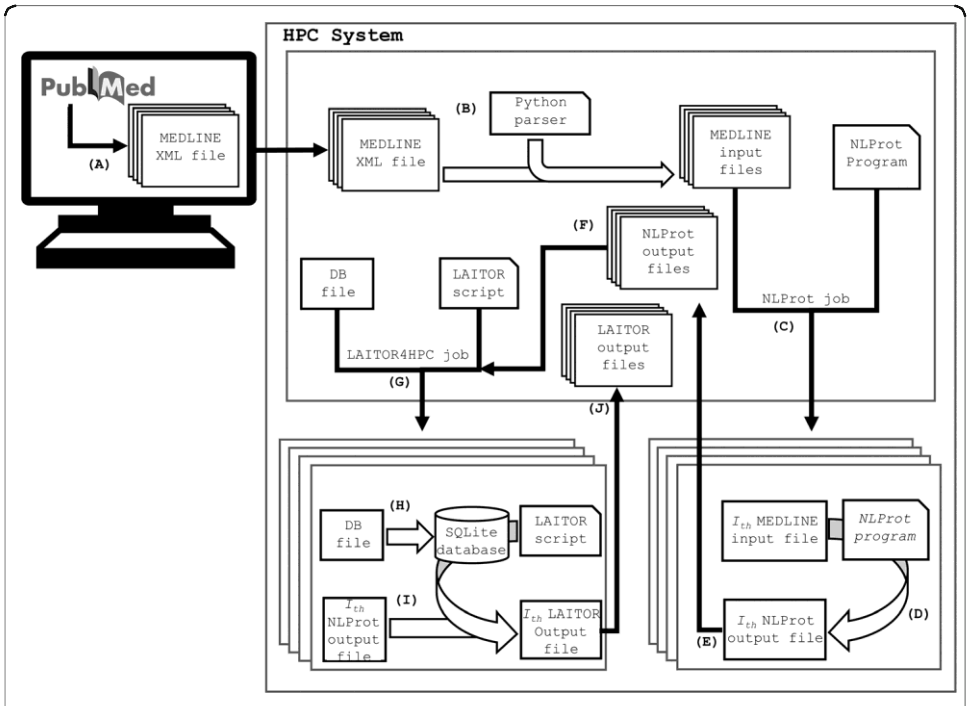


Fig. 2 Complete text mining pipeline using NLProt and LAITOR4HPC. a MEDLINE files are downloaded from NCBI FTP as XML files; b a Python parser is executed to convert the XML files into input files for NLProt which are then c transferred into the interactive (head) node of the HPC system. d A job is then started and *i* different processes are launched in parallel on 60 computing cores (where: $\{i \in Z \mid \{0 < i < 1305\}\}$). e In each core, the corresponding *i-th* MEDLINE input file is tagged by NLProt which generates f an *i-th* NLProt output file, which is then placed back to the head node together with the other outputs. g These files are used together with the DB file as input for the LAITOR4HPC job; h which loads an in-memory database before the i tagging of the bioentities and biointeraction present in the corpus. j After completion, the results are placed back to the head node and made available for downstream applications

computing core loads the DB file as an SQLite in-memory database in that node during execution (Fig. 2h). Then the LAITOR4HPC script receives the *i-th* process and analyzes it against the loaded in-memory database, which contains the bioentity and biointeraction dictionaries (Fig. 2i). Once the results are obtained, they are made available back to the head node (Fig. 2j). At the end of the job, all the LAITOR4HPC output files are retrieved back to the head node and can be copied by SCP or another similar method to a user-client computer; from there, users can further explore the text mining outputs to create co-occurrence networks, for example.

By switching from MySQL to SQLite, we avoid HPC limitations during the database querying in the HPC architecture, as previously mentioned. Using SQLite in-memory, a new database is created purely in the memory of the computing nodes. This database ceases to exist as soon as the database connection is closed. As the database is self-contained in a text file, this file needs to be distributed across the computing cores along with the input file to be analyzed.

LAITOR4HPC running time was drastically decreased by the parallelization approach, which also allowed the user to query the whole corpus and extract all its bioentity co-occurrences. In comparison to the original version used by the PESCADOR website, where only a maximum of 1000 papers could be read on-the-fly, and NLProt alone was lasting 2 s. In LAITOR4HPC context: the more articles available, the better the result. Parallel SQL limitations caused by competitive accesses on the HPC environment were avoided by loading the database in the RAM of each computing node.

After the soybean analysis performed on the overall MEDLINE corpus, the pipeline has tagged 15,788 proteins and 7894 co-occurrences along with the four occurrences types (type 1, 104; type 2, 685; type 3, 2369; type 4, 4736). The 96 non-redundant proteins were responsible for 1254 different co-occurrences in soybean. Rubisco Large subunit (rbcL) was tagged 1767 times and was present in three out of 10 most studied co-occurrences, followed by photosystem II protein A (psbA), which was tagged 983 times and present in four out of 10 of the most observed co-occurrences (Fig. 3), with475 co-occurrences of Maturase K (matK)/rbcL. Not by chance, the most abundantinteraction terms, among the non-redundant 227 terms, were *encoding* (140), *amplified* (60) and *encode* (47) (Fig. 3).

A closer look has revealed that all top 10 co-occurrences in soybean have at least one chloroplast-encoded protein related to photosystem II, and around 55% of mapped co-occurrences have the same pattern, with at least one of those being three chloroplast-encoded proteins (rbcL, psbA, matK) (Supplementary Material 2). The proteins, the co-occurrences and the terms that were only identified once or twice were considered as poorly described. Thus, for soybean, almost half of all the co-occurrences (751) and all the terms (128) were deemed as poorly characterized. All proteins, co-occurrences and terms can be found in Supplementary Material 2.

Using keywords to search all described interactions on one subject

To describe the interactions related to biotic stress in plants, the same subset of papers was used to search for information about 15 species. The chosen keywords related to biotic stress were filtered for plants (keywords are listed in Supplementary Material 3)

(A) *Glycine max* – All MEDLINE corpus

	PROTEINS	CO-OCCURRENCES	TERMS
Non-redundant	96	1254	277
Total	15788	7894	1214

(B) 10 most in Glycine max – All MEDLINE corpus

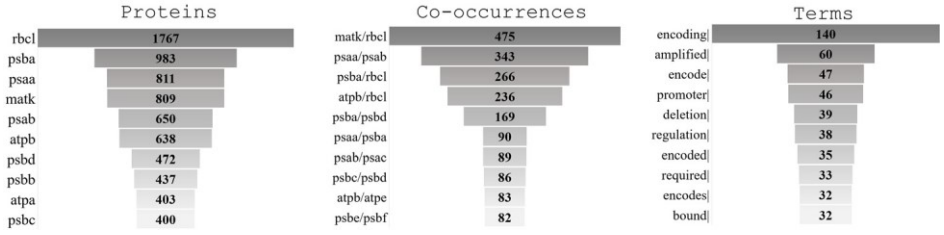


Fig. 3 Mapping of LAITOR4HPC retrieved text analysis of all *Glycine max* available abstracts (until June 2017). a Non-redundant number of identified proteins, co-occurrences, and terms followed by absolute entities number. b Top 10 most described entities in absolute number. The list with all proteins, co-occurrences, and terms is available in Supplementary Material 2

The first two steps (1) Python parsing and (2) NLProt tagging were run only once for all the analyses. The third step, LAITOR, was executed with different tax-IDs to specify the species. No co-occurrences were registered for three out of 15 species (*Medicago truncatula, Nicotiana benthamiana* and *Ricinus communis*).

The number of proteins, co-occurrences and terms varied greatly among the remaining 12 species. Considering all tagged proteins, co-occurrences and terms (Fig. 4), PR-1 (Pathogen-related) protein family was explicitly the most widespread molecular entity (5382 descriptions and 138 unique co-occurrences), followed by NPR-1 (3108 descriptions and 173 non-redundant co-occurrences). PR-1/PR-5 and RPS2/RPS4 (ribosomal protein small subunit) were the most representative, with 1019 and 914 interactions, respectively. The profile of the most annotated proteins suggests that for the biotic stress-related subject, expression of responsive genes is the main focus of study, since among the most retrieved terms are: *required* (552), *induced* (460), *induction* (299), *enhanced* (345) (complete list of proteins, co-occurrences and terms available in Supplementary Material 4).

A. thaliana and *Zea mays* are, by far, the most studied plants (Fig. 5). Using LAITOR tax-ID against the same abstract set to filter both species interactions, a total of 14,411 and 2744 co-occurrences were mapped respectively, considering all four types. Conspicuously, *A. thaliana* registered 1468 non-redundant tagged proteins, comprised of 4224 unique co-occurrences (Fig. 6a), where PR-1 alone accounted for 1470 occurrences, highlighted in four out of 10 most abundant co-occurrences (Fig. 6b). Additionally, PR-1 was present in a total of 115 co-occurrences in the selected corpus. Terms such as *regulation* (135), *induced* (116) and *enhanced* (105) are among the most traced (Fig. 6c).

Considering *Z. mays* interactions, far fewer proteins were tagged (i.e. 365, which only represent one fourth when compared to *A. thaliana*'s amount of tagged proteins). On the other hand, *Z. mays* registered almost half the number of unique co-occurrences (2742). Therefore, a more efficient network link was displayed with PR-1, as well as with the most abundant protein, totaling 486 tagged PR-1 in 32 co-occurrences. Even

(A) All searched plants – Biotic stress keywords dataset

	PROTEINS	CO-OCCURRENCES	TERMS
Non-redundant	1766	4812	813
Total	107574	53805	18807

(B) 10 most in all searched plants – Biotic stress keywords dataset

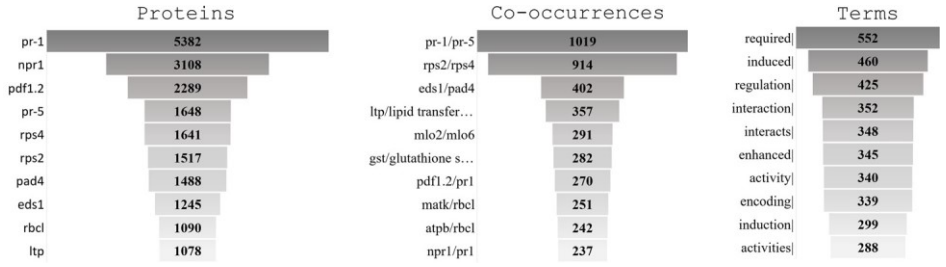


Fig. 4 Mapping of LAITOR4HPC retrieved text analysis for 12 searched plant species among manuscripts subsets. The subsets were selected with 20 biotic stress-related keywords combinations (Supplementary Material 3). a Non-redundant number of identified proteins, co-occurrences, and terms followed by the absolute number of entities. b Top 10 most described entities in absolute number. The list with all proteins, co-occurrences, and terms is available in Supplementary Material 4

though PR-1 appears only once in the 10 most cited, the interaction between PR-1 and PR-5 was registered 370 times (Fig. 6b).

In total, 38 co-occurrences were clustered, considering the 10 most relevant results for each species (Fig. 6b). Nine of those were exclusively described in *A. thaliana* (dcL2/dcL4; EDS1/PAD4; GST/glutathione S-transferase; LTP/lipid transfer protein; NPR1/PR-1; PDF1.2/PR-1; PR-1/PR-2; RIN4/RPM1; RPS4/RRS1), seven were unique in *Z. mays* (A1/A2; ARF1/GTPase; CaM/Ltp; CAT2/GBF1; Hm1/Hm2; MLO2/MLO6; MPK4/MPK6) and five were observed only in *Nicotiana tabacum* (ATP6/ATP9; ATP6/cox3; cox1/cox2; cox1/cox3; NaD1/NaD2). The interaction between matK/rbcL was the only one registered for all 12 species with similar values and, considering it regards a conserved chloroplast function, it was expected to be found in all plants. The RPS2/RPS4 co-occurrence was described for all the angiosperms searched. The pteridophyte *Selaginella moellendorffii* was the only species which did not show any RPS2/RPS4 (Fig. 6b), even after a new online keyword search on the updated 2019 MEDLINE database was performed. Despite being the least studied of all plants in the selected set, *S. moellendorffii* presents three exclusive interactions: chlB/chlL; chlB/chlN; chlL/chlN.

The other eight species (*Manihot esculenta, Cicer arietinum, Lotus japonicus, Phaseolus vulgaris, Glycine max, Brachypodium distachyon, Solanum lycopersicum, Solanum tuberosum*) revealed very similar profiles (Fig. 6). This can be explained by poorly described abstracts, missing information, or it could be due to abstracts citing more than a single species, thus causing ambiguous tagging during the NLProt process. On average, all plants have 18% of poorly studied co-occurrences (with only one or two co-occurrences registered) (Supplementary Material 5). Despite the distinct high amount of studies in *A. thaliana,* 19% of the characterized co-occurrences were poorly studied. On the other hand, only 10% of *Z. mays*' co-occurrences were considered poorly studied, a result following the inference of efficient network construction.

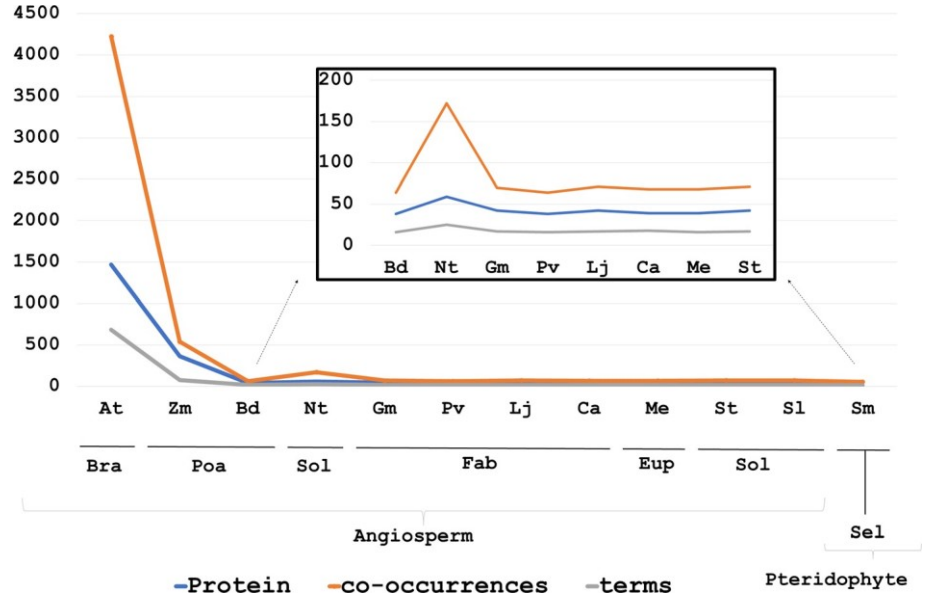


Fig. 5 Variety of proteins, co-occurrences, and terms per species. At Y axis is possible to see the amount of each entity (proteins, co-occurrences, terms) and at X axis, species are named with the initials: Angiosperm: At - *Arabidopsis thaliana* (Brassicaceae); Zm - *Zea mays* and Bd - *Brachypodium distachyon* (Poaceae); Nt- *Nicotiana tabacum,* St- *Solanum tuberosum* and Sl - *Solanum lycopersicum* (Solanaceae); Gm-*Glycine max*, Pv-*Phaseolus vulgaris,* Lj-*Lotus japonicus*, Ca-*Cicer arietinum*, Me-*Manihot esculenta* (Fabaceae); Pteridophyte: *Selaginella moellendorffii* (Selaginellaceae). The list with all proteins, co-occurrences, and terms for each plant species is available in Supplementary Material 5

Building pathways

The pathway annotation or enrichment is a challenging task in many aspects, mainly because it requires great efforts in the selection, examination and extraction of relevant information in the retrieved literature. This work can be even harder to be enriched or designed, depending on how large the pathways are [9], since a simple pathway can display many complex interactions (Fig. 7).

Considering the whole set retrieved by LAITOR4HPC, we selected 31 abstracts with type 1 only interactions to build the plant defensin pathway regulation in *A. thaliana* (PDF). It is important to highlight that the whole MEDLINE database was used as a training set, to tag the interaction terms and targets (e.g., genes, proteins) more efficiently. Thus, our pipeline was able to find feasible connections in 24 manuscripts that served to both automatic and manual annotation (Supplementary Material 6). To our knowledge, this is the first attempt at gathering information on a PDF network, focusing on building a pathway and specifying the relations among the entities. However, it must be mentioned that the gene encoding PDF has been tagged on MAPK signaling pathway at KEGG database (Entry ko04016). A correlation with proteins has also been reported on STRING, as, for instance, Octadecanoid-Responsive Arabidopsis (ORA59) and NPR1, both transcriptional activators [29, 30], which have as well been included in the present pathway.

The modeled pathway (Fig. 7, SMBL file available in Supplementary Material 7) indicates some well studied *A. thaliana* genes related to defense transcription factors. All tagged proteins and genes (except for PDF) either belong to TF class or are signaling regulators, like *coi1* and *pepr1* [31, 32]. A general overview of the defensin regulation

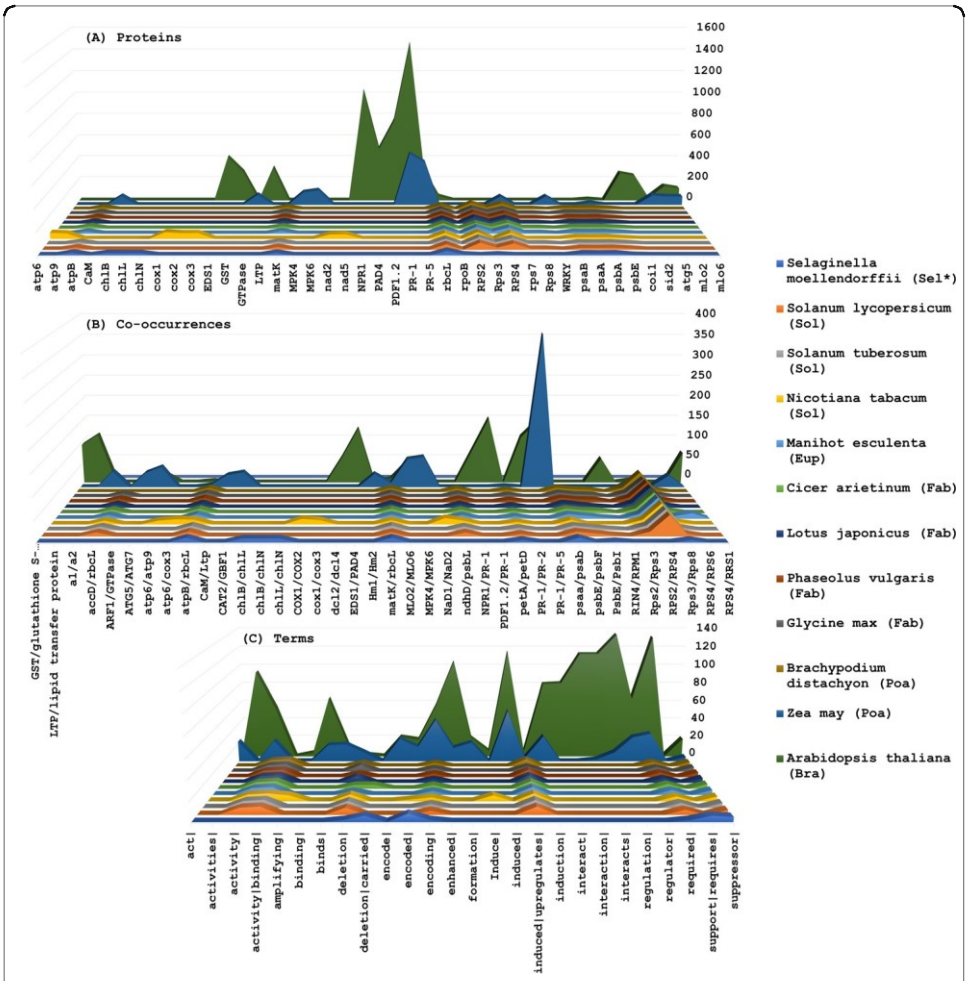


Fig. 6 Comparison of 10 most identified proteins, co-occurrences, and terms for 12 plant species. Those species were analyzed regarding the biotic stress subset of abstracts. At Y axis the amount of each entity (proteins, co-occurrences, terms), at Z species are specifies by colors and at X-axis the entities for each item. a A total of 41 proteins are displayed with respective amounts. b A total of 38 co-occurrences can be observed with the respective number of occurrences. c A total of 26 interaction terms are shown with their respective quantities. Pteridophyte: *Selaginella moellendorffii* (Selaginellaceae). Angiosperm: *Solanum lycopersicum*, *Solanum tuberosum* and *Nicotiana tabacum* (Solanaceae); *Manihot esculenta* (Euphorbiaceae); *Cicer arietinum*, *Lotus japonicus*, *Phaseolus vulgaris* and *Glycine max* (Fabaceae); *Brachypodium distachyon* and *Zea mays* (Poaceae); *Arabidopsis thaliana* (Brassicaceae)

pathway in *A. thaliana* allows the division of its whole structure into three main groups: signaling, regulation factors and defense response itself. For the signaling group, three hormones play a role as positive effectors: nitric oxide (NO), jasmonic acid (JA) and ethylene (ET) [33]. The second group (regulation factors) regards the transcription factors and receptors, and the third group (defense response) regards the PDF genes (Fig. 7).

Scalability

In the NLProt parallelized analysis, the run-time varied from 12 min (three cores and three files) to 29 min (one core and three files), tagging 71,969 proteins considering all results. For the non-parallelized analysis, the run-time varied from 10 min to 29 min for

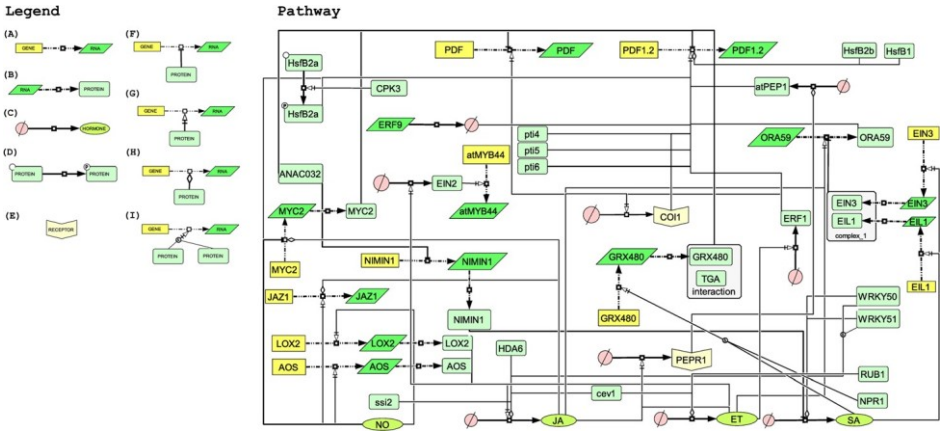


Fig. 7 *Arabidopsis thaliana*'s defensin (PDF) pathway. On the left a representation of all interaction types recovered in the present case study: a Transcription; b Translation; c State transition; d Phosphorylation; e Receptor; f Inhibition; g Activation; h Modulation; i Boolean sign "&" indicating that both proteins activate a transcription. The PDF gene pathway represents the co-occurrences recovered by LAITOR4HPC when using the terms "Plant AND Defensin". The presented data regards the summary of 31 curated abstracts containing type 1 interactions. Yellow rectangles represent genes; Bright green parallelograms represent RNA transcripts; Light green rectangles are proteins; Light yellow arrow-like figures represent a receptor; Light pink circles are inactive hormone forms; Green oval-shaped forms regard the active form of each hormone. Abbreviations: PDF: Plant Defensin; NO: nitric oxide; JA: jasmonic acid; ET: ethylene; SA: salicylic acid

the largest file. However, on the file containing all the 3000 abstracts, the number of tagged proteins was 84,284, due to the SVM optimization performed by NLProt (Fig. 8).

For the LAITOR4HPC the number of interactions/running time varied from approximately 4000 interactions in 159 min for *H. sapiens* to approximately 6000 interactions in 132 min for *C. elegans*. The fact that the worm had more proteins and interactions validated in less time is due to the number of proteins and redundancies retrieved in other organisms in its abstracts set (Fig. 9).

Discussion

The first case study identified 96 non-redundant proteins responsible for 1254 different co-occurrences in soybean, in which chloroplast-encoded protein are abundant. Besides their functional importance, chloroplast-encoded proteins are, together, widely used as a barcode for species and population studies in Fabaceae [34, 35]. As a consequence, it is not by chance that the most abundant interaction terms were: *encoding*, *amplified* and *encode*. Additionally, all top 10 co-occurrences in soybean are somehow related to photosystem II, whose proteins are part of a thylakoid structure and can be affected by high salt levels. Plants that can avoid a decrease in such proteins during stress may tolerate the stress with higher success [36]. Photosystem II proteins' efficiency can also limit biomass [37]. For that reason, improving these proteins on plants of agronomical importance, such as soybean, is of great interest.

Fifteen species were selected, however, three had no co-occurrences described, probably due to one of the two reasons: either (1) there is no description available for the searched protein interactions, or (2) the protein interactions are not well described in the paper's abstract. Such a flaw could lead to a false-negative result, since the main

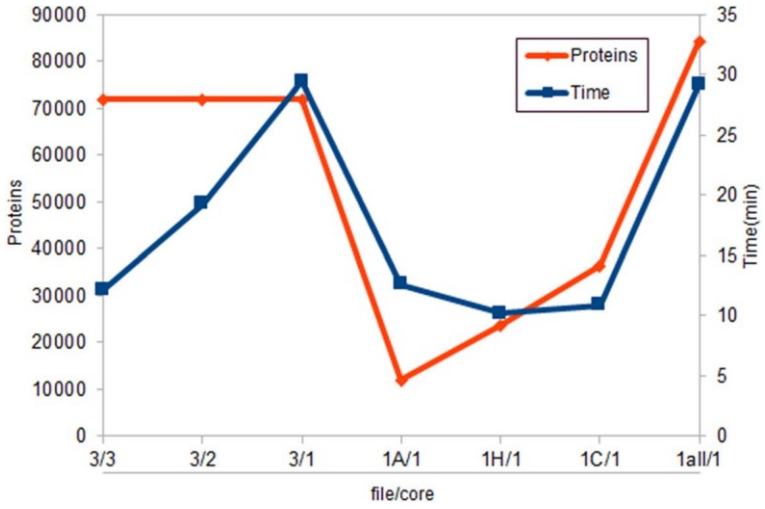


Fig. 8 Relation among number of proteins, running time and file used to run the NlProt. 1A - *Arabidopsis thaliana;* 1H - *Homo sapiens;* 1C - *Caenorhabditis elegans;* 1all - all the abstracts gathered in one file

text mining tools, such as LAITOR4HPC, PESCADOR [9], STRING [14–16], iHop [38], only access the abstracts. The preference for accessing abstracts is due to the difficulties in parsing full paper texts, which can include images and tables. The more objective the abstract is, including the key information, the more efficient the text mining tool will be in retrieving the results.

Considering the remaining 12 plants from the original set, PR-1 and NPR-1 proteins are the most described bioentites. The former was identified in the 1970s, and it still is largely studied and vastly induced during plant defense response. Since there is also evidence of PR-1 activity in growth and development besides stress response, its full biological role has not been completely clarified. As a consequence, the number of studies on this behalf keeps growing [39]. On the other hand, NPR-1 is a transcriptional regulator of plant stress response that is regulated by stress-released hormones recognized by plant receptors [40]. The two most representative interactions were PR-1/PR-5 and RPS2/RPS4, which contain the PR and RPS plant-disease resistance (*R*) genes, with a specific bacterial-resistant response [41, 42]. Together with the most annotated terms (*required*, *induced*, *induction*, *enhanced*), this suggests that expression profile is one of the most studied topics for plant biotic stress response.

First suggested as a model plant in 1943, *A. thaliana* has been studied for approximately 70 years, cited in more than 54,000 manuscripts until 2016 and considered a benchmark on the understanding of plant-pathogen responses, helping to enlight higher plants research. Nevertheless, this model is still an important source to fully understand stress response in flowering plants, considered an entry point for elucidating or identifying still uncovered plant-protein interaction [43–46]. Thus, it is not a surprise that this small plant stood out as the most researched plant and that the not yet completely clarified PR-1 protein is being exhaustively studied in the model plant as well.

Despite the abundance of repetitive sequences and complex genome, *Z. mays* was the second in the number of available data, exhibiting fewer tagged proteins, but almost half of the unique co-occurrences when compared with *A. thaliana*; therefore, displaying a more efficient network link. *Z. mays* is one of the primary sources for food

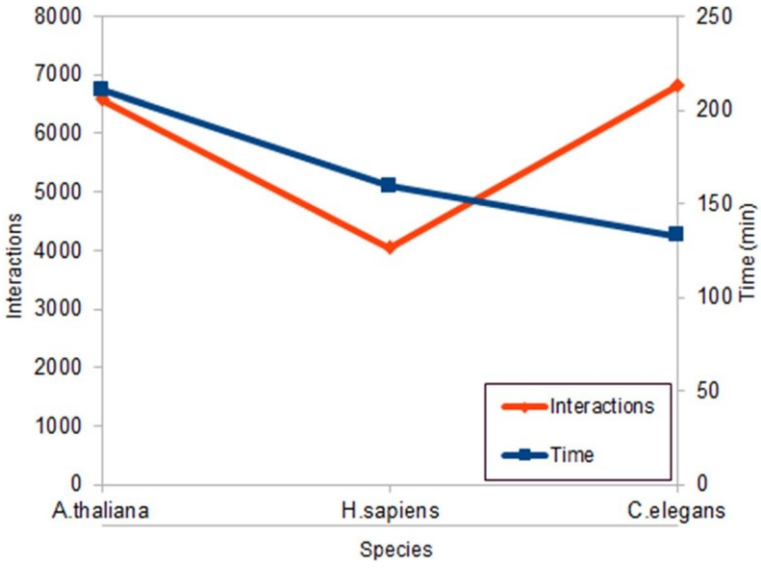


Fig. 9 Relation among the identified interactions, running time and the species for each set of abstracts

security and one of the most studied plants when it comes to breeding studies aiming to boost productivity, seed protein quality and, especially, to raise resistance to pathogens [46–48]. Therefore, it is not a coincidence that, as in *A. thaliana*, PR-1 is also the most abundant protein in *Z. mays*.

In general, plant hormones are involved in a wide range of defense-related signaling pathways [44]. In the presented case study, four manually curated hormones (NO, JA, ET and salicylic acid; SA) interact regulating defense response. The JA hormone works like a positive effector, by activating regulation factors as COI1 and PEPR1, which is also activated by ET. In turn, PEPR1 modulates at PEP1 (as COI1), involved in *PDF1.2* transcription induction [31, 45, 46]. Additionally, JA controls defensin expression by inducing specific transcription factors, as ORA59 [29], or by modulating the transcription of *MYC2* that inhibits PDF [47].

Another signaling molecule that plays an essential role in the pathway is NO, first because it inhibits *MYC2* transcription and, second, because it induces *PDF* expression. Besides, NO also activates the JA signaling positive effectors *LOX2* and *AOS* [47]. Thus, it plays a role in the pathway, not only by enabling transcription factors to induce the defense response, but also by regulating JA signaling intensity. Finally, ET signaling hormone can induce *PDF* transcription indirectly by activating ORA59 and ERF1 [29, 46] (Fig. 7). Both ORA59 and ERF1 are positive effectors to *PDF1.2* transcription, despite activating PEPR1 receptor, an indirect positive regulator of defensin transcription, as aforementioned [31].

The only hormone mapped as a negative regulator of *A. thaliana* defensin expression was SA, by inducing the *EIN3* and *EIL1* transcription, which become a complex EIN3/EIL1. This complex is responsible for inhibiting the positive regulator ORA59 [29] as mentioned before. Additionally, SA hormone and NPR1 protein activate *GRX480* transcription. Once active, it forms an interaction complex with TGA and inhibits *PDF1.2* transcription [48] (Fig. 7). These results show how complex the plant defense regulation can be and shed some light to understand the cross factors that may occur. Thus,

in this case study, the pipeline was very effective, not only in retrieving information automatically, but also in providing a significant and pertinent abstract set for manual annotation. Such a combination of approaches allowed specifying the correlation among the entities in an efficient way, giving a more detailed view of the defensin regulation pathway in *A. thaliana*.

Most of the current text mining tools are either online, like PESCADOR [9], STRING [14–16] and PPIcurator [49]; or very specific, such as FamPlex [50], for human proteins, MPTM [51], for post-translation modification in humans, and PaperBLAST [52], for homology search. LAITOR4HPC and STRING updates are the only programmatic text mining tools available that came to our knowledge. Nevertheless, both have different approaches. STRING focuses on co-occurrences within neighborhood genes and uses protein names for keyword searches. Text mining functionality is directed to corroborate interactions in their database and, when used separately, it retrieves only the top tagged proteins [14–16]. On the other hand, LAITOR4HPC pipeline is intended to retrieve information from a different point of view, thus providing flexibility in research topics. Our pipeline is prepared to search all the interactions of a given PubMed XML corpus, retrieving data for a comprehensive network design. Besides, LAITOR4HPC can help spotting co-occurrences that have already been exhaustively studied, as well as highlight some that have been poorly studied or that still have not been considered.

The parallelization has sped up the analyses, since it avoids the pilling up of files. The time rate comparison revealed a speed improvement of more than double from the parallelized version to the non-parallelized one.

Conclusion

The improvement of LAITOR [8] and development of LAITOR4HPC has decreased computing time significantly, due to the implementation of parallelization. Such an increase resulted not only in much faster run time, but also maintained the consistency and reliability of previously LAITOR implementations. Time will vary accordingly, depending on available hardware resources, specially regarding memory capacity and the number of available cores. Since this improved online tool includes only data from abstracts, it is essential to consider manual data curation to confirm predicted protein-protein interactions from co-occurrences terms.

Despite its economic importance and intensive research investments, most soybean publications are focused on chloroplast-encoded proteins, rather than on stress-responsive proteins. On the other hand, in the case study that analyzed the biotic stress terms, PR-1 was the most representative protein, and probably some effort should be applied to clarify other genes/proteins related to the biotic stress response. A more comprehensive subset of described interactions can fill gaps in the understanding of PR-1 role and in other relevant pathways related to the biotic stress response. Using manual and automatic annotation, the pipeline provided a very detailed pathway with literature support, evidencing the components of plant defensin signalling and modula- tion. Thus, it maintains the accuracy of PESCADOR with the improved possibility to analyze big data in a short time.

LAITOR4HPC is suitable for establishing or enriching new interaction pathways. It has shown to be efficient in retrieving reliable information, providing an

overview for a given target, or even for a given keyword associated with an organism of interest. It is important to highlight that the pipeline was able to retrieve all the relevant sets of papers for the searched topics in a more efficient way than just digging into the list of MEDLINE publications. Since the number of manuscripts is increasing quickly, new approaches for linking information are demanded to enable a fast, reliable and prompt way of fully understanding the targeted taxon's or organism's systems biology.

As take-home message, for more efficient development and application of tools, such as LAITOR4HPC in Systems Biology, future publications should include some 'minimal information about publication of interaction data' (MIAPID) preferably in a tabular format. This summary of identified and validated interactions will simplify the data recovery and integration to generate or enrich existing pathways.

Datasets availability

Project name:LAITOR4HPC

Project home page: https://zenodo.org/record/1717329

Operating system(s): e.g., Linux (Ubuntu 16.04+) and macOS 10+

Programming language: Python v.3, PHP v.7

Other requirements: Perl v.5.22, SQLite v.3

License: GNU Affero General Public License (AGPL 3.0)

Any restrictions to use by non-academics: No restriction

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03620-4.

Additional file 1. PMID list. Unformatted text file containing the list of PMIDs used to the soybean analysis. The list represents all the MEDLINE PMIDs available until July 2017.

Additional file 2. Searching bioentities co-occurrences in all available abstracts for one specie (First case study report). Excel file containing information about the total of proteins, co-occurrences and terms of interaction mapped by LAITOR4HPC in *Glicine max*. For each category there are information about the absolute number of each element category, the total number of elements and the number of non-redundant elements. Additionally, one section is dedicated to the total for each interaction type, INT_1 (more likely to be effective) to INT_4 (less likely to be effective).

Additional file 3. Keywords. PDF file listing the 20 keywords used to filter PMID related to biotic stress on case study 3. Each keyword was run on LAITOR4HPC 15 times, one time for each plant species tax-ID (*Arabidopsis thaliana*, *Zea mays*, *Brachypodium distachyon, Nicotiana tabacum, Solanum tuberosum, Solanum lycopersicum,* Gm-*Glycine max, Phaseolus vulgaris, Lotus japonicus*, *Cicer arietinum*, *Manihot esculenta, Selaginella moellendorffii, Medicago truncatula, Nicotiana benthamiana,* and *Ricinus communis*).

Additional file 4. Using keywords to look for all described interactions on one subject, all plants summary (Second case study report). Excel file containing information about the total of proteins, co-occurrences and terms of interaction mapped by LAITOR4HPC in 12 plant species considering the biotic stress related PMIDs. For each category there are information about the absolute number of each element category, the total number of elements and the number of non-redundant elements. Additionally, one section is dedicated to the total for each interaction type, INT_1 (more likely to be effective) to INT_4 (less likely to be effective).

Additional file 5. Using keywords to look for all described interactions in *Arabidopsis thaliana* (Third case study report). Excel file containing information about the total of proteins, co-occurrences and terms of interaction mapped by LAITOR4HPC in *Arabidopsis thaliana*. For each category, there are information about the absolute number of each element category, the total number of elements and the number of non-redundant elements. Additionally, one section is dedicated to the total for each interaction type, INT_1 (more likely to be effective) to INT_4 (less likely to be effective).

Additional file 6. Building pathways (Third case study report). Excel file containing a curated list of interaction identified with LAITOR4HPC in *Arabidopsis thaliana* for the "plant AND defensing" keywords filtered PMIDs. This list embraces automatic and manual annotation of the mapped data.

Additional file 7. *Arabidopsis thaliana pathway.* The SMBL file is a xml-based file format storing the computational biological model of the pathway. The file contains the defensing interaction network of *Arabidopsis thaliana*, and is compatible with CellDesigner tool for biological network visualization and creation.

Abbreviations

Symbol: Abbreviation meaning; A1: Dihydroflavonol 4-reductase; A2: Leucoanthocyanidin dioxygenase; AOS: Allene oxide synthase; ARF: Auxin Response Factors; ATP: Adenosine Triphosphate; CaM: Calmodulin; CAT: Catalase; chl: Chloroplast; coi: Coronatine Insensitive; cox: Cytochrome oxidase; DB: Database; dcL: Dicer-like; EDS: Enhanced disease susceptibility; EIL: Ethylene Insensitive-Like; EIN: Ethylene Insensitive; ERF: Ethylene Response Factor; ET: Ethylene; GBF: G-Box Factors; GRX: Glutaredoxin; GST: Glutatione s-transferase; Hm1: NADPH HC toxin reductase; Hm2: *Helminthosporium carbonum* susceptibility; HPC: High Performance Computing; IR: Information Retrival; JA: Jasmonic acid; LOX: Lypoxigenase; LTP: Lipid transfer protein; MAPK: Mithogen activated protein kinase; matK: Maturase K; MLO: Mildew resistance Locus; MPK: Mitogen-Activated Protein Kinase; MYC: Jasmonate Insensitive; NaD: *Nicotiana alata* Defensin; NCBI: National Center for Biotechnology Information; NO: Nitric oxide; NPR: Noneexpressor of Pathogenesis-related Gene; ORA59: Octadecanoid-Responsive Arabidopsis; PAD: Phytoalexin Deficient; PDF: Plant Defensin; pepr: Perception of the Arabidopsis Danger Signal Peptide; PMID: PubMed ID; PR: Pathogenesis-related; psb: Photosystem II protein; rbcL: Rubisco Large Subunit; RIN: RPM-Interacting Protein; RPM1: *Pseudomonas syringae pv maculicola*1; RPS: Ribosomal protein small subunit; SA: Salicylic acid; SCP: Secure copy protocol; SSH: Secure Shell; tax-ID: Taxonomy ID (NCBI); TGA: Transcription Factor TGA

Acknowledgments

The authors thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil) and FACEPE (Fundação de Amparo à Pesquisa do Estado de Pernambuco, Brazil) for valuable financial support and fellowships. The experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg (https://hpc.uni.lu). We thank the R3 group from UL's Bioinformatics core for given support, especially Yohan Jarosz, Noua Toukourou, Maharshi Vyas and Christophe Trefois. We also thank National Laboratory of Scientific Computing (LNCC, Petrópolis, Rio de Janeiro, Brazil) where scalability tests were runned and Prof. Dr. J. Miguel Ortega (Biodados Laboratory at UFMG, Belo Horizonte, Brazil) for allocating Sagarana HPC resources (CPAD-ICB-UFMG) for testing LAITOR4HPC. Part of this paper was presented and awarded as poster presentation during the X-meeting (2016) and was also included as part of the PhD. thesis of the author Marx Oliveira-Lima.

Authors' contributions

BP was responsible for LAITOR's HPC implementation and parallelization, as much as the case study design. ML designed and executed the case studies. SD provided support for Gaia cluster usage and software installation. ABI and ABV designed the case studies. RS designed the HPC implementation project. ABS designed the HPC implementation project and has coordinated it. BP, ML, ABI, ABV and ABS have participated in drafting the manuscript. All authors have read and approved the final manuscript.

Funding

The authors of this work are funded by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil) and FACEPE (Fundação de Amparo à Pesquisa do Estado de Pernambuco, Brazil) with valuable financial support and fellowships of the researchers and umbrella project (Rede InterSys) in which this work was supported.

Availability of data and materials

The datasets analyzed for this study can be found in Zenodo repository [zenodo.org], as well as the files required for LAITOR4HPC installing process. The dataset can be found by searching for a repository called LAITOR4HPC [https://zenodo.org/record/1717329#.XAXWpmhKhPY] and is registered at DOI: https://doi.org/10.5281/zenodo.1717329. All datasets generated and analyzed for this study are included in the manuscript and the supplementary files.

Ethics approval and consent to participate

The present work involves no sensitive data that demand ethical approval procedures. The study and the manuscript involve no human and no animal subjects. Additionally, no details, no images, no videos related to individual or collective data of people were accessed, produced, or made available in the scope of the present publication. All material analyzed regarded article abstracts publicly available, and a consent form is not applicable. Therefore, we consent on making all manuscript data available.

Consent for publication

Since no details, images or videos relating to individual person was accessed or made available with this research and instead and all the material (abstracts and programs) are of public domain, consent forms are not applicable. Therefore, we consent on making all manuscript data available.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details

[1]Genetics Department, Laboratório de Genética e Biologia Vegetal, Universidade Federal de Pernambuco, Recife, Pernambuco, Brazil. [2]University of Luxembourg, Luxembourg Centre for Systems Biomedicine, Bioinformatics Core, Esch-sur-Alzette, Luxembourg. [3]Queen Mary University of London, Centre for Translational Bioinformatics, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Charterhouse Square, London, UK.

Received: 14 June 2019 Accepted: 19 June 2020
Published online: 24 August 2020

References

1. Andrade MA, Bork P. Automated extraction of information in molecular biology. FEBS Lett. 2000;476:12–7.
2. Resource NCBI. Coordinators. Database resources of the National Center for biotechnology information. Nucleic Acids Res. 2017;45:D12–7.
3. Kitano H. Systems biology: a brief overview. Science. 2002;295:1662–4.
4. Gawron P, Ostaszewski M, Satagopam V, Gebel S, Mazein A, Kuzma M, et al. MINERVA - a platform for visualization and curation of molecular interaction networks (in revision). Syst Biol Appl. 2016;2:1–6 https://doi.org/10.1038/npjsba.2016.20.
5. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet. 2012;13:829–39.
6. Ghosh S, Matsuoka Y, Asai Y, Hsin K-Y, Kitano H. Software for systems biology: from tools to integrated platforms. Nat Rev Genet. 2011;12:821–32.
7. Pavlopoulos GA, Malliarakis D, Papanikolaou N, Theodosiou T, Enright AJ, Iliopoulos I. Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. Gigascience. 2015;4:38.
8. Barbosa-Silva A, Soldatos TG, Magalhães ILF, Pavlopoulos GA, Fontaine J-F, Andrade-Navarro MA, et al. LAITOR--literature assistant for identification of terms co-occurrences and relationships. BMC Bioinform. 2010;11:70.
9. Barbosa-Silva A, Fontaine J-F, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro MA. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. BMC Bioinform. 2011;12:435 https://doi.org/10.1186/1471-2105-12-435.
10. Krallinger M, Valencia A. Text-mining and information-retrieval services for molecular biology. Genome Biol. 2005;6:224.
11. Mika S, Rost B. NLProt: Extracting protein names and sequences from papers. Nucleic Acids Res. 2004;32:634–7 WEB SERVER ISS.
12. Fluck J, Hofmann-Apitius M. Text mining for systems biology. Drug Discov Today. 2014;19:140–4 https://doi.org/10.1016/j.drudis.2013.09.012.
13. Trindade D, Orsine LA, Barbosa-Silva A, Donnard ER, Ortega JM. A guide for building biological pathways along with two case studies: hair and breast development. Methods. 2015;74:16–35.
14. Snel B, Lehmann G, Bork P, Huynen MA. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. 2000;28:3442–4.
15. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009;37(Database issue):D412–6.
16. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45:D362–8.
17. Chakradhar T, Hindu V, Reddy PS. Genomic-based-breeding tools for tropical maize improvement. Genetica. 2017;145:525–39.
18. Moreno M, Segura A, García-Olmedo F. Pseudothionin-St1, a potato peptide active against potato pathogens. Eur J Biochem. 1994;223:135–9 http://www.ncbi.nlm.nih.gov/pubmed/9885189.
19. Nawrot R, Barylski J, Nowicki G, Broniarczyk J, Buchwald W, Goździcka-Józefiak A. Plant antimicrobial peptides. Folia Microbiol (Praha). 2014;59:181–96 https://doi.org/10.1007/s12223-013-0280-4.
20. Pelegrini PB, Lay FT, Murad AM, Anderson MA, Franco OL. Novel insights on the mechanism of action of α-amylase inhibitors from the plant defensin family. Proteins Struct Funct Genet. 2008;73:719–29.
21. Parisi K, Shafee TMA, Quimbar P, van der Weerden NL, Bleackley MR, Anderson MA. The evolution, function and mechanisms of action for plant defensins. Semin Cell Dev Biol. 2019;88:107–18 https://doi.org/10.1016/j.semcdb.2018.02.004.
22. Varrette SP, Bouvry P, Cartiaux H, Georgatos F. Management of an academic HPC cluster: The UL experience. Bologna: International Conference on High Performance Computing & Simulation (HPCS); 2014. p. 959–67.
23. Tange O. GNU parallel - the command-line power tool. USENIX; 2011.
24. Mika S, Rost B. ROSTLAB –NLProt; 2004.
25. PHP. PHP: Hypertext Preprocessor (Version 7). https://www.php.net/. Accessed 20 July 2020.
26. MySQL. MySQL (Version 8). https://www.mysql.com/. Accessed 20 July 2020.
27. SQLite. SQLite (Version 3). https://www.sqlite.org/. Acessed 20 July 2020.
28. Funahashi A, Morohashi M, Kitano H, Tanimura N. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. Biosilico. 2004;1:159–62.
29. He X, Jiang J, Wang C, Denesh K. ORA59 and EIN3 interaction couples jasmonate-ethylene synergistic action to antagonistic salicylic acid regulation of PDF expression. J Integr Plant Biol. 2017;59:275–87.
30. Wu Y, Zhang D, Chu JY, Boyle P, Wang Y, Brindle ID, et al. The Arabidopsis NPR1 protein is a receptor for the plant defense hormone salicylic acid. Cell Rep. 2012;1:639–47.
31. Huffaker A, Ryan CA. Endogenous peptide defense signals in Arabidopsis differentially amplify signaling for the innate immune response. Proc Natl Acad Sci U S A. 2007;104:10732–6.
32. Yan J, Zhang C, Gu M, Bai Z, Zhang W, Qi T, et al. The Arabidopsis CORONATINE INSENSITIVE1 protein is a jasmonate receptor. Plant Cell. 2009;21:2220–36.
33. Pieterse CMJ, Leon-Reyes A, Van Der Ent S, Van Wees SCM. Networking by small-molecule hormones in plant immunity. Nat Chem Biol. 2009;5:308–16.
34. Wu F, Ma J, Meng Y, Zhang D, Pascal Muvunyi B, Luo K, et al. Potential DNA barcodes for Melilotus species based on five single loci and their combinations. PLoS One. 2017;12:e0182693.
35. Gao T, Ma X, Zhu X. Use of the psbA-trnH region to authenticate medicinal species of Fabaceae. Biol Pharm Bull. 2013;36:1975–9.
36. He Y, Yu C, Zhou L, Chen Y, Liu A, Jin J, et al. Rubisco decrease is involved in chloroplast protrusion and Rubisco-containing body formation in soybean (*Glycine max*.) under salt stress. Plant Physiol Biochem. 2014;74:118–24.
37. Vitlin Gruber A, Feiz L. Rubisco assembly in the chloroplast. Front Mol Biosci. 2018;5:24.
38. Fernández JM, Hoffmann R, Valencia A. iHOP web services. Nucleic Acids Res. 2007;35:W21–6 Web Server issue.

39. Breen S, Williams SJ, Outram M, Kobe B, Solomon PS. Emerging insights into the functions of pathogenesis-related protein 1. Trends Plant Sci. 2017;22:871–9.
40. Luo J, Xu Z, Tan Z, Zhang Z, Ma L. Neuropeptide receptors NPR-1 and NPR-2 regulate Caenorhabditis elegans avoidance response to the plant stress hormone methyl salicylate. Genetics. 2015;199:523–31.
41. Hatsugai N, Hillmer R, Yamaoka S, Hara-Nishimura I, Katagiri F. The μ subunit of Arabidopsis adaptor Protein-2 is involved in effector-triggered immunity mediated by membrane-localized resistance proteins. Mol Plant-Microbe Interact. 2016;29:345–51.
42. Gassmann W, Hinsch ME, Staskawicz BJ. The Arabidopsis RPS4 bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. Plant J. 1999;20:265–77.
43. Nishimura MT, Dangl JL. Arabidopsis and the plant immune system. Plant J. 2010;61:1053–66.
44. Bari R, Jones JDG. Role of plant hormones in plant defence responses. Plant Mol Biol. 2009;69:473–88 https://doi.org/10.1007/s11103-008-9435-0.
45. Adams E, Turner J. COI1, a jasmonate receptor, is involved in ethylene-induced inhibition of Arabidopsis root growth in the light. J Exp Bot. 2010;61:4373–86.
46. Cerrudo I, Keller MM, Cargnel MD, Demkura PV, de Wit M, Patitucci MS, et al. Low red/far-red ratios reduce Arabidopsis resistance to Botrytis cinerea and Jasmonate responses via a COI1-JAZ10-dependent, salicylic acid-independent mechanism. Plant Physiol. 2012;158:2042–52 https://doi.org/10.1104/pp.112.193359.
47. Mira MM, Wally OSD, Elhiti M, El-Shanshory A, Reddy DS, Hill RD, et al. Jasmonic acid is a downstream component in the modulation of somatic embryogenesis by Arabidopsis class 2 phytoglobin. J Exp Bot. 2016;67:2231–46.
48. Ndamukong I, Al AA, Thurow C, Fode B, Zander M, Weigel R, et al. SA-inducible Arabidopsis glutaredoxin interacts with TGA factors and suppresses JA-responsive PDF1.2 transcription. Plant J. 2007;50:128–39.
49. Li M, He Q, Ma J, He F, Zhu Y, Chang C, et al. PPICurator: a tool for extracting comprehensive protein-protein interaction information. Proteomics. 2019;19:e1800291.
50. Bachman JA, Gyori BM, Sorger PK. FamPlex: a resource for entity recognition and relationship resolution of human protein families and complexes in biomedical text mining. BMC Bioinform. 2018;19:248.
51. Sun D, Wang M, Li A. MPTM: a tool for mining protein post-translational modifications from literature. J Bioinforma Comput Biol. 2017;15:1740005.
52. Price MN, Arkin AP. PaperBLAST: Text Mining Papers for Information about Homologs. mSystems. 2017;2:1–10.

Publisher's Note