UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

GENOMIC AND TRANSCRIPTOMIC ANALYSIS APPLIED TO AGRIBUSINESS: FOCUS ON BIOTIC AND ABIOTIC STRESS

PhD THESIS

FABRÍCIO BARBOSA MONTEIRO ARRAES

PORTO ALEGRE – 2020

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

ANÁLISE DE DADOS DE GENÔMICA E TRANSCRIPTOMICA APLICADA AO AGRONEGÓCIO: FOCO EM ESTRESSES BIÓTICOS E ABIÓTICOS

Fabrício Barbosa Monteiro Arraes

Tese de Doutorado submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS como requisito parcial para a obtenção do título de Doutor

Orientadora: Dra. Maria Fátima Grossi de Sá

PORTO ALEGRE
JANEIRO – 2020

“O que sabemos é uma gota; o que ignoramos é um oceano”.

Isaac Newton

Este trabalho foi concluído no Laboratório de Interação Molecular Planta-Patógeno (EMBRAPA - Recursos Genéticos e Biotecnológica - CENARGEN), Laboratório de Biotecnologia Vegetal (EMBRAPA/SOJA). O projeto foi subsidiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e pela Fundação de Amparo à Pesquisa do Distrito Federal (FAPDF).

AGRADECIMENTOS

Àquele que descubro um pouco a cada dia e que inquieta meu espírito, instiga o coração e me leva sempre a buscar novas perguntas e respostas;

À minha amada família, principalmente meus pais Maria de Fátima Barbosa Monteiro e Joel Arraes† que sempre acreditaram e investiram em mim. Amo vocês!

À minha orientadora, Dra. Maria Fátima Grossi de Sá, pela oportunidade, confiança no meu trabalho, amizade, compreensão e aconselhamento durante esse período. Também ao Dr. Etienne G. J. Danchin e toda sua equipe no Instituto de Pesquisa Agronômica (INRA – França), pela contribuição científica e pelo acolhimento durante meu Doutorado Sanduíche (2018-2019).

À Dra. Carolina Morgante, pela amizade, parceria e contribuição significativa com o trabalho, além de ser para mim um exemplo de conduta de vida. Também ao Dr. Roberto Coiti Togawa e toda sua equipe, pela amizade e conhecimentos na área de Bioinformática.

A todos no Laboratório de Interação Molecular Planta-Praga (LIMPP) e a toda equipe da Embrapa Recursos Genéticos e Biotecnologia por também terem sido minha família durante todo o período de conclusão do Doutorado.

Ao amigo Diogo Martins de Sá, pela colaboração, ajuda, pelas noites de estudo em claro e pelas extensas e acaloradas discussões científicas.

À minha amiga e irmã de consideração Raquel Sampaio que tanto me deu forças nos momentos mais difíceis, iluminando meu caminho com seu coração puro.

Ao PPGBCM, mais especificamente ao Coordenador e aos professores do programa que contribuíram tanto com a minha formação acadêmica. Também à Sílvia Centeno pelo acolhimento no período do Doutorado. Sem a sua alegria e ajuda tudo teria sido muito mais difícil.

Aos órgãos de fomento CAPES, CNPq, EMBRAPA e FAPDF pelo apoio financeiro, estrutura e pela bolsa de estudos, essenciais para a conclusão deste projeto.

Aos membros da minha Comissão de Acompanhamento (Dr. Diego Bonatto e Dra. Érika Valéria Albuquerque Freire) e aos membros da Banca Examinadora (Dr. Rogério Margis, Dr. Henrique Marques-Souza, Dra. Maria Helena Zanettini e Dr. Giancarlo Pasquali) pelo tempo disponibilizado e pela significativa contribuição com o trabalho.

Agradeço imensamente a todos!

SUMMARY

RESUMO GERAL

O aumento nos danos causados pelos estreses bióticos e abióticos vem apresentando um profundo impacto na produtividade das culturas de interesse agronômico. Dentre os principais fatores bióticos que acometem o agronegócio, os insetos-praga constituem a principal classe de patógenos que causam grandes perdas em cultivares como soja, milho e algodão. Devido ao uso intensivo de inseticidas e consequente favorecimento da seleção de populações de insetos resistentes, o desenvolvimento de novas formas de controle sustentável de insetos-praga se faz cada vez mais necessária. Além dos fatores bióticos, estresses abióticos também influenciam negativamente a agricultura, reduzindo consideravelmente a produção. Prevê-se que a competição por recursos hídricos se intensificará ainda mais nas regiões agrícolas, constituindo-se em fator chave na bioeconomia mundial. Dessa forma, a Biotecnologia possui um papel fundamental no aprimoramento e desenvolvimento de novas tecnologias visando o melhoramento de plantas. Dentre as principais fontes de conhecimento da atualidade para tal fim destacam-se a Genômica e a Transcriptômica. Desta forma, a presente tese de Doutorado avaliou dados em bancos de genômica e transcriptômica disponíveis para desenvolver e melhorar estratégias para aumentar a tolerância/resistência das culturas tanto a estresses biótico (insetos-pragas) e abióticos (seca).

No Capítulo 01 foram avaliados estado atual do conhecimento sobre a via de miRNA e siRNA em 168 espécies de insetos e a estrutura de domínios conhecidos abrangendo as principais proteínas do mecanismo. Este estudo é de grande relevância visto ao potencial de utilização do mecanismo de RNAi para o controle de insetos-praga. Os elementos analisados foram identificados em bancos de dados públicos de genomas e transcritomas de espécies da ordem Coleoptera, Diptera, Hemiptera, Hymenoptera e Lepidoptera. Dentre os domínios analisados, dsrm, PAZ, Plataforma, Ribonuclease III (RIIID) e Helicase foram os que forneceram mais informações sobre a variabilidade identificada. Também foi possível concluir que a estabilidade dos complexos de microprocessadores responsáveis pela produção de miRNA e siRNA em insetos é o ponto chave na eficiência da biogênese desses pequenos RNAs.

No Capítulo 02 foi avaliada a importância do fitormônio etileno em resposta à seca. Para isso, foram identificados *in silico* 176 genes de soja descritos como participantes tanto na biossíntese quanto na transdução de sinal mediada por este fitormônio. A partir de genes expressos diferencialmente no banco de dados do transcriptoma, foi analisada a expressão relativa por qPCR de alguns genes selecionados em cultivares de soja tolerantes e suscetíveis à déficit hídrico. Nas mesmas amostras, altos níveis de produção de etileno foram detectados e

foram diretamente correlacionados com os níveis de fração livre do seu precursor. Sendo assim, a análise *in silico*, combinada com a quantificação da produção de etileno (e seu precursor) e experimentos de RT-qPCR, permitiu uma melhor compreensão da importância do etileno em nível molecular nesta cultura, bem como de seu papel na resposta a estresses abióticos.

Assim, como os dados demonstraram, a análise de dados de genômica e transcritômica pode contribuir significativamente para o desenvolvimento do agronegócio global, principalmente por fornecer conhecimento para a geração e otimização de ferramentas que visam o melhoramento sustentável das culturas de interesse agronômico, principalmente em resposta à estresses bióticos e abióticos.

GENERAL ABSTRACT

The increasing damage caused by biotic and abiotic stresses has had a profound impact on crop yields worldwide. Among the main biotic factors affecting agribusiness, insect-pests constitute the main class of pathogens that cause large losses in cultivars such as soybean, maize and cotton. Due to the intensive use of insecticides and the consequent favoring of the selection of resistant insect populations, the development of new forms of sustainable pest control is becoming increasingly necessary. In addition to biotic factors, abiotic stresses also negatively influence agriculture, considerably reducing production. Competition for water resources is expected to intensify further in agricultural regions, becoming a key factor in the world bioeconomy. Thus, Biotechnology has a fundamental role in the improvement and development of new technologies aiming at plant breeding. Among the main sources of current knowledge for this purpose are Genomics and Transcriptomics. Thus, the present PhD thesis evaluated data in available genomic and transcriptomic databases to develop and improve strategies to increase crop tolerance/resistance to both biotic (insect pest) and abiotic (drought) stresses.

In Chapter 01 it was evaluated the current state of knowledge about miRNA and siRNA pathway in 168 insect species and the structure of known domains covering the main proteins of the mechanism. This study is of great relevance given the potential use of the RNAi mechanism for insect-pest control. The analyzed elements were identified in public databases of genomes and transcriptomes from species belonging to Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera insect orders. Among the domains analyzed, dsrm, PAZ, Platform, Ribonuclease III (RIIID) and Helicase provided the most information on the identified variability. It was also possible to conclude that the stability of microprocessor complexes responsible for miRNA and siRNA production in insects is the key point in the biogenesis efficiency of these small RNAs.

In Chapter 02, the importance of ethylene phytohormone in response to drought was evaluated. For this, 176 soybean genes described as participating in both biosynthesis and signal transduction mediated by this phytohormone were identified *in silico*. From genes differentially expressed in the transcriptome database, the relative expression by qPCR of some selected genes in tolerant and susceptible to water deficit soybean cultivars was analyzed. In the same samples, high levels of ethylene production were detected and were directly correlated with the free fraction levels of its precursor. Thus, *in silico* analysis, combined with the quantification of ethylene production (and its precursor) and RT-qPCR experiments, allowed a better

understanding of the importance of molecular ethylene in soybean, as well as its role in the response to abiotic stresses.

Thus, as the data have shown, genomic and transcriptomic data analysis can contribute significantly to the development of global agribusiness, mainly by providing knowledge for the generation and optimization of tools aimed at the sustainable improvement of crops of agronomic interest, mainly in response to biotic and abiotic stresses.

1. GENERAL INTRODUCTION

Climate change, an increasing world population, and genetic erosion are the main factors that alert us to the need to improve crop adaptation, tolerance, and productivity. There is therefore a continuing need to develop novel cultivars better adapted to different biomes, with improved tolerance to biotic and abiotic stresses, and with superior yield and quality (Arzani and Ashraf, 2016). Classic plant breeding, despite being a slow and usually difficult process, has made great contributions over the years. This method has been used mainly to add traits to an already otherwise variety/cultivar. In contrast, genetic engineering has provided a complementary tool to introduce desirable genes horizontally for traits of interest in crop plants. The association between genetic engineering tools and classic plant breeding has accelerated crop improvement to a more accurate and efficient technique. Additionally, the development of new biotechnological tools increases agricultural sector competitiveness in internal and external markets (Limera et al., 2017).

The advances in functional genomics and other omics technologies over the years have revealed the biological function and features of innumerable elements of genetic engineering. The exploration of these elements has allowed the obtaining of a greater number of elite events in significantly reduced time. Thus, several genes of interest have been associated with agronomic traits of great economic interest, and several new biotechnological tools have been developed to overcome the main limitations in the agricultural sector.

Thus, it is impossible to talk about genetic engineering and biotechnology without addressing the scientific advances related to recombinant DNA technology. Since the elucidation of its chemical structure in 1953, DNA has become one of the most studied molecules in the world (Goodwin et al., 2016; Kulski, 2016). The DNA molecule is composed of nucleotides linked in specific and unique combinations that can be identified by sequencing methodology. DNA sequencing became common in the 1970s - 1980s with chemical degradation methods (Maxam and Gilbert) and dideoxynucleotide (ddNTP) chain termination (Sanger Method). Currently, the Sanger method is not widely used because it is quite labor intensive, expensive and time consuming (Maxam and Gilbert, 1977; Sanger et al., 1977).

In 2005, the first DNA sequencing technologies known as *Next Generation Sequencing* (NGS) emerged, enabling a new approach to *High Throughput Sequencing* (HGS). In subsequent years, several NGS platforms were developed, based mainly on the methodologies: (i) *pyrosequencing* (454-Roche) (Siqueira et al., 2012) (ii) *linkage sequencing* (SOLiD)

(Valouev et al., 2008); (iii) *semiconductor methodology* (Ion Torrent) (Rothberg et al., 2011); (iv) *synthesis sequencing* (Illumina) (Guo et al., 2008); and (v) *long-reads sequencing* (Pacific Biosciences and Oxford Nanopore) (Clarke et al., 2009; Eid et al., 2009).

Currently, the most widely used technologies for genome and transcriptome sequencing are: (i) synthesis sequencing (Illumina); and (ii) long reads sequencing (van Dijk et al., 2018). In Illumina technology, after solid phase template enrichment, a mixture of primers, DNA polymerase and modified nucleotides are added to a flow cell. Each nucleotide is blocked and labeled with a cleavable fluorophore specific to each base. During each cycle, fragments in each cluster incorporated only one nucleotide with the 3' group blocked, thus avoiding additional incorporations. After nucleotide incorporation, unincorporated bases are washed, and the lane is photographed by total internal reflection fluorescence microscopy to identify the base that was incorporated into each cluster. The fluorophore is then cleaved and the 3'-OH is regenerated at the beginning of a new cycle (Goodwin et al., 2016; Guo et al., 2008). With technology upgrades (NovaSeq) it is possible to sequence with high quality (less than 1 % of error) up to 3 Tb per flow cell with single- or paired-reads with the length ranging from 50 to 250 nucleotides.

Long-read sequencing can be divided according to the technology used. In Pacific Biosciences (PacBio) sequencing methodology, templates are prepared and linked to adapters, resulting in a circular DNA molecule with single stranded DNA (ssDNA) regions at each end and double stranded (dsDNA) in the middle, which pass through a length selection protocol for removing large and small DNA fragments. The SMRTbell templates (primers and an efficient DNA polymerase plus selected DNA fragments) are added to *zero-mode waveguides* (ZMWs), which are nanoscale observation chambers that should be loaded with exactly one SMRTbell template. During chain elongation, the polymerase within each ZMW incorporates fluorescently labeled nucleotides, and emit a fluorescent signal that is recorded by a real-time camera (Ardui et al., 2018; Eid et al., 2009). Depending on the insert length and sample quality, the PacBio Sequel II System (SMRT Cell 8M) can generate up to 4,000,000 high fidelity (HiFi) reads with more than 99.0 % accuracy, where the longest read can reach a length up to 175 Kb. It is noteworthy that the high accuracy promised by the PacBio is not yet observed in the day-to-day of the laboratory.

On the other hand, unlike other sequencing methods based on Polymerase Chain Reaction (PCR), the Oxford Nanopore Technologies (ONT) relies on reading a DNA molecule as it passes through a nanopore in a membrane. For library construction with 8-10 Kb DNA fragments, two different adapters are attached at each end of the dsDNA fragment, one to direct the DNA to the nanopore and the other to direct the DNA to the membrane surface. As DNA translocates through the nanopore, a voltage change is observed, with a specific intensity for each nucleotide. Several parameters, including the magnitude and duration of the shift, are recorded, and can be interpreted as a *k*-mer sequence. Once the DNA continues to translocates through the pore adapter and onto the complement strand, it is possible to create a consensus sequence called a *2D read* (Clarke et al., 2009). According to Oxford Nanopore, the current technology associated to PromethION 48 (48 flow cells) can direct sequencing at real time of native DNA/RNA samples, with any length (short to ultra-long) and generate up to 7.6 Tb per run (around 159 Gb per flow cell).

PacBio and ONT associated with novel long-range assays have revolutionized *de novo* genome assembly by automating the reconstruction of reference-quality genomes. Associated with these technologies, a new approach of long-range contact information, the Hi-C sequencing, which was originally proposed to study the 3D genome organization, is becoming an economical method for generating chromosome-scale scaffolds (Ghurye et al., 2019, 2017; Pal et al., 2019). The Hi-C technique protocol starts with restriction digestion of a cross-linked genome, followed by fill in and repair of digested ends with the incorporation of biotin-linked nucleotides. The repaired ends are then re-ligated. Finally, the cross-linking is reversed, and associated proteins are removed. The resulting DNA fragments are used as templates for the construction of Illumina paired-end libraries to be subsequently mapped into the original assembly. This information can provide linkage information among the original scaffolds and increase the assembly resolution in a megabase scale (Ghurye et al., 2019; Pal et al., 2019).

Thus, all NGS methodologies previously described have the great advantage of providing direct and parallel sequencing of thousand to millions of DNA molecules, considerably increasing the scale and resolution of genome and transcriptome analyzes. In addition, the reduced sample amount required (*e.g.*, single cell sequencing) and the significant decrease in the cost per sequenced nucleotide are the major advances achieved with NGS, which can be directly applied to sequence complete genomes, metagenomes, RNA-Seq, exomes, long and/or small non-coding RNAs, amplicons, ChIP-Seq (chromatin immunoprecipitation) and several other applications.

The increasing use of NGS technologies in a wide and routine way has become an important tool in understanding the biological diversity of different ecosystems and can be directly applied in plant breeding, especially in transgenic approaches. This statement is reflected in data deposited on National Center for Biotechnology Information (NCBI) database until January 2020, in which 44,642 Bioprojects are directly or indirectly related to genomic and transcriptomic research, with species belonging to 58 different plant orders. (Table 1). Among the plant orders with the largest number of annotated genome assemblies, stand out Poales, Fabales and Malvales, including the five world's main crops: *Zea mays*, Triticum spp. and *Avena sativa* (Poales); *Glycine max* (Fabales) and *Gossypium hirsutum* (Malvales). These available data provide important information about the physiology, development and production of these plants in the most diverse conditions. It also helps to characterize new mechanisms of tolerance/resistance or susceptibility to the most diverse types of biotic and abiotic stresses (Afzal et al., 2020; Meena et al., 2017; Schreiber et al., 2018).

Since plant pathogens are an important concern to agribusiness, the development of disease-resistant plants through biotechnological approaches is important. It is desirable the development of economically important crops through Genetically Modified (GM) lines that not only exhibit durable and broad resistance spectrum to several plant pathogens, but also are biosafe to the environment and consumers (Feliciano, 2019). To achieve this goal, it is important to elucidate molecular processes related to the physiology and development of these pathogens, as well as their adaptive mechanisms aiming at circumventing host plant defenses during a compatible interaction. Genomics and transcriptomics are standard technologies to elucidate such molecular processes and provide a large set of biomolecules that can be applied to development of GM disease-resistant/less susceptible crops. In view of this, insect-pests stand out as one of the major classes of phytopathogens due to its direct association with substantial crop losses worldwide through direct damage and transmission of plant diseases (Douglas, 2018).

Similarly to plant orders, it is possible to observe in NCBI database a large number of Bioprojects (9,192) associated with genome and transcriptome sequencing of most known insect orders, highlighting Coleoptera, Hemiptera and Lepidoptera, whose species can cause losses mainly from mechanical damage or transmission of other pathogens (Table 2). These data provide valuable information for the development of strategies to pest control, such as the identification of possible targets for RNA interference (RNAi)-mediated gene knockdown (Darrington et al., 2017; Kanakala and Ghanim, 2016; Li et al., 2019; Vogel et al., 2018; Xu et al., 2016)

Table 1. Plant genome and transcriptome data based on National Center for Biotechnology Information (NCBI)[1]

Plant Orders	Bioprojects[2]			Assemblies[3]	Annotation
	Genomes	Transcriptomes	Others		
Acorales	0	2	1	0	0
Alismatales	232	56	12	3	2
Amborellales	2	4	1	1	1
Apiales	23	105	9	3	1
Aquifoliales	0	6	2	0	0
Arecales	11	52	29	5	2
Asparagales	57	665	27	8	4
Asterales	26	253	36	11	5
Austrobaileyales	0	3	2	0	0
Berberidopsidales	0	1	1	0	0
Brassicales	1,642	4,555	517	61	17
Buxales	0	1	2	0	0
Canellales	0	1	2	0	0
Caryophyllales	37	618	42	19	4
Celastrales	0	18	4	0	0
Ceratophyllales	0	0	1	0	0
Chloranthales	0	0	2	0	0
Commelinales	0	5	2	1	0
Cornales	3	16	3	1	1
Crossosomatales	0	1	1	0	0
Cucurbitales	29	178	32	12	7
Cycadales	1	5	4	0	0
Dioscoreales	5	6	1	4	0
Dipsacales	2	16	6	0	0
Ericales	27	190	26	11	3
Fabales	122	1,246	98	36	25
Fagales	20	141	37	22	3
Garryales	1	3	2	0	0
Gentianales	15	115	26	7	3
Geraniales	11	9	4	0	0
Ginkgoales	1	25	5	0	0
Gnetales	0	3	4	0	0
Gunnerales	0	1	1	0	0
Lamiales	103	287	35	22	8
Laurales	3	29	2	2	1
Liliales	4	43	8	0	0
Magnoliales	7	24	4	2	0
Malpighiales	2,518	4,426	78	15	8
Malvales	42	265	21	20	18
Myrtales	46	91	18	10	6
Nymphaeales	4	4	2	1	1
Oxalidales	3	4	3	2	1
Pandanales	0	4	2	1	0
Petrosaviales	2	0	0	0	0
Pinales	14	227	105	6	1
Piperales	1	22	3	0	0
Poales	8,940	11,951	656	51	28
Proteales	3	14	3	2	1
Ranunculales	200	67	8	5	3
Rosales	69	403	68	35	19
Santalales	4	11	3	1	0
Sapindales	41	250	24	14	6
Saxifragales	11	398	7	3	0
Solanales	192	844	134	32	15
Trochodendrales	0	0	2	0	0
Vitales	15	241	24	7	2
Zingiberales	8	67	17	5	6
Zygophyllales	1	1	2	0	0
Total	**14,498**	**27,973**	**2,171**	**441**	**202**

[1] Counting by NCBI taxonomy IDs and updating in January 2020.

[2] Filtered by data type: Genome Sequencing, Transcriptomes and Others.

[3] Genome assemblies derived from surveillance or anomalous projects were excluded. Only representatives were kept.

Table 2. Insect genome and transcriptome data based on National Center for Biotechnology Information (NCBI)[1]

Insect Orders	**Bioprojects**[2]			**Assemblies**[3]	**Annotation**
	Genomes	**Transcriptomes**	**Others**		
Acerentomata	1	4	0	0	0
Archaeognatha	0	28	1	1	0
Blattodea	2	110	16	4	3
Coleoptera	31	357	66	22	15
Collembola	0	51	4	4	2
Dermaptera	0	21	1	0	0
Diplura	0	20	0	2	0
Diptera	241	2,506	549	161	57
Ephemeroptera	1	11	3	2	0
Grylloblattodea	0	6	0	0	0
Hemiptera	819	564	49	35	17
Hymenoptera	75	645	64	96	56
Lepidoptera	61	581	120	67	26
Mecoptera	0	14	0	0	0
Neuroptera	2	70	0	0	0
Odonata	0	133	0	2	0
Orthoptera	10	120	11	3	0
Phasmatodea	1	54	12	13	0
Phthiraptera	621	9	1	1	0
Plecoptera	2	18	1	3	0
Psocoptera	32	32	0	0	0
Siphonaptera	0	7	0	1	0
Strepsiptera	0	5	0	1	0
Thysanoptera	783	159	2	2	1
Trichoptera	1	61	2	6	0
Zygentoma	2	18	1	0	0
Total	**2,685**	**5,604**	**903**	**426**	**177**

[1] Counting by NCBI taxonomy IDs and updating in January 2020.

[2] Filtered by data type: Genome Sequencing, Transcriptomes and Others.

[3] Genome assemblies derived from surveillance or anomalous projects were excluded. Only representatives were kept.

Thus, it is clear the importance of genomics and transcriptomics for the development of elite crops, not only for the identification of genes related to traits of interest, but also for the identification of new genetic elements that can be used as biotechnological tools as gene promoters, terminators, among others. The generation and analysis of new sequence databases integrated with existing ones is essential for the advancement of global agribusiness in an eco-sustainable manner.

REFERENCES

Afzal, M., Alghamdi, S.S., Migdadi, H.H., Khan, M.A., Nurmansyah, Mirza, S.B., El-Harty, E., 2020. Legume genomics and transcriptomics: from classic breeding to modern technologies. Saudi J Biol Sci 27, 543–555. doi:10.1016/j.sjbs.2019.11.018

Ardui, S., Ameur, A., Vermeesch, J.R., Hestand, M.S., 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 46, 2159–2168. doi:10.1093/nar/gky066

Arzani, A., Ashraf, M., 2016. Smart engineering of genetic resources for enhanced salinity tolerance in crop plants. CRC. Crit. Rev. Plant Sci. 35, 146–189. doi:10.1080/07352689.2016.1245056

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., Bayley, H., 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. 4, 265–270. doi:10.1038/nnano.2009.12

Darrington, M., Dalmay, T., Morrison, N.I., Chapman, T., 2017. Implementing the sterile insect technique with RNA interference - a review. Entomol Exp Appl 164, 155–175. doi:10.1111/eea.12575

Douglas, A.E., 2018. Strategies for enhanced crop resistance to insect pests. Annu. Rev. Plant Biol. 69, 637–660. doi:10.1146/annurev-arplant-042817-040248

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., et al., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138. doi:10.1126/science.1162986

Feliciano, D., 2019. A review on the contribution of crop diversification to Sustainable Development Goal 1 “No poverty” in different world regions. Sust. Dev. doi:10.1002/sd.1923

Ghurye, J., Pop, M., Koren, S., Bickhart, D., Chin, C.-S., 2017. Scaffolding of long read assemblies using long range contact information. BMC Genomics 18, 527. doi:10.1186/s12864-017-3879-z

Ghurye, J., Rhie, A., Walenz, B.P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A.M., Koren, S., 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput. Biol. 15, e1007273. doi:10.1371/journal.pcbi.1007273

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351. doi:10.1038/nrg.2016.49

Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D.H., Sano Marma, M., Meng, Q., Cao, H., Li, X., Shi, S., Yu, L., Kalachikov, S., Russo, J.J., Turro, N.J., Ju, J., 2008. Four-color DNA sequencing with 3’-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc. Natl. Acad. Sci. USA 105, 9145–9150. doi:10.1073/pnas.0804023105

Kanakala, S., Ghanim, M., 2016. RNA interference in insect vectors for plant viruses. Viruses 8. doi:10.3390/v8120329

Kulski, J.K., 2016. Next-generation sequencing — an overview of the history, tools, and “omic” applications, in: Kulski, J.K. (Ed.), Next Generation Sequencing - Advances, Applications and Challenges. InTech. doi:10.5772/61964

Li, F., Zhao, X., Li, M., He, K., Huang, C., Zhou, Y., Li, Z., Walters, J.R., 2019. Insect genomes: progress and challenges. Insect Mol. Biol. 28, 739–758. doi:10.1111/imb.12599

Limera, C., Sabbadini, S., Sweet, J.B., Mezzetti, B., 2017. New biotechnological tools for the genetic improvement of major woody fruit species. Front. Plant Sci. 8, 1418. doi:10.3389/fpls.2017.01418

Maxam, A.M., Gilbert, W., 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74, 560–564. doi:10.1073/pnas.74.2.560

Meena, K.K., Sorty, A.M., Bitla, U.M., Choudhary, K., Gupta, P., Pareek, A., Singh, D.P., Prabha, R., Sahu, P.K., Gupta, V.K., Singh, H.B., Krishanani, K.K., Minhas, P.S., 2017. Abiotic stress responses and microbe-mediated mitigation in plants: the omics strategies. Front. Plant Sci. 8, 172. doi:10.3389/fpls.2017.00172

Pal, K., Forcato, M., Ferrari, F., 2019. Hi-C analysis: from data generation to integration. Biophys. Rev. 11, 67–78. doi:10.1007/s12551-018-0489-1

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., Hoon, J., Simons, J.F., Marran, D., Myers, J.W., Davidson, J.F., Branting, A., Nobile, J.R., Puc, B.P., Light, D., Clark, T.A., Huber, M., Branciforte, J.T., Stoner, I.B., Cawley, S.E., Lyons, M., Fu, Y., Homer, N., Sedova, M., Miao, X., Reed, B., Sabina, J., Feierstein, E., Schorn, M., Alanjary, M., Dimalanta, E., Dressman, D., Kasinskas, R., Sokolsky, T., Fidanza, J.A., Namsaraev, E., McKernan, K.J., Williams, A., Roth, G.T., Bustillo, J., 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352. doi:10.1038/nature10242

Sanger, F., Nicklen, S., Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA 74, 5463–5467. doi:10.1073/pnas.74.12.5463

Schreiber, M., Stein, N., Mascher, M., 2018. Genomic approaches for studying crop evolution. Genome Biol. 19, 140. doi:10.1186/s13059-018-1528-8

Siqueira, J.F., Fouad, A.F., Rôças, I.N., 2012. Pyrosequencing as a tool for better understanding of human microbiomes. J. Oral Microbiol. 4. doi:10.3402/jom.v4i0.10743

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., Sidow, A., Fire, A., Johnson, S.M., 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res. 18, 1051–1063. doi:10.1101/gr.076463.108

van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., Thermes, C., 2018. The third revolution in sequencing technology. Trends Genet. 34, 666–681. doi:10.1016/j.tig.2018.05.008

Vogel, E., Santos, D., Mingels, L., Verdonckt, T.-W., Broeck, J.V., 2018. RNA interference in insects: protecting beneficials and controlling pests. Front. Physiol. 9, 1912. doi:10.3389/fphys.2018.01912

Xu, J., Wang, X.-F., Chen, P., Liu, F.-T., Zheng, S.-C., Ye, H., Mo, M.-H., 2016. RNA interference in moths: mechanisms, applications, and progress. Genes (Basel) 7. doi:10.3390/genes7100088

2. JUSTIFICATIVE

Insect-pests, mainly from Coleoptera and Lepidoptera orders, are the main biotic factor limiting the crop production in the world, causing serious qualitative and quantitative damages. Thus, chemical (insecticide) and biological (*e.g.*, viruses, bacteria, fungi, nematodes and protozoa) control methods are often used to minimize this damage. Due to the indiscriminate use of chemical pesticides and their harmful consequences for the environment, as well as the non-uniform field distribution and coverage of biological agents, these modalities of insect control are becoming ineffective through the selection of resistant insect populations. To address this problem, RNAi-mediated gene silencing has emerged as an important biotechnological tool to be used to control pest insect populations. Theoretically, RNAi is a specific, low cost and efficient technology in which double strand RNA (dsRNA) molecules are administered to target insects: (i) by feeding, mainly by the expression in genetically modified plants; or (ii) topical use, with administration of the naked or protected dsRNA molecules into nanoparticles. Unfortunately, this technology has extremely variable efficiency among the major insect pest orders, both by external factors and by the intrinsic variability of each population. Since most of the studies that characterized RNAi pathway elements were developed in model species, such as *Drosophila melanogaster*, any study analyzing the particularities of this metabolic pathway in different species or insect orders is relevant, to contribute for a better understanding and optimization of this important biotechnological tool.

On the other hand, abiotic stresses, like drought, can also significantly reduce crop yields and restrict crop cultivation worldwide. The implications are many, since not only products, but the whole society is affected. Therefore, understanding drought tolerance and how to exploit plants should be judged not only as agronomic, physiological, or ecological problems, but also as an international goal of economic and political significance. In addition, agribusiness is the sector of society that consumes most freshwater in Brazil. According to the Food and Agriculture Organization of the United Nations (FAO) and the National Water Agency (ANA), in 2019, agriculture consumed around 70 % of the amount of freshwater used in the country. The impacts of that fall mainly on the ecosystem, as groundwater and rivers suffer with a lack of rainfall and may dry up over the years. Thus, the development of cultivars more tolerant to drought is desirable enabling the reduction of water use in agriculture and the cultivation of plants in regions with less water availability. Plant tolerance to drought is a species-specific trait controlled by several factors that play singly and together to tolerate periods of water deficit. Therefore, the identification and understanding of drought tolerance mechanisms are fundamental in the development of new crops, more tolerant to water deficit. For example, the gene expression in tolerant genotypes can be used to study drought tolerance mechanisms and to identify other genotypes with similar traits. More specifically, studies evaluating the participation of phytohormones in the drought response is a key point in this understanding process, aiming at subsequent plant breeding to this stress condition.

Finally, analysis of both plant and insect genome and transcriptome databases can contribute to innovations in Plant Biotechnology in a significant way for Brazilian and world agribusiness, releasing plants less susceptible to both biotic and abiotic stresses on the market.

3. OBJECTIVES

3.1. General Objective

The present study aimed to use available genomics and transcriptomics data to develop and improve strategies to increase the tolerance/resistance of crops to biotic (insect-pests) and abiotic (drought) stresses.

3.2. Specific Objectives

In Chapter 01, entitled "Dissecting variability in the core RNA interference machinery of five insect orders", the study aimed the *in silico* evaluation of structural and phylogenetic variability of the core elements of RNAi machinery in 168 insect species of Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera orders, which represent the majority of important insect species that interact with crops. For this purpose, public databases of genomes or transcriptomes of selected species were used. The integration of these data aimed to identify order-specific differences that could justify the variability in gene knockdown efficiency observed in previous studies.

In Chapter 02, entitled "Implication of Ethylene Biosynthesis and Signaling in Soybean Drought Stress Tolerance", the study initially aimed to identify and characterize *in silico* the members related to biosynthesis and the metabolic pathway signaled by ethylene, based on previous knowledge imported from model plants such as *Arabidopsis thaliana* and *Oryza sativa*. Then, the results were correlated with previous transcriptome data obtained under water deficit conditions to evaluate the expression of identified genes in soybean cultivars (susceptible and drought tolerant). After validation, *in vitro* and *in vivo* analysis allowed to infer a putative role of ethylene in the selected soybean cultivars, cultivated in hydroponic system and submitted to drought.

4. CHAPTER 01: DISSECTING VARIABILITY IN THE CORE RNA INTERFERENCE MACHINERY OF FIVE INSECT ORDERS

Published in	**:**	RNA Biology
Impact Factor	**:**	5.35
DOI	**:**	10.1080/15476286.2020.1861816

RNA BIOLOGY
https://doi.org/10.1080/15476286.2020.1861816

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

Dissecting protein domain variability in the core rna interference machinery of five insect orders

Fabricio Barbosa Monteiro Arraes [a,b], Diogo Martins-de-Sa [c], Daniel D. Noriega Vasquez[b,d], Bruno Paes Melo[b,e], Muhammad Faheem[b,f], Leonardo Lima Pepino de Macedo[b], Carolina Vianna Morgante[b,g,h], Joao Alexandre R. G Barbosa [c], Roberto Coiti Togawa[b], Valdeir Junio Vaz Moreira[a,b,c], Etienne G. J. Danchin [h,i], and Maria Fatima Grossi-de-Sa[b,d,h]

[a]Biotechnology Center, Brazil; [b]Plant-Pest Molecular Interaction Laboratory (LIMPP), Brasilia, Brasília-DF, Brazil; [c]Departamento De Biologia Celular, Universidade De Brasília, Brasília-DF, Brazil; [d]Catholic University of Brasília, Brasília-DF, Brazil; [e]Viçosa University, UFV, Viçosa-MG, Brazil; [f]Department of Biological Sciences, National University of Medical Sciences, Punjab, Pakistan; [g]Embrapa Semiarid, Petrolina-PE, Brazil; [h]National Institute of Science and Technology, Jakarta Embrapa-Brazil; [i]INRAE, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

ABSTRACT

RNA interference (RNAi)-mediated gene silencing can be used to control specific insect pest populations. Unfortunately, the variable efficiency in the knockdown levels of target genes has narrowed the applicability of this technology to a few species. Here, we examine the current state of knowledge regarding the miRNA (micro RNA) and siRNA (small interfering RNA) pathways in insects and investigate the structural variability at key protein domains of the RNAi machinery. Our goal was to correlate domain variability with mechanisms affecting the gene silencing efficiency. To this end, the protein domains of 168 insect species, encompassing the orders Coleoptera, Diptera, Hemiptera, Hymenoptera, and Lepidoptera, were analysed using our pipeline, which takes advantage of meticulous structure-based sequence alignments. We used phylogenetic inference and the evolutionary rate coefficient (*K*) to outline the variability across domain regions and surfaces. Our results show that four domains, namely dsrm, Helicase, PAZ and Ribonuclease III, are the main contributors of protein variability in the RNAi machinery across different insect orders. We discuss the potential roles of these domains in regulating RNAi-mediated gene silencing and the role of loop regions in fine-tuning RNAi efficiency. Additionally, we identified several order-specific singularities which indicate that lepidopterans have evolved differently from other insect orders, possibly due to constant coevolution with plants and viruses. In conclusion, our results highlight several variability hotspots that deserve further investigation in order to improve the application of RNAi technology in the control of insect pests.

ARTICLE HISTORY
Received 27 July 2020
Revised 21 November 2020
Accepted 2 December 2020

KEYWORDS
*In silic*o analysis; structure-function relationship; dicer; drosha; pasha; argonaute; r2d2; loquacious; dsRBDs; protein evolution

1. Introduction

Even in the age of genome editing, the discovery of small non-coding RNAs (sncRNAs) represents one of the most exciting frontiers in molecular biology and biotechnology. Molecular pathways related to these molecules were first described in *Caenorhabditis elegans* [1,2], plants [3] and *Drosophila melanogaster* [4], with a focus on the regulation of gene expression and viral infections [5–8].

Specifically in insects, sncRNAs can be categorized into three main families based on their size and the RNA interference (RNAi) pathway that generates them: (i) *micro RNAs* (miRNAs), which are 22-nucleotide endogenous sncRNAs that participate in the regulation of gene expression via degradation or translational repression of mRNAs [9,10]; (ii) *small interfering RNAs* (siRNAs), which vary around 21 nucleotides in length and can be generated from either exogenous or endogenous double-stranded RNA (dsRNA) to counteract viral infections [11]; and finally (iii) *piwi-interacting RNAs* (piRNAs), which are sncRNAs spanning 25–31 nucleotides in length that interact with PIWI-related proteins and are required for processes ranging from the maintenance of germline stem cells in flies to retro-transposon silencing in eukaryotes [12,13]. For biotechnological purposes that target host-parasite interactions, miRNAs – and siRNAs-based approach es are the most widely adopted.

The characterization of the miRNA and siRNA pathways in *D. melanogaster* coupled with the mass sequencing of genomes and transcriptomes from several insect species have led to the wide use of the RNAi technology in the development of biotechnological resources aimed at controlling the populations of insect pests and virus vectors [14–17] (Fig. 1, *see* Supplementary Text ST1 – The miRNA and siRNA pathways in insects: An overview). However, the efficiency of RNAi knockdown is highly variable across insect orders, especially due to differences in the delivery, processing, and stability of sncRNAs. In the case of agriculture-driven RNAi-

CONTACT Maria Fatima Grossi-de-Sa (fatima.grossi@embrapa.br). Embrapa Genetic Resources and Biotechnology 02372, Brasília-DF, Brazil.

[§]Authors contributed equally to the work.

Supplemental data for this article can be accessed here.

RESUMO

A maquinaria de RNA de interferente (RNAi) nos insetos se distingue de outros metazoários pelo fato de que as vias de miRNA (*micro RNA*) e siRNA (*small interfering RNA*) são teoricamente independentes e cujos componentes são similares mas distintos. No presente estudo, foi avaliado o estado atual do conhecimento sobre estas duas vias em insetos através da análise filogenéticas e variabilidade, associadas com estudos com a estrutura linear e tridimensional de domínios conhecidos presentes em proteínas chaves destas rotas metabólicas.

Inicialmente, a seleção dos elementos relacionados ao mecanismo de miRNA e siRNA de insetos foi feita de acordo com estudos anteriores com a espécie modelo *Drosophila melanogaster*. Todas as análises *in silico* posteriores foram realizadas com sequências ortólogas às de *D. melanogaster*, identificadas em bancos de dados de 168 espécies de insetos, abrangendo as ordens Coleoptera, Diptera, Hemiptera, Hymenoptera e Lepidoptera (149 genomas e 20 transcriptomas). Acredita-se que todas as espécies possuam todos os elementos básicos das máquinas miRNA e siRNA, mas devido à baixa qualidade de um grande número de genomas disponíveis em banco de dados público, associados às limitações das metodologias de detecção de ortólogos, alguns elementos podem ter sido perdidos em algumas espécies.

Uma vez que a taxa de evolução (parâmetro K) é essencial para a avaliação da evolução molecular computacional e a análise de variabilidade, esse parâmetro foi calculado para as proteínas e domínios analisados neste estudo utilizando alinhamentos curados baseados na estrutura e em reconstruções filogenéticas por Máxima Verossimilhança. Estes dados forma input o algoritmo *Likelihood Estimation of Individual Site Rates* (LEISR), a fim de calcular a taxa de evolução diretamente dos dados de proteínas. A análise filogenética das oito proteínas completas revelou para as cinco ordens de insetos analisadas uma árvore da vida com padrão compatível com o proposto na literatura. A variabilidade e a distância filogenética podem ser evidências de que existem particularidades suficientes nos elementos da maquinaria de RNAi das espécies da ordem Lepidoptera que as diferenciam das demais.

Tais análises, associadas com estudos estruturais mostram que os domínios dsrm, PAZ, Plataforma, Ribonuclease III (RIID) e Helicase apresentaram a maior variabilidade, principalmente nas regiões de *loop* (sejam ordem-específicas ou não). A estrutura destes *loops* é difícil resolução e, quase sempre, não estão representados em modelos estruturais disponíveis em bancos de dados públicos. Mesmo com essas características, essas regiões influenciam

significativamente a atividade das proteínas nas quais estão inseridas e podem diminuir a afinidade do substrato, bloquear os sítios catalíticos ou até interagir com outras proteínas.

Além disso, a análise *in silico* dos domínios também mostrou uma variabilidade diferencial entre os elementos da via miRNA e siRNA analisados nas cinco ordens de insetos, destacando a ordem de Lepidoptera como a ordem mais distante quando comparada com as demais. Além disso, a variabilidade observada nos elementos siRNA é encontrada principalmente em domínios funcionais, o que não é amplamente observado nos elementos da rota de miRNA.

Outro fator importante a considerar é a variabilidade nos mecanismos de regulação transcricional dos elementos analisados. Com relação a este aspecto, os estudos *in silico* de sintenia aqui apresentados são importantes para ajudar a identificar se os mecanismos de regulação da expressão dos elementos analisados podem ser compartilhados entre espécies da mesma ordem, principalmente pela semelhança entre sequências que flanqueiam cada gene analisado, incluindo sua região promotora.

Finalmente, pode-se concluir que os estudos de variabilidade genética e estrutural de elementos da via RNAi em insetos mostraram, até certo ponto, como os mecanismo de desenvolvimento e resposta a infecções virais evoluíram nessas espécies e como podem tais informações podem ser aplicadas no controle de populações de insetos-praga, principalmente como ferramenta biotecnológica.

ABSTRACT

RNA interference (RNAi)-mediated gene silencing can be used to control specific insect pest populations. Unfortunately, the variable efficiency in the knockdown levels of target genes has narrowed the applicability of this technology to a few species. Here, we examine the current state of knowledge regarding the miRNA (micro-RNA) and siRNA (small interfering RNA) pathways in insects and investigate the structural variability at key protein domains of the RNAi machinery. Our goal was to correlate domain variability with mechanisms affecting the gene silencing efficiency. To this end, the protein domains of 168 insect species, encompassing the orders Coleoptera, Diptera, Hemiptera, Hymenoptera, and Lepidoptera, were analyzed using our pipeline, which takes advantage of meticulous structure-based sequence alignments. We used phylogenetic inference and the evolutionary rate coefficient (*K*) to outline the variability across domain regions and surfaces. Our results show that four domains, namely dsrm, Helicase, PAZ and Ribonuclease III, are the main contributors of protein variability in the RNAi machinery across different insect orders. We discuss the potential roles of these domains in regulating RNAi-mediated gene silencing and the role of loop regions in fine-tuning RNAi efficiency. Additionally, we identified several order-specific singularities which indicate that lepidopterans have evolved differently from other insect orders, possibly due to constant coevolution with plants and viruses. In conclusion, our results highlight several variability hotspots that deserve further investigation in order to improve the application of RNAi technology in the control of insect pests.

Keywords: *in sili*co analysis; structure-function relationship; Dicer; Drosha; Pasha; Argonaute; R2D2; Loquacious; dsRBDs; protein evolution.

INTRODUCTION

Even in the age of genome editing, the discovery of small non-coding RNAs (sncRNAs) represents one of the most exciting frontiers in molecular biology and biotechnology. Molecular pathways related to these molecules were first described in *Caenorhabditis elegans* (Fire et al., 1998; Olsen and Ambros, 1999), plants (Napoli et al., 1990) and *Drosophila melanogaster* (Kavi et al., 2008), with a focus on the regulation of gene expression and viral infections (Chow and Kagan, 2018; Leggewie and Schnettler, 2018; Li et al., 2002; Swevers et al., 2018).

Specifically in insects, sncRNAs can be categorized into three main families based on their size and the RNA interference (RNAi) pathway that generates them: (i) *micro RNAs* (miRNAs), which are 22-nucleotide endogenous sncRNAs that participate in the regulation of gene expression via degradation or translational repression of mRNAs (Mallory and Vaucheret, 2010; Sempere et al., 2004); (ii) *small interfering RNAs* (siRNAs), which vary around 21 nucleotides in length and can be generated from either exogenous or endogenous double-stranded RNA (dsRNA) to counteract viral infections (Okamura and Lai, 2008); and finally (iii) *piwi-interacting RNAs* (piRNAs), which are sncRNAs spanning 25-31 nucleotides in length that interact with PIWI-related proteins and are required for processes ranging from the maintenance of germline stem cells in flies to retro-transposon silencing in eukaryotes (Brennecke et al., 2007; Lin and Spradling, 1997). For biotechnological purposes that target host-parasite interactions, miRNAs- and siRNAs-based approaches are the most widely adopted.

The characterization of the miRNA and siRNA pathways in *D. melanogaster* coupled with the mass sequencing of genomes and transcriptomes from several insect species have led to the wide use of the RNAi technology in the development of biotechnological resources aimed at controlling the populations of insect pests and virus vectors (Airs and Bartholomay, 2017; Joga et al., 2016; Mamta and Rajam, 2017; Zhang et al., 2017a) (Figure 1, *see* Supplementary Text ST1 - The miRNA and siRNA pathways in insects: An overview). However, the efficiency of RNAi knockdown is highly variable across insect orders, especially due to differences in the delivery, processing, and stability of sncRNAs. In the case of agriculture-driven RNAi-based technologies, delivery can be achieved either through the use of transgenic plants expressing long dsRNAs, artificial miRNAs (amiRNAs), or through topical sncRNA administration (*e.g.*, naked, or nanoparticle-borne dsRNA/amiRNA) (Agrawal et al., 2015; Joga et al., 2016; Saini et al., 2018; Sharath-Chandra et al., 2019; Whitten, 2019; Yogindran and Rajam, 2016; Yu et al., 2013). The main disadvantage of transgenic plant-based approaches is that sncRNAs are

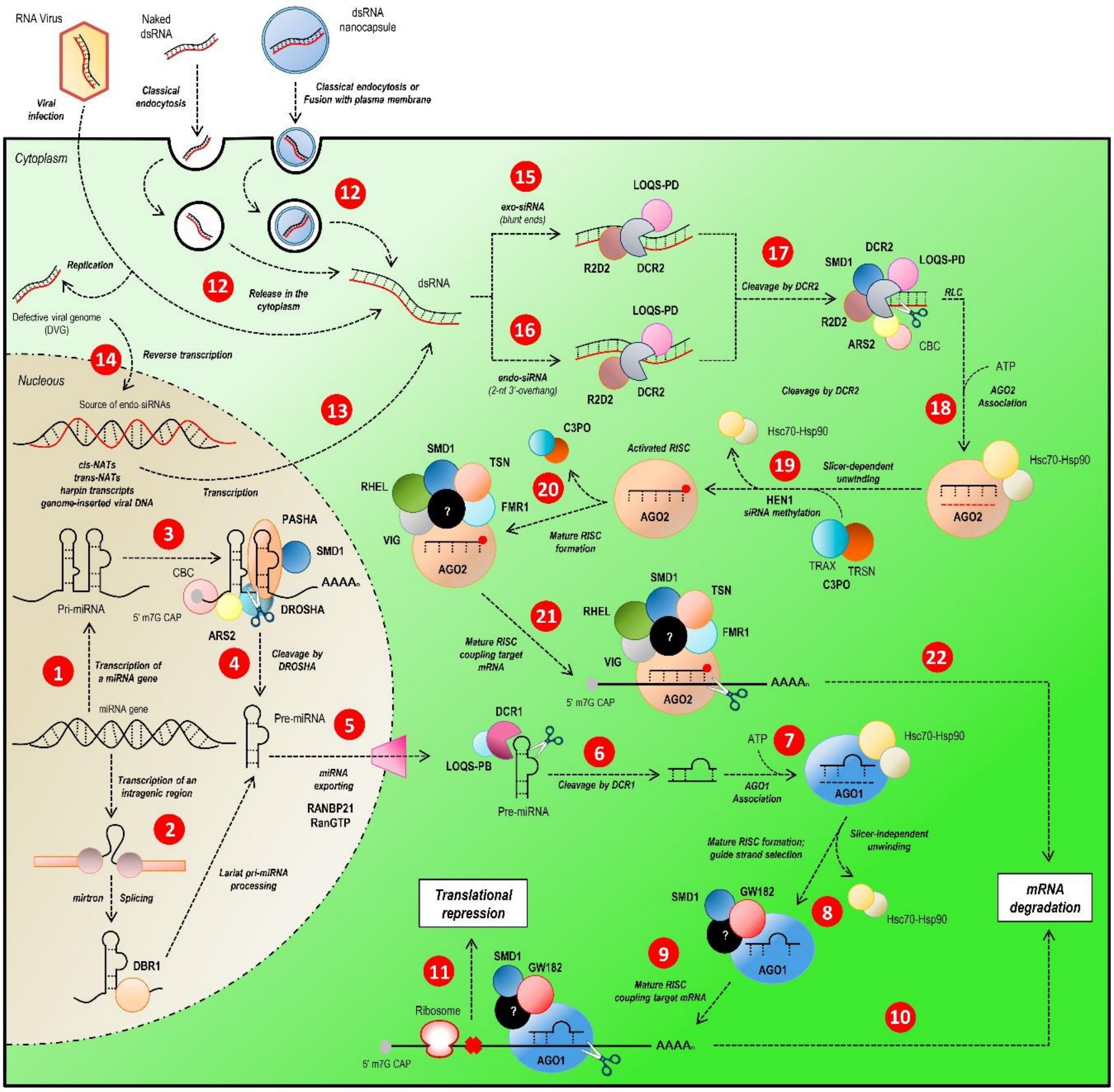

Figure 1. Overview of miRNA/siRNA gene silencing pathways in *D. melanogaster*. The sncRNAs can be categorized in three groups, according to their size and the processing-pathway they participate. The miRNAs (22 nucleotides) and siRNAs (21 nucleotides) display some differences in biogenesis follow independently processing pathway for gene silencing by translational repression or mRNA degradation. The miRNA biogenesis starts with the transcription of a primary transcript (pri-miRNA) with some structural peculiarities (hairpin loop domains, 5' cap and poly-A tail) (**step 1**). Intragenic regions can generate miRNAs; the loop present on spliceosome is recognized and processed by DBR1 (**step 2**), generating a pre-miRNA. The pri-miRNA loops are recognized by the DROSHA-PASHA complex associated with ARS2, CBC and SMD1, essential proteins in complex recruitment and pri-miRNA structural elements recognition (**step 3**). The pri-miRNA is cleaved by DROSHA (**step 4**) and the pre-miRNA exported to the cytoplasm by RANBP21 (**step 5**). In cytoplasm, the pre-miRNA is processed by DCR1 (**step 6**) in association with LOQS-PB and its loop is removed, generating a double-strand miRNA which is loaded on AGO1 (**step 7**), where one strand of miRNA duplex is selected as mature miRNA (**step 8**) and will constitute a mature RISC complex (**step 9**), which attaches to target mRNA directed by miRNA-mRNA base pairing, culminating in mRNA degradation (**step 10**) or translational repression (**step 11**). Unlike miRNA pathway, who biogenesis follows an endogenous-starting pathway, the siRNA starts, mainly, from an exogenous dsRNA source (as virus or some artificial source) or an endogenous-alternative source of dsRNA incorporated on host cell genome (**steps 12, 13 and 14**). According to the origin of dsRNA, it follows different, but remarkably similar, processing-pathways.

Figure 1. (cont.) The long exogenous dsRNA (exo-siRNA) is recognized by R2D2-DCR2 complex (**step 15**) and endo-siRNA is recognized by a complex of R2D2, DCR2 and LOQS-PD in association (**step 16**). Both siRNAs are cleaved by DCR2 stimulated by ARS2 and SMD1 (**step 17**) and associated with AGO2 (**step 18**). The selection of the guide-strand of mature siRNA is optimized by the association of AGO2 with the C3PO complex and its stabilization is acquired by siRNA methylation by HEN1 (**step 19**) until the mRNA target attachment. The mature RISC complex formation is dependent of the association of many proteins which enhance mRNA recognition and structure-changing, such as RHEL, SMD1, TSN and FMR1 (**step 20**). Once mRNA is attached to the mature RISC complex (**step 21**), the gene silencing is reached by mRNA degradation (**step 22**).

processed by the plant RNAi machinery prior to their delivery. For effective dsRNA uptake by insect cells, the optimal size of dsRNA ranges from 100-200 nucleotides; in contrast, after pre-processing by the plant's RNAi machinery, what remains for herbivorous insects are Argonaute-coupled single-stranded siRNAs and low levels of intact transgenic sncRNA, which jeopardizes efficient gene knockdown (Bally et al., 2018; Jin et al., 2015; Ulvila et al., 2006; Wang et al., 2019). This problem can be solved by the transgenic expression of sncRNA in plastids, such as chloroplasts. Chloroplasts are present in large numbers in plant cells (approximately 100 per leaf cell, depending on plant species) and display a compact genome that lacks classical elements of the RNAi machinery. Thus, sncRNAs expression in chloroplasts can provide high levels of intact transgenic sncRNA to the target insect population, thereby increasing the silencing efficiency (Bally et al., 2018; Jin et al., 2015). On the other hand, the technical complications related to non-transgenic RNAi-based approaches, such as the use of sncRNA nanocapsules, can be exemplified by the difficulty in choosing the best polymer for nanoparticle preparation. Delivery of sncRNA must be efficient while keeping the dsRNA molecule intact; in parallel, the production method must be low cost and adverse effects, such as high toxicity, must not be observed in non-target species.

Since sncRNA are mainly delivered to insects through nutrient absorption, the stability of exogenous sncRNAs in the insect midgut and hemolymph is another important factor that must be considered for successful gene knockdown. Several studies involving different species and insect orders have shown the presence of more than one nuclease isoform capable of degrading exogenous dsRNA (dsRNAses) in both the midgut and hemolymph (Almeida-Garcia et al., 2017; Liu et al., 2012; Peng et al., 2018; Prentice et al., 2019; Song et al., 2019; Spit et al., 2017; Wynant et al., 2014). These dsRNAses are highly stable (acting on acidic pHs) and do not present sequence specificity. In addition, transcriptional repression of these enzymes shows, in most cases, a considerable increase in the RNAi-mediated silencing efficiency of target insect populations (Almeida-Garcia et al., 2017; Liu et al., 2012; Peng et al., 2018; Prentice et al., 2019; Song et al., 2019; Spit et al., 2017; Wynant et al., 2014). Recently, a study involving lepidopteran species demonstrated the presence of a specific dsRNAse, REase, whose activity was associated with the low efficacy of RNAi-based gene silencing observed in this insect order (Guan et al., 2018a, 2018b).

A third factor to consider when evaluating gene silencing efficiency in insects is the uptake and transport of sncRNA across insect cells, the latter of which is a crucial feature of systematic RNAi. In *C. elegans*, such a process is mediated by the proteins SID1 and SID2 (Systemic RNA Interference-Deficient 1 and 2), which are transmembrane proteins responsible for binding and internalizing long sncRNAs; SID2 mediates tissue-specific endocytosis of exogenous sncRNA present in the intestine of *C. elegans*, whereas SID1 mediates vesicle release of sncRNAs into the cytoplasm and acts as a transmembrane channel that directly imports sncRNAs from tissues other than the intestine (McEwan et al., 2012; Winston et al., 2002). Even though the RNAi response as a cellular mechanism is highly conserved among eukaryotes, the systemic aspect of it is not. This situation can be observed among species of different insect orders, insofar as no orthologues of *C. elegans* SID2 protein have been identified, and possible orthologues of SID1 protein (SID1-like proteins; SIL) are generally associated with cholesterol transport rather than with sncRNA uptake (Méndez-Acevedo et al., 2017; Tomoyasu et al., 2008; Vélez and Fishilevich, 2018). Consistent with these observations, previous studies involving *D. melanogaster* and *Tribolium castaneum* have shown that exogenous sncRNA uptake in these two insect species occurs through the clathrin-dependent endocytosis pathway. Exogenous long sncRNAs are recognized by a membrane receptor (scavenger receptor) and later internalized into endosome vesicles, which in turn fuse tardily with lysosomes (Ulvila et al., 2006; Xiao et al., 2015). To become available to the RNAi machinery in the cytosol, the dsRNA needs to escape from the early-to-late endosomes before they fuse with lysosomal compartments (Dominska and Dykxhoorn, 2010). Problems during the release of sncRNA into the cytoplasm can lead to their accumulation in vesicles, which dramatically reduces the RNAi-mediated silencing efficiency, as observed in studies with *Spodoptera frugiperda* Sf9 cells (Yoon et al., 2017).

In light of the factors aforementioned, we hypothesized that the variability present within the core proteins of the insect RNAi machinery may also influence the success of RNAi-mediated gene silencing to control insect pests. Herein, we report a thorough *in silico* analysis of key proteins of the miRNA and siRNA pathways identified in genomes and transcriptomes from species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera). In particular, we focused on dissecting the sequence and structure variability present at the functional domains which compose the eight core proteins of the miRNA and siRNA pathways (AGO1-2, DCR1-2, DROSHA, LOQS, PASHA and R2D2). Given that proteins never function in isolation, and to put our analyses into context, we additionally present a compact and updated overview regarding the mechanisms of miRNA and siRNA biogenesis in the Supplementary Materials (Supplementary Text ST1 - The miRNA and siRNA pathways in insects: An overview). Our results identified several variability hotspots that might be associated to the different sensitivities to gene silencing mechanisms exhibited by insects. We found that all substantial variability hotspots can be mapped to loop regions within the functional domains of the RNAi core proteins (while milder variability is present in some of the secondary structural elements). We discuss the possible implications of the different locations and biochemical composition of these loops, as well as some of the idiosyncrasies pertaining to specific insect orders. Finally, our analysis revealed that some proteins that were thought to be lacking specific domains actually harbor them; furthermore, these domains appear to retain their canonical structures with very few exceptions that amount to loop regions.

METHODS

Database construction and phylogenetic analysis

The selection of proteins involved in insect miRNA and siRNA machinery was made according to previous studies with the model species *D. melanogaster*. The selection of 149 genomes and 20 transcriptomes (168 different species) belonging to the 5 insect orders analyzed in this study (Coleoptera, Diptera, Hemiptera, Hymenoptera, and Lepidoptera) was made according to the following parameters: (i) agronomic importance, including insect pests, as well as virus vectors; (ii) genomes and transcriptomes with a completeness value greater or equal than 95 % obtained by analysis with the *BUSCO* software (version 3; genome and protein modes; insect dataset odb9) (Waterhouse et al., 2018a); (iii) genomes with high N50 values. Model species with the most advanced genomes were chosen for each insect order and used as reference to search for orthologues in insects within the same order. The selected model species were: Coleoptera: *T. castaneum*; Diptera: *D. melanogaster*; Hemiptera: *Bemisia tabaci*; Hymenoptera: *Apis melífera* and Lepidoptera: *Manduca sexta*. Ortholog selection of the 8 selected proteins (AGO1-2, DCR1-2, DROSHA, LOQS, PASHA and R2D2) in genomes was made using the NCBI's *Basic Local Alignment Search Tool* for proteins (BLASTp; in BLAST package; version 2.8) (Altschul et al., 1997) with default parameters and *e*-value threshold of

10^{-5} through the *Best Bidirectional Hit* (BBH) methodology with modifications (Zhang and Leong, 2010). Due to the high level of duplication present in hexapod genomes (Li et al., 2018), we evaluated the best hit in BBH analysis in order to prevent the loss of orthologues (Dalquen and Dessimoz, 2013; Ward and Moreno-Hagelsieb, 2014). Regarding the transcriptomes, the initial search for orthologues was made with *tBLASTn* from the NCBI BLAST package (Gertz et al., 2006). Once the possible orthologues were selected, the open read frames (ORFs) were predicted for each transcript with the *ORF finder tool* (Rombel et al., 2002) and the correct ORF was selected and translated in the correct frame with the same tool. Thus, all subsequent phylogenetic and structural analyses were performed with the predicted protein sequences from all genomes and transcriptomes. All data concerning genomes and transcriptomes, and the ID of all selected sequences are summarized in Table S1. The protein sequences deduced from transcriptomes assembled in our lab (*Anthonomus grandis, Diatraea saccharalis, Hypothenemus hampei* and *Telchin licus licus*) are available in PDF format (Supplementary data). The protein sequences from other Metazoa phyla used for phylogenetic analysis (Figure 2; Chordata, Cnidaria, Nematoda and Platyhelminthes) were selected with the same BBH pipeline used for selection of insect sequences (*see* Table S2). In addition, the initially selected orthologues were quality-filtered according to the following criteria: (i) all selected protein sequences should start with methionine and their corresponding gene must end with a stop codon; (ii) The alignment coverage between the model species (query) and the target species (subject) should be greater or equal than 80 %. Subsequently, each selected protein was submitted for annotation of domains, which was performed locally using the *Hidden Markov Models* tool (HMMER; version 3.2) (Eddy, 2009) against the *Protein family* (Pfam) database (version 32.0 with 17,929 domain families) and default parameters, as well as the online platform *Simple Modular Architecture Research Tool* (SMART; version 8.0; http://smart.embl-heidelberg.de/) in normal mode including the option *Outlier homologues and homologues of known structure* (Letunic and Bork, 2018). Posteriorly, the protein domains limits were manually curated using multiple alignments and protein structures from the *Protein Data Bank* (PDB; https://www.rcsb.org). Prior to phylogenetic analysis, both complete proteins and their individual domains were aligned separately using the *MAFFT* software (version 7.402, via *Conda* repository) with -*auto* option and then manually curated (Katoh and Standley, 2013). Regarding protein domains, extremely discrepant sequences were removed from later analysis since they can represent errors in genome/transcriptome assemblies. Spurious sequences or poorly aligned regions identified from all multiple alignments from complete proteins and domains were removed with *trimAl* software (version 1.2) with -*gt* value equal to 0.5 (columns with gaps in at least 50 % of the sequences were eliminated) (Capella-Gutiérrez et al., 2009). With curated multiple alignments, the next step was the phylogenetic analysis itself using *Maximum Likelihood* method. The software used for such analyses was *Randomized Axelerated Maximum Likelihood* (RAxML; version 8.2.12) with options *-# autoMRE* (the software decided how many bootstrap replicates must be run) and *-m PROTGAMMAAUTO* (the fittest protein substitution model was selected by the software) (Stamatakis, 2014). The obtained phylogenetic trees were analyzed and annotated using the online tool *Interactive Tree of Life* (iTOL; version 4; *https://itol.embl.de/*), where all phylogenetic trees presented in this study are deposited (Letunic and Bork, 2019). The phylogenetic trees of the complete proteins (AGO1-2, DCR1-2, DROSHA, LOQS, PASHA and R2D2) are available as Supplementary material in TRE format.

Relative evolutionary rate inference

Site-wise relative evolutionary rates (*K*) are essential for computational molecular evolution and variability analysis. To investigate these evolutionary rates, the curated alignments and phylogenetic trees of complete proteins and individual domains were used as input for the program *Likelihood Estimation of Individual Site Rates* (LEISR), which is implemented in the software package *Hypothesis Testing Using Phylogenies* (HyPhy; version 2.5.1) and used for calculating the evolution rate directly from protein data (Spielman and Kosakovsky Pond, 2018; Sydykova et al., 2017). LEISR was run in protein mode with *LG* as the protein substitution model (Le and Gascuel, 2008) and four-category discrete gamma distribution to optimize branch lengths. The raw data was normalized with the average of all individual *K* values obtained for each site and box plots of the evolutionary rates were generated to assess the data distribution.

Sequence clusterization

Given that structure is much more conserved than sequence, modeling all proteins would implicate a redundant effort. To eliminate redundancy, proteins were repeatedly clustered using identity cutoffs; after every round of clusterization the largest sequence of each cluster was chosen as the representative of that cluster. We created a non-redundant dataset of sequences for each type of domain (*e.g.*, PAZ/PAZ-like), wherein the domain sequences within each dataset could have originated from different classes of proteins (*e.g.*, DCR1, DCR2, DROSHA, AGO1 and AGO2). Each of these datasets were first clustered using 95 % identity cut-off to eliminate near redundant domain sequences and then using 55 % identity as cut-off in the *CD-HIT suite web-server* (Huang et al., 2010); 55 % identity is considered a safe threshold to guarantee structure-function relationship between homologous proteins. Clusters containing only one sequence were regarded as outliers. If after these two clusterization steps the quantity of non-outlier clusters (those with two or more sequences) were bigger than 25 (square the number of insect orders evaluated), new rounds of clusterization were performed using continuously smaller identity cut-offs (in 5 % steps). Once the number of non-outlier clusters reduced to at most 25, clusters were manually verified. The representative sequence of clusters comprising non-redundant, non-outlier domain sequences from each insect order were selected for homology modeling and structural assessment.

Structure-based sequence alignment and homology modeling

The structure-based alignment of domains was performed in the following way: the representative cluster sequences were submitted to the *SAS* (Milburn et al., 1998), *LOMETS* (Wu and Zhang, 2007), *FFAS* (Jaroszewski et al., 2011), *GeneSilico* (Kurowski and Bujnicki, 2003), *MMseq2* (Mirdita et al., 2019) and *SEEKQUENCER* (*https://sysimm.org/seekquencer/*) servers with the purpose of finding templates for homology modeling. The most recurrent structures appearing in the results from these servers were selected as templates. The templates were structurally aligned using the sequence-independent mode of the *MaxCluster program* (*http://www.sbg.bio.ic.ac.uk/~maxcluster/index.html*) and also by means of the *POSA server* (Li et al., 2014). The superimposed structures outputted from *MaxCluster* and *POSA* were used to generate two refined structure-based *MSA* by employing the *STACATTO program* (Shatsky et al., 2006); sequence fragments that were not present in the structures were removed (*e.g.*, 6BUA had large portions of its sequence unresolved in the pdb file). We compared the structure-based sequence alignments originating from the superposition of both methods and, where divergent, manually selected the one that best captured our visual inspection of the superposed structures. Thus, at the end of this step, we were equipped with a curated structure-based sequence alignment of the template structures for each domain. The representative sequences of each domain were aligned to the curated structure-based *MSA* via the "*MAFFT – addfragments*" *algorithm* (Katoh and Frith, 2012) and an all-vs-all identity matrix was calculated using *UGENE* (Okonechnikov et al., 2012). The representative sequences were individually modeled using the template structure with which they shared the highest identity and at least 85 % coverage (when the latter condition was not satisfied, the highest coverage was used regardless of the identity); to this end, a pairwise target-template alignment was submitted as input to the *SWISSMODEL server* (Waterhouse et al., 2018b). The best quality model originating from the representative sequences of each domain were chosen for posterior structure analyses (*e.g.,* RNA-binding sites).

Multiple sequence alignments

The alignment of the remaining non-representative sequences from each domain (Figures S5-S32) were performed through two steps. First, we generated individual alignments for each group of insect order-protein-domain subunit using a combination of the *TCOFFEE* and *Probcons algorithm* in the *TCOFFEE server* (Armougom et al., 2006). For example, an individual alignment can encompass the sequences from the second RIIID subunit of DCR1 proteins from coleopterans, while another can encompass the first RIIID subunit of DCR1 proteins from coleopterans. This step is important to better align loop regions from each domain. The individual alignments were then sequentially merged with the parent alignment containing the template and representative sequences by means of the *MAFFT -merge algorithm* (Katoh and Frith, 2012). Given that the sequences have been previously clustered, every group of sequences within an individual alignment has at least one representative sequence in the parent alignment. Since the merge of an alignment can influence how the next one will be merged, the order in which the alignments were merged corresponded to their representative sequence's identity to the template structure. Thus, the alignment bearing sequences from the cluster with the highest identity to one of the template structures was added first, and then the alignment with highest average identity to the previously merged alignment was added next, and so forth. This hierarchical procedure guarantees a better alignment of loop regions by gradually decreasing the identity of groups of sequences. The canonical (Q, I, Ia, Ib, Ic, II, III, IV, IVa, V, Va and VI) and non-canonical (IVb) conserved-sequence motifs, important to ATP binding and hydrolysis, RNA binding, and in the communication between ATP and RNA binding sites were identified in Helicase domains using *MEME suite* (Bailey et al., 2009). All protein domain alignments are available as Supplementary material in FASTA format.

Statistical analysis

Statistical analyses of *K* values were performed using the median test for non-parametric data. To assess normality of the data, a *Kolmogorov Smirnov* test was performed beforehand (Wallot and Leonardi, 2018). All statistical tests were made by using the software *IBM SPSS Statistics*© version 25 (*https://www.dmss.com.br/produtos/statistics/statistics1.html*).

RESULTS & DISCUSSION

Phylogenetic overview of whole protein sequences

To identify potential sources of variability in the insect RNAi machinery, an *in silico* screening was performed through phylogenetic and structural analyses of both the complete proteins and their individual protein domains. Thus, a total of 1,211 sequences representing the core proteins of the insect siRNA and miRNA pathways were selected, namely the proteins AGO1, AGO2, DCR1, DCR2, DROSHA, LOQS, PASHA and R2D2. These proteins were chosen because they are directly associated with dsRNA processing and considerably influence the efficiency of RNAi-mediated gene silencing events, particularly those induced by environmentally introduced RNAs (environmental RNAi). Furthermore, many of the domains present in these proteins have at least one representative atomic structure deposited in the RCSB Protein Databank (Burley et al., 2019). This allowed us to produce structure models of homologous sequences and to map any variation to their three-dimensional context within the protein's structure. We identified representatives of all eight core proteins in species of the five insect orders we proposed to study: Coleoptera (*e.g.*, beetles), Diptera (*e.g.*, mosquitos and flies), Hemiptera (*e.g.*, cicadas and bugs), Hymenoptera (*e.g.*, bees and wasps) and Lepidoptera (*e.g.*, butterflies and moths). This verified that both pathways are ubiquitous in insects (Rubio et al., 2018). After the identification of orthologues by the BBH approach, the first important observation was the presence of putative paralogues of some of the core proteins in species of specific insect orders; specifically, we observed paralogues for AGO1 (in Hemiptera), AGO2 (in Diptera, Hemiptera, and Hymenoptera), LOQS (in Diptera, specifically in the Anopheles and Bactrocera genera) and PASHA (in Hemiptera) (Table S1).

Phylogenetic analysis of the eight complete proteins revealed topologies consistent with the insect tree-of-life proposed by Misof and coworkers (Misof et al., 2014) for the five insect orders analyzed (Figure 2A-D). Moreover, such an analysis also enabled us to evaluate the phylogenetic relationships between proteins that perform similar functions, mainly because

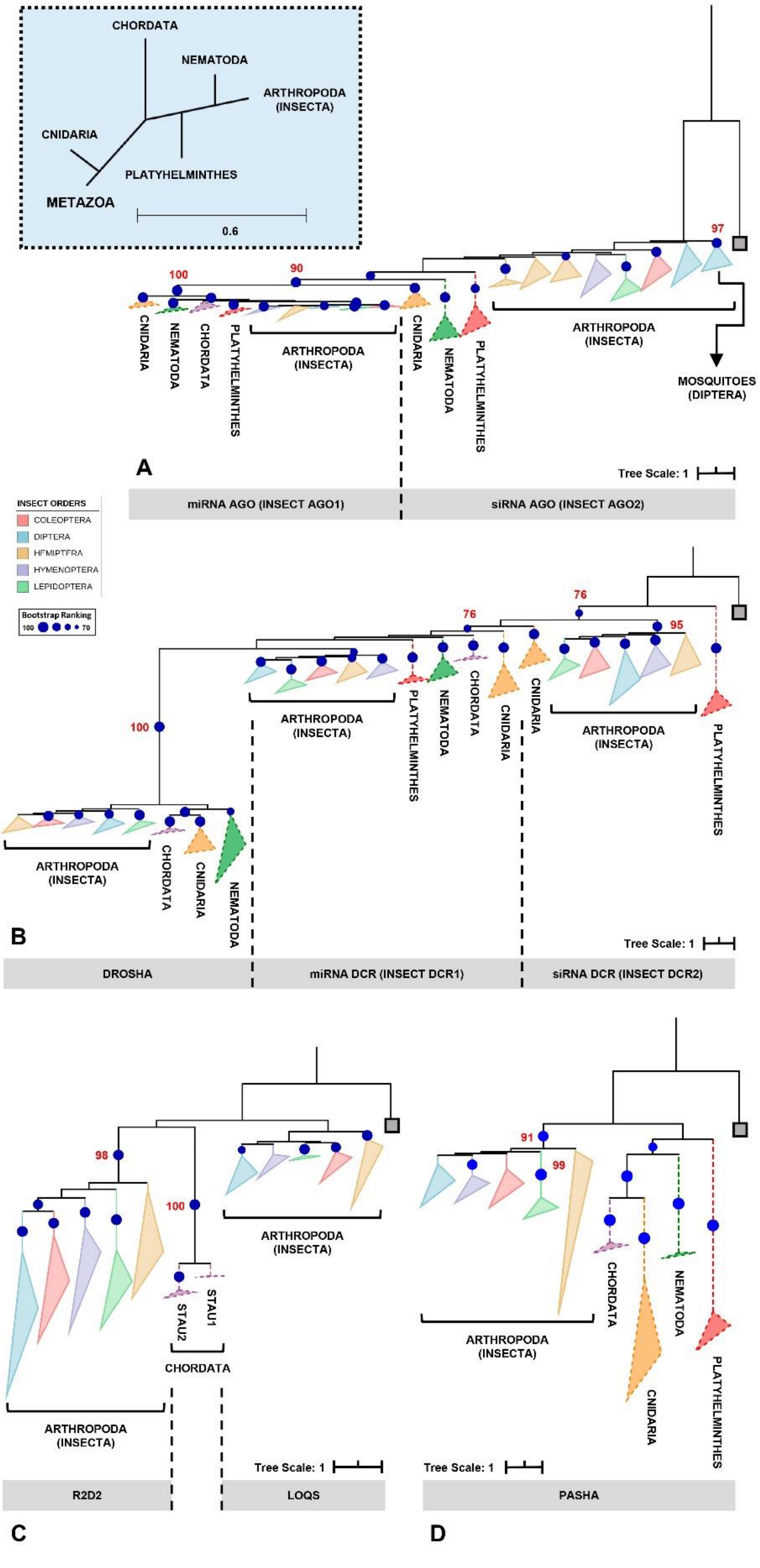

Figure 2. Phylogenetic analysis of the main RNAi machinery core elements in five different insect orders. (**A-D**) phylogenetic trees (Maximum Likelihood) showing the relationship among complete proteins from the basic core of miRNAs and siRNAs pathways in five insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera, represented by colored triangles – full lines).

Figure 2. (cont.) (**A**) AGO proteins; (**B**) RNAse III proteins (DCR1-2 and DROSHA); (**C**) DCR partners (LOQS and R2D2; dsrm-containing proteins); and (**D**) PASHA. The gray square on each phylogenetic tree represents the selected outgroup: (**A**) *Exiguobacterium* sp. ACQ71053.1 (bacteria); (**B**) *Batrachochytrium dendrobatidis* XP_006676691.1 (fungi); (**C**) *Homo sapiens* NP_599150.1 (TARBP2); and (**D**) *Rhodamnia argentea* XP_030526936.1 (plant). The cutoff value for bootstrap was 70 (represented by dark blue circles). The big blue square (dashed line) on the top represents the evolutionary relationship expected to each Metazoa phylum presented on the analysis. The dashed triangles represent the outgroup phyla (*purple* – Chordata; *orange* – Cnidaria; *green* – Nematoda; and *red* – Platyhelminthes). All phylogenetic tree files (.tre) can be found in Supplementary Section, as well as the selected species and the respective protein IDs (*see* Tables S1 and S2).

they share the same functional domains and probably the same ancestor. Four distinct maximum likelihood phylogenetic trees were produced for this purpose: (i) one including AGO1 and AGO2 proteins (Figure 2A); (ii) another comprising RNAse or RIIID-bearing endonucleases (DCR1-2 and DROSHA) (Figure 2B); (iii) the third consisting of insect-exclusive LOQS and R2D2, which are composed of double-stranded RNA-binding motif (dsrm) domains (Figure 2C); and (iv) the last consisting of DROSHA's partner protein, PASHA (Figure 2D). Insect AGO1 proteins formed a monophyletic group (bootstrap value: 100), with shorter branches and thus less variability than AGO2 proteins. The phylogenetic reconstruction of metazoan AGO proteins shown in Figure 2A corroborates previous phylogenetic studies that show two conserved AGO proteins between basal metazoans (represented here by cnidarians) and invertebrates (arthropods - insects, and nematodes), while Chordata phylum maintained only one type of AGO, closer to insect AGO1 (Wynant et al., 2017). Note that the Nematocera AGO2 (*e.g.*, species of the Aedes and Anopheles genera) clustered in a clade separate from the other dipterans (Figure 2A; bootstrap value: 97). This observation is extremely relevant in studies aimed at controlling the population of these viral vectors because of the "mutualistic" relationship between mosquitoes and viruses and the importance that the AGO2 protein has in the siRNA-mediated response to viral infection. RIIID endonucleases showed a characteristic pattern in which DCR1 and DROSHA proteins clustered in the same monophyletic clade, which was divided in two subclades, one for each protein class (bootstrap value for insect DROSHA clade: 100), whereas insect DCR2 proteins formed a separate monophyletic clade (bootstrap value: 95). These findings corroborate the hypothesis that DROSHA proteins may have evolved from the duplication of a common DCR ancestor and later specialized in the miRNA pathway (Cerutti and Casas-Mollano, 2006; Kwon et al., 2016; Moran et al., 2017). Overall, we observed that sequences of AGO1-2, DCR1-2 and DROSHA clustered in monophyletic groups according to their protein family rather than species (*e.g.*, one might have expected AGO1 and AGO2 sequences from the same species to be found in the same clade). This corroborates a canonical model of evolution in which the lineage-specific duplication of these proteins occurred, at least, before the speciation of insects (de Jong et al., 2009). However, robust support exists for a model in which the duplication of these genes occurred during deep metazoan diversification, concomitant with the origin of multicellularity and long before the diversification of the Arthropoda (Kosik, 2010; Mukherjee et al., 2013). Coupled with these analyses, the distribution of evolutionary rate (*K* value) for each protein family confirmed what was observed in the phylogenetic trees, wherein AGO1 orthologues showed the lowest variability among the eight core proteins ($p = 0.013$); in contrast, the AGO2 and DCR2 orthologues displayed the highest *K* values ($p = 0.031$ and 0.049, respectively) (Figure 3).

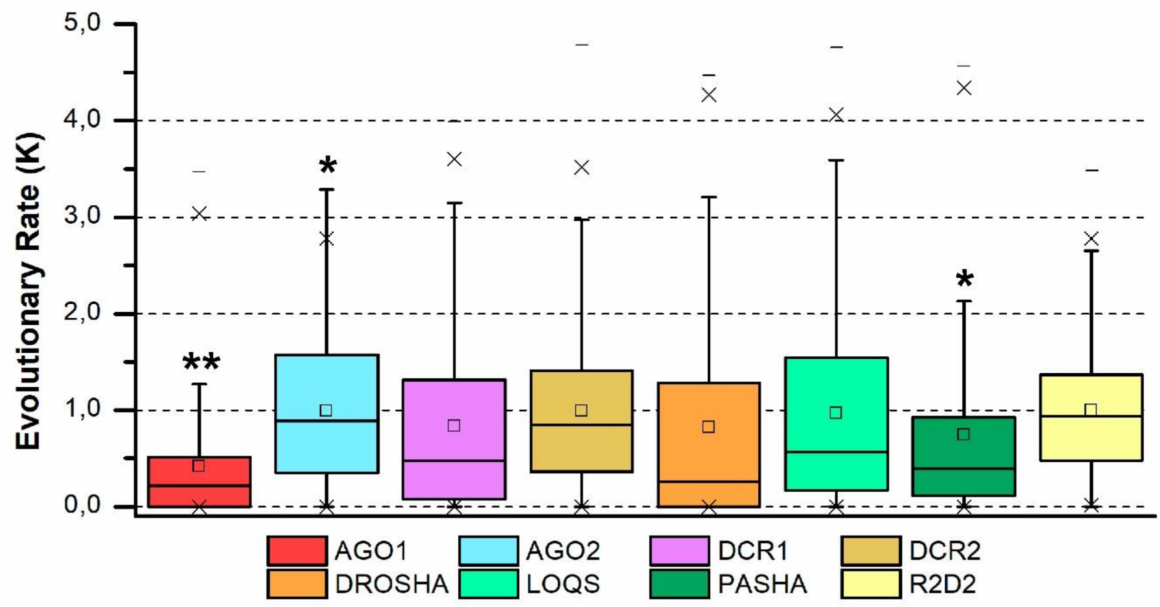

Figure 3. <u>Evolutionary rate evaluation of the main RNAi machinery core elements in five different insect orders.</u> The graph shows the distribution of the evolutionary rate (*K* value) in each alignment position for all protein classes analyzed. *Box plot interpretation:* The line in the middle of the box represents the *median* (mid-point of the data). Each part of the box divided by the median line represents 25 % of the data distribution. In this way, the box represents 50 % of the data. The unfiled small square inside the boxes represents the *average* value. The *whiskers* (upper and lower) represent scores outside of the 50 % represented by the box. The region delimited by each whisker until the limit of the box represents respectively 25 % (lower whisker) and 95 % (upper whisker) of the data. The dashes (-) at the ends represent the *maximum* and *minimum* values. The "exes" (x) represent outliers. The number of asterisks (*) indicates a statistically significant difference according to the non-parametric median test among insect orders (* $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$).

Among the protein families classified as double-stranded RNA binding proteins (dsRBPs), LOQS and R2D2, which are found exclusively in arthropods and considered essential for RNAi-mediated gene silencing in insects, appear to have evolved distinctly from other metazoan proteins of this class (Murphy et al., 2008). Our phylogenetic analysis (Figure 2C) showed both LOQS and R2D2 in different monophyletic clades (bootstrap value for insect R2D2 clade: 98), with R2D2 being more closely related to the Staufen proteins (STAU) of the Chordata phylum. Initially characterized in *D. melanogaster*, STAU proteins are widely distributed in several phyla in the Metazoa kingdom and can participate in both the transport and silencing of mRNAs, as well as in the control of their translation (St Johnston et al., 1992; Wickham et al., 1999).

Across most of the domains we analyzed, lepidopterans presented the highest phylogenetic distance compared to the other insect orders, especially in the analyses involving proteins of the siRNA machinery (AGO2, DCR2 and R2D2 proteins; Figure 2A-C). Specifically, regarding the high variability, and even absence, of R2D2 in the Lepidoptera (note the long branch in Figure 2C, R2D2 clade), some studies have suggested that the function of this protein may be carried out by LOQS in species of this order (Dowling et al., 2016). In summary, phylogenetic analyses of complete proteins showed highly conserved elements in the insect miRNA machinery when compared to the significantly more variable siRNA proteins. It is noteworthy that this variability is mainly observed across different insect orders but is remarkably reduced among species of the same order (Figure 3). This observation is important because most of the knowledge related to RNAi-mediated gene silencing in insects was initially obtained in studies involving *D. melanogaster* and later transferred to other insect species. Our analyses suggest that even though the primary domain functions are conserved within the miRNA and siRNA pathways, each insect order, or even species, may present idiosyncrasies that influence the RNAi-mediated gene silencing efficiency (*e.g.*, virus vectors). This premise is an important factor to be considered when RNAi is exploited as a biotechnological tool.

Upon observing variability between insect orders in our phylogenetic analyses, two questions need to be addressed: (i) are there "variability hotspots" within the sequences of each of the core RNAi proteins? and (ii) if so, is the hotspot region and its respective variability sufficient to cause structural and functional differences that could explain the RNAi efficiency/sensitivity in a given insect species? To answer these questions, it is important (and easier) to analyze the individual functional domains that make up the eight core proteins. Thus, we performed individual analyses of each domain by employing optimized structure-based sequence alignments, which are arguably more accurate than sequence-based alignments and also mitigate potential phylogenetic errors that may arise when examining the evolutionary history of said domains. Furthermore, structure-based sequence alignments allow us to use the calculated evolutionary rate of all sites in a domain's sequence to confidently pinpoint variability hotspots and conserved regions. The evolutionary rate of a given site informs us about the significance of the different amino acid substitutions at that position and allows direct comparison between other sites or regions (since the values are normalized). Thus, the detection of variability hotspots and, conversely, of slowly evolving sites is important for mapping functionally significant regions onto the three-dimensional structure of a domain; the structure, on the other hand, allows us to associate regions that are otherwise distant from each other at the sequence level but in close proximity within the three-dimensional and, therefore, functional context.

Domain architecture of core RNAi proteins

To analyze the intrinsic variability of each protein domain, our first step was to identify all known functional domains present in each of the eight core proteins of all 168 insect species. This step was initially achieved by annotating domains using HMM profiles from the Pfam database and then performing a data survey of protein structures deposited in the PDB that are involved in RNA interference. Bioinformatics analyses typically rely on the automatic annotation of domains using specialized databases, such as Pfam, CDD and SMART. While false positive hits are uncommon during these annotations, the same cannot be said about false negatives, these may result from indels, domain insertion, gene truncations or sequence saturation (excess of mutations) present in the query sequence. Notably, the atomic structures of proteins involved in miRNA biogenesis indicate the presence of domains that are not readily detected by automatic annotation databases, such as the Platform-PAZ-Connector domains within DROSHA (PDB ID: 5B16) and the Rhed and CTD domains in PASHA (PDB ID: 3LE4) (Kwon et al., 2016; Senturia et al., 2010). Even though structural data for some of these domains have been available for a while now, recent papers still fail to acknowledge them due to their reliance on automatic domain annotation servers (Davis-Vogel et al., 2018; Sharma and Mohanty, 2018). By thoroughly analyzing these protein structures, as well as reviewing their associated papers and comparing our results with the DASH database (Rozewicki et al., 2019), we were able to not only confidently expand the initial annotation using HMM profiles but also to define the precise boundaries of all annotated domains within each of our selected sequences. In total, 20 different domains were identified in the eight-core RNAi proteins: ArgoL1 (PF08699.8), ArgoL2 (PF16488.3), ArgoMid (PF16487.3), ArgoN (PF16486.3), Helicase domain (DEAD/ResIII; PF00270.27/PF04851.13, Hel2i, Helicase C; PF00271.29, and Pincer), Dicer Dimer (PF03368.12), Double-Stranded RNA-binding Motif (dsrm; PF00035.24), Piwi, Argonaute and Zwille (PAZ; PF02170.20, and PAZ-Like), P Element Induced Wimpy Testis (Piwi; PF02171.15), Ribonuclease III (RIIID; PF00636.24, and RIIID-like; PF14622.4), RNA-binding heme domain (Rhed), C-terminal domain (CTD), Platform, Connector and Staufen C-terminal domain (hereafter named Staufen C; PF16482.3) (Figure 4A and Figure S1-S4).

The analysis of *K* values for individual domains showed that those involved in the miRNA pathway presented lower *K* values than the ones involved in the siRNA pathway (Figures 4B-I). The AGO1 protein domains were those with the lowest *K* values (especially the PAZ domain; $p = 0.007$), while the domains of the AGO2 (*e.g.*, ArgoL2 and PAZ domains; $p = 0.038$ and $p = 0.041$, respectively) and DCR2 proteins (*e.g.*, Platform-Connector, RIIIDs and dsrm domains) exhibited significantly higher values ($p \leq 0.05$). Considering that the *K* values are directly proportional to the variability levels in our analyses, we can say that the protein domains from the siRNA pathway of lepidopteran species are the most permissive to mutations (Figures 5-12; Figures S2-S4).

Next, we further analyzed five protein domains whose functions are relevant to the biogenesis of sncRNAs and which presented regions with characteristic variability (high or low *K* values). The following domains were selected: (i) dsrm, which interacts with dsRNA molecules and is present in DCR1-2, DROSHA, LOQS, PASHA and R2D2 proteins (Burd and Dreyfuss, 1994; St Johnston et al., 1992); (ii) PAZ domain, which actively participates in the selection and correct orientation of miRNA/siRNA strands in AGO proteins and which is also crucial for the discrimination and length fidelity of substrates in DCR proteins (Cerutti et al., 2000; Hall, 2005; Kandasamy and Fukunaga, 2016); (iii) Platform domain, which recognizes the 5' phosphate moiety of dsRNA substrates and acts as a scaffold for the PAZ domain in DCR and DROSHA proteins (Kwon et al., 2016); (iv) RIIID domain, identified in DCR1-2 and DROSHA proteins, which displays exquisite cleavage specificity towards A-form dsRNA molecules (Blaszczyk et al., 2004, 2001; Conrad and Rauhut, 2002); and (v) Helicase domain, present in DCR proteins, which interacts with other RNAi-related proteins (*e.g.*, LOQS) in order to modulate the specificity of DCR2 for dsRNA substrates of the endo- or exo-siRNA pathways (Cenik et al., 2011; Sinha et al., 2018; Trettin et al., 2017; Ye et al., 2007).

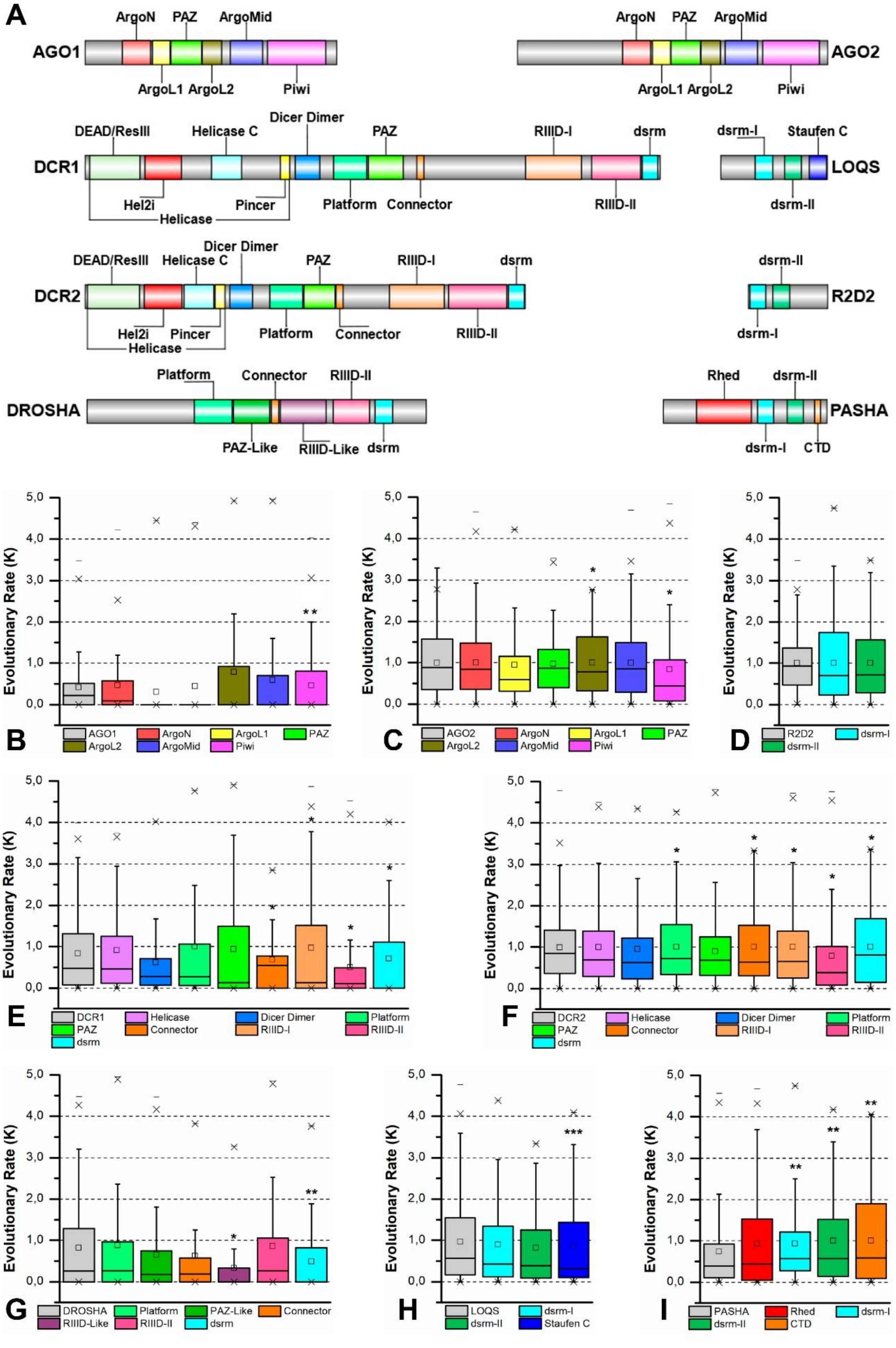

Figure 4. Protein domains from RNAi core proteins. (**A**) In-scale diagram of protein domains identified *in silico* in the classes of analyzed proteins. (**B-I**) Distribution of the evolutionary rate (*K* value) of each identified domain for all protein: (**B**) AGO1; (**C**) AGO2; (**D**) R2D2; (**E**) DCR1;

Figure 4. (cont.) (**F**) DCR2; (**G**) DROSHA; (**H**) LOQS and (**I**) PASHA. Asterisks (*) show statistical analysis of the data distribution of each domain compared to the complete protein (gray boxes). The number of asterisks (*) indicates statistically significant difference according to the non-parametric median test among insect orders (* $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$). *Box plot interpretation:* The line in the middle of the box represents the *median* (mid-point of the data). Each part of the box divided by the median line represents 25 % of the data distribution. In this way, the box represents 50 % of the data. The unfiled small square inside the boxes represents the *average* value. The *whiskers* (upper and lower) represents scores outside of the 50 % represented by the box. The region delimited by each whisker until the limit of the box represents respectively 25 % (lower whisker) and 95 % (upper whisker) of the data. The dashes (-) at the ends represent the *maximum* and *minimum* values. The “exes” (x) represent outliers.

Variability within dsrm and dsrm-like domains

We identified the canonical dsrm domain in most proteins and found it to be present in either one copy (DCR1-2, DROSHA and PASHA) or two copies (LOQS and R2D2) (Figure 4A; Figure 5). Due to its structural similarity (α-β-β-β-α topology), we classified the Dicer Dimer and Staufen C domains as *dsrm-like* domains, although previous studies have shown that they can interact with ssRNA and other proteins (such as DCR2) (Kurzynska-Kokorniak et al., 2016; Trettin et al., 2017). The dsrm domain yielded by far the highest *e*-values in our HMM-Pfam analysis, which demonstrates some sequence variability among the orthologues that have been annotated and deposited in public databases. This high variability may be the reason why several studies have failed to detect the C-terminal dsrm domain present in DCR1 proteins, even though it is highly conserved across insects (Figure 4A). Interestingly, the dsrm domains from different proteins of the miRNA machinery (DCR1, DROSHA, and PASHA) showed a highly conserved primary structure across all of insect orders we analyzed, especially when compared to the elements of the siRNA machinery (Figures 4D-I; Figure 5).

In general, despite exhibiting a conserved structure, we found that dsrm domains display a remarkable sequence variability in the loop between strands β1 and β2, a region that has been shown to directly interact with the dsRNA minor groove (Figures 6A-B) (Gan et al., 2006). We observed several amino acid substitutions at this site (Figures S5-S16), as well as several insertions of neutral and positively charged amino acids, mainly in species of the Anopheles genus and Lepidoptera order (Figures S10 and S13, respectively). The plasticity we observed for the β1-β2 loop (Figure 5E) may directly influence the interaction of these domains with dsRNA and consequently impact the efficiency/sensitivity of RNAi-mediated gene silencing.

The dsrm domains exhibit two different functions: they bind dsRNA molecules and/or facilitate protein-protein interactions, primarily in association with DCR, mammalian PKR or through the formation of dimers (Laraki et al., 2008; Wilson et al., 2015; Yang et al., 2010).

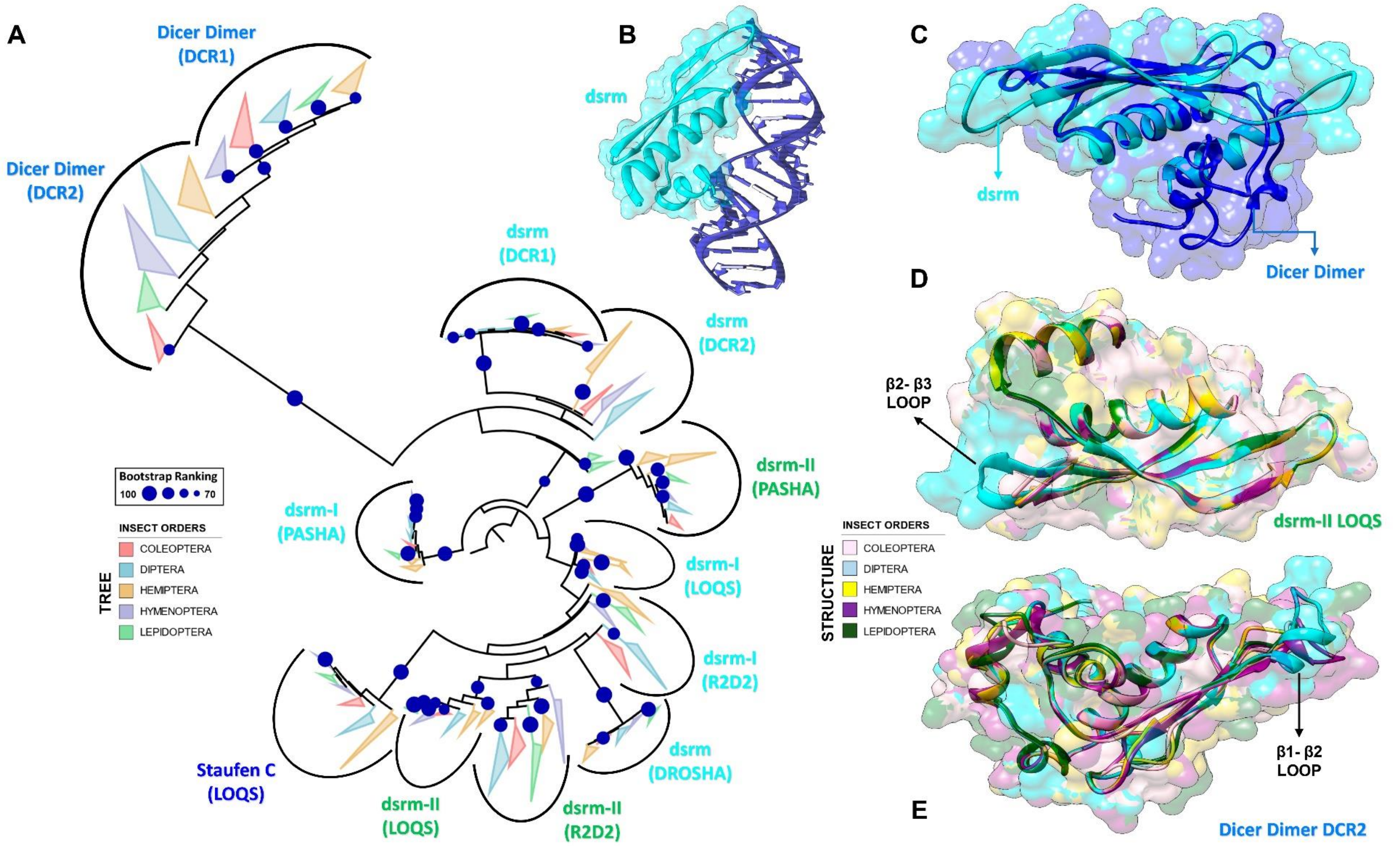

Figure 5. Structural and phylogenetic analysis of dsrm domains. (**A**) Maximum likelihood analysis including all domains with similar structure to dsrm present in the proteins DCR1, DCR2, DROSHA, LOQS, PASHA and R2D2 from species belonging to the five insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera). Dicer Dimer and Staufen C domains were inserted on this analysis due to have high structural

Figure 5. (cont.) similarity with dsrm. Each triangle represents an insect order, according to the color legend presented, and it is proportional to the number of branches present. The outgroup (hidden) used was the dsrm domain from human DROSHA (PDB ID: 5B16) and the bootstrap values are represented by dark blue circles (minimum 70). (**B**) Structural model of dsrm domain from human DROSHA (PDB ID: 5B16, B), interacting with RNA molecule, and (**C**) the same domain from human DROSHA superimposed with a Dicer Dimer from *Arabidopsis thaliana* DCL protein (PDB ID: 2KOU), highlighting the differences and similarities between these two domains. (**D** and **E**) Superposition of the models from LOQS dsrm-II and DCR2 Dicer Dimer domains, representing dsrm domains that hypothetically can interact preferentially with dsRNAs and proteins, respectively. In (**D**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC011666); **Diptera:** *D. melanogaster* (FBpp0080075); **Hemiptera:** *B. tabaci* (Bta01704); **Hymenoptera:** *A. melífera* (GB47214); and **Lepidoptera:** *M. sexta* (Msex2.00134). In (**E**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC001108); **Diptera:** *D. melanogaster* (FBpp0086061); **Hemiptera:** *B. tabaci* (Bta10685); **Hymenoptera:** *A. melífera* (GB48923); and **Lepidoptera:** *M. sexta* (Msex2.04462). In both (**D**) and (**E**) were highlighted the main variability spots.

According to our analysis, dsrm domains that bind to dsRNA (*e.g.*, those common to LOQS-PB and LOQS-PD) display contrasting variability hotspots compared to dsrm-like domains that are predicted to bind to proteins (*e.g.*, Dicer Dimer and Staufen C). While we found dsRNA-binding dsrms to accumulate most of their mutations in the β1 strand and β2-β3 loop (and marginally at the end of α2 helix) (Figures 6B and S13), protein-binding dsrm-like domains accumulate most mutations in the β1-β2 and β3-α2 loops (and marginally at the beginning of α1 helix) (Figures 6B and S16). The dsrm fold is highly conserved across animals and plants, and our observations corroborate previous studies, which show that dsrm-dsRNA interaction occurs primarily through two interfaces: (i) a canonical histidine, present on the β1-β2 loop, which inserts the dsRNA minor groove; and (ii) a cluster of basic residues at the beginning of α2, which stabilize the dsRNA backbone at an adjacent major groove (Ryter and Schultz, 1998; Vuković et al., 2014). Thus, it stands to reason that dsRNA-binding dsrms should not accumulate mutations in these regions, which would directly affect their capability to bind dsRNA molecules (stabilizing selection). In accordance with this reasoning, Dias and coworkers (2017) have shown that concerted amino acid substitutions in the dsrm β1-β2 loop and α2 region have been responsible for repeated gains and losses of dsRNA affinity during the evolution of animal and plant double-stranded RNA binding proteins (dsRBPs), and these regions are therefore considered "hotspots" for "tinkering" with dsrm-dsRNA interactions (Dias et al., 2017). Furthermore, the authors show that changes in dsrm-RNA affinity occurred often and could produce significant shifts in K_d through specific structural mechanisms: either by establishing/interfering with the critical histidine-RNA contact or by altering dsrm-dsRNA polar contacts within the β1-β2 loop and α2 region. Thus, if dsRNA-binding dsrms are to avoid these drastic shifts in affinity, as can be concluded from the low evolutionary rates we observed in these regions, it is likely that the β1-β2 loop and α2 region are under purifying selection. Conversely, protein-binding dsrms do not require the maintenance of dsRNA-binding residues (*e.g.*, histidine) in these hotspots and, accordingly, are able to accumulate many of the "tinkering" mutations reported by Dias (2017) without apparent fitness cost. It would seem that these amino acid substitutions are responsible for the domain's distinctive loss of dsRNA binding affinity relative to that of canonical double-stranded RNA binding domains (dsRBDs) (Dias et al., 2017). This observation raises the question of whether the same reasoning could be applied to putative protein-binding regions of dsrms; *i.e.*, will dsRNA-binding dsrms accumulate more mutations in protein-binding regions, as opposed to protein-binding dsrms displaying a purifying selection in the same regions? Hence, the contrasting pattern of evolutionary rates that we observed in the sequences of dsRNA- and protein-binding dsrms may provide us with a map for the identification of protein-binding interfaces in dsrms. Dias and colleagues (2017) pointed out that "although dsrms have been shown to directly mediate interactions with DCRs in animals and plants (Dias et al., 2017; Kurihara et al., 2006; Wilson et al., 2015), the extent to which dsrm-dsRNA and dsrm-protein binding may involve evolutionary trade-offs in specialization is not clear". It appears from our results that the "trade-offs" are significant despite different regions being involved with each type of interaction. The three-dimensional structure of dsrms shows that these regions are on opposite sides of the domain's long axis, which led us to propose a model wherein dsrm domains display two interaction-prone surfaces: one specialized in dsRNA recognition and another capable of binding proteins. The putative protein-binding surface (Figure 6A) is composed by the β1 strand, β2-β3 loop (including half of each β-strand) and the C-terminus of α2 helix (*e.g.*, DCR2's Dicer Dimer and LOQS' Staufen C domains; Figures S7 and S16, respectively); in some cases, the participation of β1 in protein binding appears to be relegated in preference to the α1-β1 loop (*e.g.*, DCR1's Dicer Dimer domain) (Figure S6). Nevertheless, we found that the β2-β3 loop contains a conserved (L/M)P$(X)_{2\text{-}3}$(S/C) motif in the Dicer Dimer and Staufen C domains of DCR1-2 and LOQS-PB, respectively (see alignment positions 39, 40 and 44 in Figure 6B). Considering these observations, we hypothesized that other dsrm domains might also share a similar pattern of accumulated mutations depending on whether they bind protein or dsRNA molecules. Accordingly, all other dsrm domains fell under the dsRNA-binding pattern, with the exception of the second dsrm subunit (dsrm-II) from PASHA. In this case, the prediction was slightly ambiguous, as mutations have accumulated in a large region that encompasses both the β1 strand and the β1-β2 loop (Figure S14); however, since most of the insect species retain the

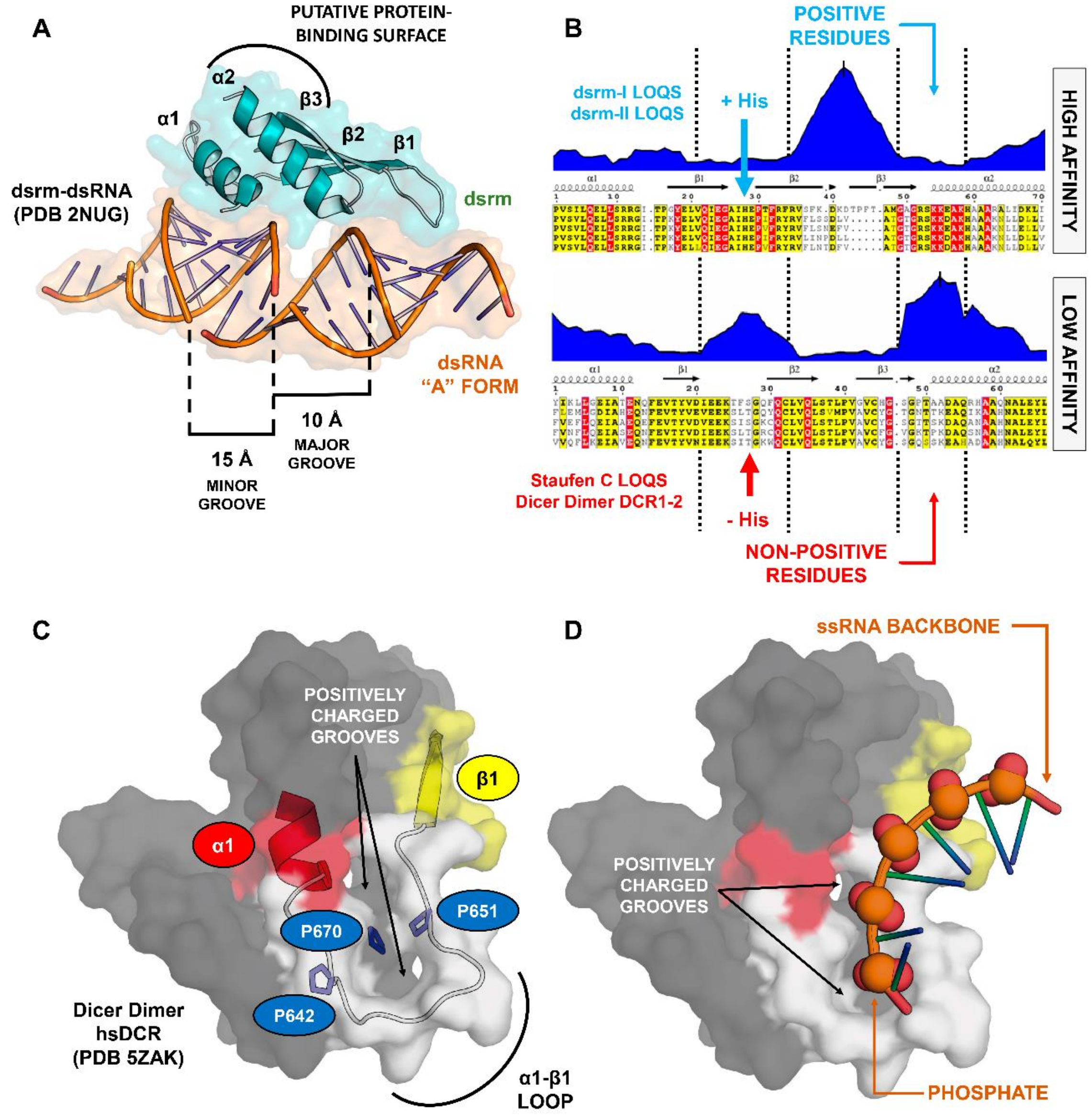

Figure 6. <u>RNA recognition by dsrm and dsrm-like domains.</u> (**A**) Canonical dsrm domains bind to one major groove and its two adjacent minor grooves by means of the β1-β2 hairpin and the N-terminal regions of helices α1 and α2. (**B**) The dsrm fold may present high or low affinity for dsRNA, depending on whether the conserved histidine and positively charged residues are present in the β1-β2 loop and α2 helix, respectively. Furthermore, protein-binding dsrms and dsRNA-binding dsrms display contrasting patterns of sequence conservation (*see* Figures S8 and S13 for complete alignment). (**C**) The α1-β1 loop of the Dicer Dimer domain from human Dicer (PDB ID: 5ZAK) forms two well-structured grooves which are separated by three proline residues; these proline residues are conserved in insect Dicer proteins. (**D**) Proposed model for the interaction of Dicer Dimer domains and ssRNA molecules. While the function of the two Dicer Dimer grooves are unknown, they present a positive electrostatic potential and are distanced such that two adjacent phosphate groups of a ssRNA backbone can be modeled to fit them (RNA template was retrieved from PDB ID: 4A36). This model was proposed to account for the Dicer Dimer's ability to bind single-stranded nucleic acids and promote base-pairing between complementary RNA/DNA molecules *in vitro* (Kurzynska-Kokorniak et al., 2016).

dsRNA-binding histidine residue in the β1-β2 loop and the positively charged residues in the N-terminus of α2 helix, we believe this dsrm domain may have higher affinity for dsRNA while also interacting with proteins via the β2-β3 loop and the C-terminus of helix α2. An extensive literature review allowed to confirm that our predictions for the Dicer-Dimer and Staufen C domains were, in fact, accurate. The Staufen C-like domain from human TRBP [a dsRBP that partners with human DCR (hsDCR) and is equivalent to LOQS-PD in Drosophila; PDB ID: 4WYQ] was shown to bind the helicase Hel2i domain via the β1 strand, β2-β3 loop and the C-terminus of α2 helix, all regions displaying low evolutionary rates and which we predicted to bind proteins (Figure S16) (Wilson et al., 2015). The cryo-EM reconstruction of hsDCR (PDB ID: 5ZAK) also enabled us to perform a comparative assessment of the Dicer Dimer protein-binding interface: it binds the junction between the RIIIDs and the Helicase domain mainly by means of its α1-β1 and β2-β3 loops, confirming our prediction and suggesting it shares functional similarity with its counterpart in Drosophila DCR1. However, we found the predicted binding of α2 was relegated in preference to the α3 helix (a unique feature of Dicer Dimer domains, which have an additional C-terminal extension containing two helices) (Liu et al., 2018). The Dicer Dimer has also been shown to bind single-stranded nucleic acids and to promote base-pairing between complementary RNA/DNA molecules *in vitro* (Kurzynska-Kokorniak et al., 2016). Thus, we also investigated whether the α1-β1 and β2-β3 loops from hsDCR could display other potential interaction surfaces. Strikingly, we found that the α1-β1 loop creates a flat surface on which two well-structured grooves are exposed (Figure 6C). These grooves are maintained and separated from each other through three conserved proline residues that are aligned in between them (*see* alignment positions 18, 27 and 47 in Figures S6 and S7). Both grooves are of sufficient size to accommodate phosphate anions, so we experimented modelling a single-stranded RNA (ssRNA) fragment onto the Dicer Dimer domain. The distance between the center of both grooves fits the exact distance between two adjacent phosphate oxygens of an A-form RNA backbone (Figure 6D). While this finding is very promising, it is still unclear whether our model can accurately predict the nature of dsrm binding partners (*i.e.*, either protein or nucleic acid) or even be extrapolated to dsrm domains outside the miRNA and siRNA pathways. Further investigations are needed to validate this model and effectively determine the structural interface of dsrm-dsrm, dsrm-protein and dsrm-ssRNA contacts.

Based on the study of Dias and coworkers (2017), we were also able to make predictions about the affinity of dsrm domains participating in the RNAi machinery. If a dsRNA-binding dsrm presented both the canonical histidine residue in β1-β2 and positively charged residues in α2, we categorized it as "high affinity"; accordingly, if a dsrm lacked both of these characteristics, we categorized it as "low affinity" (Figure 6B). We did not make assumptions about dsrms lacking just one of the characteristics, which boiled down to the two dsrms from R2D2 (Figures S10 and S15, respectively). Thus, the putative dsrm domains that we predicted to bind to dsRNA with high affinity were the dsrm II from PASHA (Figure S14) and dsrms I and II from LOQS (Figures S8 and S13, respectively), while those predicted to bind with low affinity were the dsrm I from PASHA (Figure S9) and the C-terminal dsrms from DROSHA, DCR1 and DCR2 (Figures S5, S11 and S12, respectively). In the case of DROSHA and DCR1, the presence of mismatches, small bulges and loops in the pri-miRNA and pre-miRNA substrates might explain the lack of high affinity residues in their dsrm domains; more importantly, it has been experimentally demonstrated that the C-terminal dsrm domain of DROSHA shows low affinity for dsRNA and that the insertion of LTLR(T/S)(M/V)(D/E) residues between α1 and β1 is important for this recognition (Figure S5) (Zhang et al., 2017b). As for DCR2 dsrm (Figure S12), the indication that it binds with low affinity to dsRNA is somewhat surprising; given its specialized role in antiviral RNAi, we would expect the C-terminal dsrm of DCR2 to bind dsRNA with high affinity, especially since we could not make affinity predictions on the dsrms of its partner protein, R2D2. While it might be the case that our prediction is entirely wrong, the lack of alternative highly conserved residues (Figure S12) in the three canonical RNA-binding regions (N-terminus of α1, β1-β2 loop and C-terminus of α2) further supports the low affinity binding of DCR2 dsrm to dsRNA.

Variability within PAZ and PAZ-like domains

The PAZ domains within proteins of the miRNA machinery (AGO1 and DCR1) displayed low variability between the insect species we analyzed (both *p* values lower than 0.05) (Figure 4 and 7; Figures S17-S21); however, we found that the PAZ-like domain from DROSHA contains a large insertion where the canonical β-hairpin module is predicted to be located (alignment positions 46-80; in DCR1-2, the β-hairpin is found between β2 and α1, while in AGO1-2 it is found between β3 and α3). The β-hairpin region is part of the 3'-pocket and interacts directly with the terminal 2-nt 3'-overhang via a conserved aromatic residue that establishes a π-stacking interaction between DCR proteins and the last nitrogenous base (Tian et al., 2014); this residue is classically a phenylalanine, which shows a preference for binding to U or G (Wilson et al., 2016). We found that phenylalanine can also be substituted by a

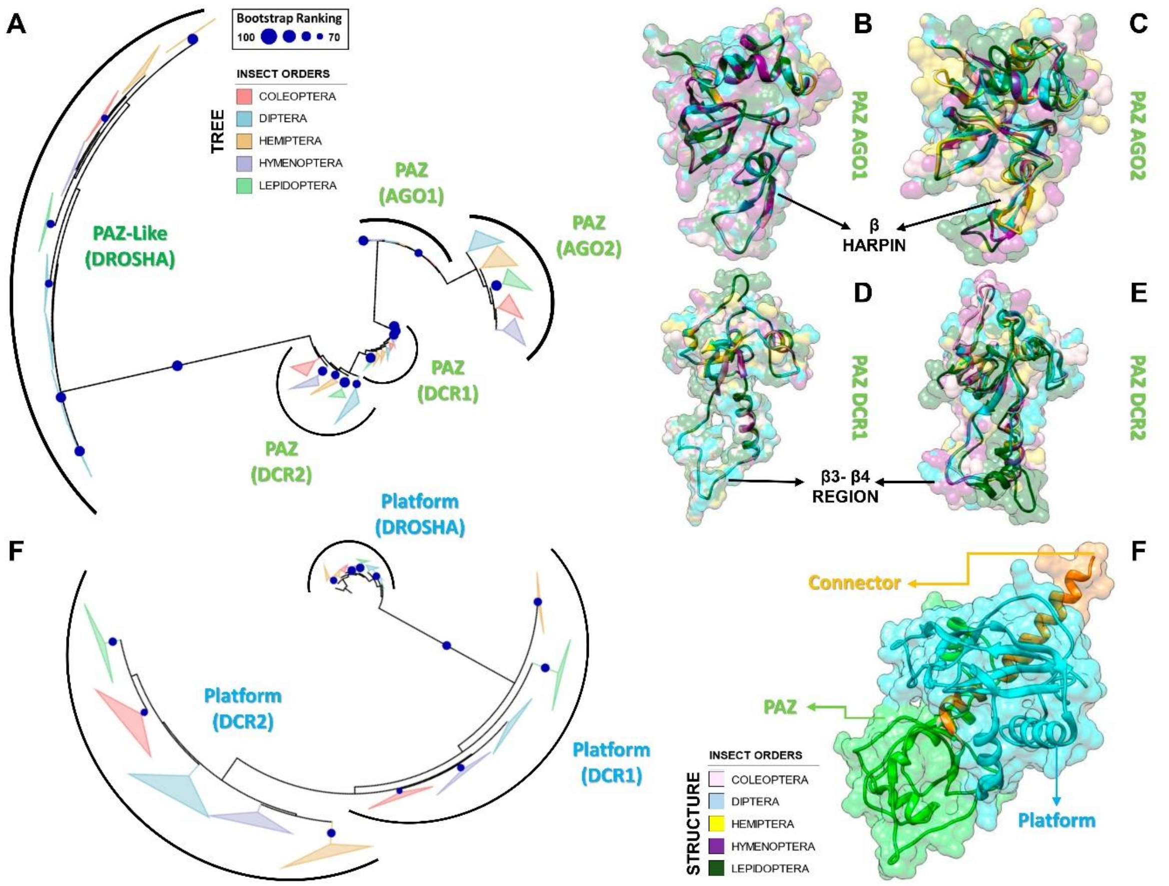

Figure 7. <u>Structural and phylogenetic analysis of PAZ and Platform domains.</u> (**A** and **B**) Maximum likelihood analysis of the PAZ domain presents in the proteins AGO1, AGO2, DCR1, DCR2 and DROSHA (PAZ-like) (**A**) and Platform (**B**) domain presents in the proteins DCR1, DCR2 and DROSHA, both from species belonging to the five insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera). Each triangle represents an insect order, according to the color legend presented, and it is proportional to the number of branches present. The outgroup (hidden) used to the PAZ domain tree was human DCR1 (PDB ID: 4NGD) and the Platform tree was human DROSHA (PDB ID: 5B16). The bootstrap values are represented by dark blue circles (minimum 70). (**B-F**) Superposition of the models from AGO and DCR PAZ domains, highlighting the main variability spots. No model was found for modeling the PAZ-like domain from DROSHA proteins. In (**B**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC005857); **Diptera:** *D. melanogaster* (FBpp0294043); **Hemiptera:** *B. tabaci* (Bta01840); **Hymenoptera:** *A. melífera* (GB48208); and **Lepidoptera:** *M. sexta* (Msex2.06997). In (**C**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC011525); **Diptera:** *D. melanogaster* (FBpp0075312); **Hemiptera:** *B. tabaci* (Bta00938); **Hymenoptera:** *A. melífera* (GB50955); and **Lepidoptera:** *M. sexta* (Msex2.05578). In (**D**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC001750); **Diptera:** *D. melanogaster* (FBpp0083717); **Hemiptera:** *B. tabaci* (Bta12886); **Hymenoptera:** *A. melífera* (GB44595); and **Lepidoptera:** *M. sexta* (Msex2.10734). In (**E**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC001108); **Diptera:** *D. melanogaster* (FBpp0086061); **Hemiptera:** *B. tabaci* (Bta10685); **Hymenoptera:** *A. melífera* (GB48923); and **Lepidoptera:** *M. sexta* (Msex2.04462). (**F**) Illustrative representation of Platform-PAZ-Connector domains from human DCR 5ZAK PDB model.

tyrosine or histidine, in the PAZ-like domain of DROSHA (alignment position 56 in Figure S21). Specifically, the 3'-pocket in DCR1-2 is composed of three main regions of the PAZ domain: (i) the loop between β1-β2 (β2-β3 in AGO1-2), (ii) the β-hairpin region + α1 (α3 in AGO1-2), and (iii) the β4 strand (β7 in AGO1-2) (Figures S19 and S20) (Tian et al., 2014). Remarkably, although we observed these regions might display increased evolutionary rates in both AGO and DCR proteins, they all retain the canonical residues (or similar) responsible for the recognition of the 2-nt 3'-overhang (YR-29, FP-53, F60, YY-64, KY-68, and QIL-125; *see* Figure S19, 4NGD sequence). This finding corroborates the notion that 3' dsRNA recognition is an ancestral characteristic of PAZ domains (Mukherjee et al., 2013). The PAZ domain may also participate in 5'-phosphate recognition together with the Platform domain (Tian et al., 2014). However, this characteristic is only observed in DCR proteins and is enabled due to a DCR-specific insertion between β3 and β4 (equivalent to β6 and β7 in the PAZ domain of AGO1-2; Figures S17 and S18). This insertion can form a dsRNA-interacting helix that is not critical for DCR processing but has been associated with the release and transfer of the cleaved dsRNA molecule into AGO proteins (Figure 8A) (Tian et al., 2014). In DCR2, we found that the PAZ residues that potentially interact with the 5'-phosphate (positions H85, S87, R89, and R96 of 4NGD sequence in Figures S19 and S20) display considerable variability when compared to DCR1, as illustrated by their contrasting evolutionary rates (Figure S20, the region between β3 and β4). This observation may reflect the fact that siRNA biogenesis in insects is mediated by the Helicase domain in DCR2, which preferentially recognizes long dsRNAs (≥ 38 bps) without the requirement of a specific 5' terminal structure (*i.e.*, it is permissive to blunt or 5'-non-monophosphorylated ends); in contrast, miRNA biogenesis is mediated by the PAZ domain in DCR1, which evolved to specifically recognize the 2 nt 3'-overhang and 5'-monophosphorylated ends of short dsRNAs (< 38 bps) (Fukunaga et al., 2014). Thus, while the DCR-specific insertion in the PAZ domain may mediate the release/transfer of the product in both DCR1 and DCR2 (Fukunaga et al., 2014), the conservation of key residues that we observed in DCR1 correlates with its role in the specific recognition of 5'-monophosphorylated ends, as exemplified by the "5' counting rule" observed during the pre-miRNA cleavage carried out by human and Drosophila DCR1 (Park et al., 2011).

Interestingly, *in vitro* studies have shown that the DCR2 PAZ domain of Drosophila species has regained the ability to specifically recognize the 5'-phosphate (Jia et al., 2017; Kandasamy and Fukunaga, 2016). We observed that this Drosophila domain bears mutations at sites adjacent to those typically participating in the 5'-phosphate recognition carried out by the

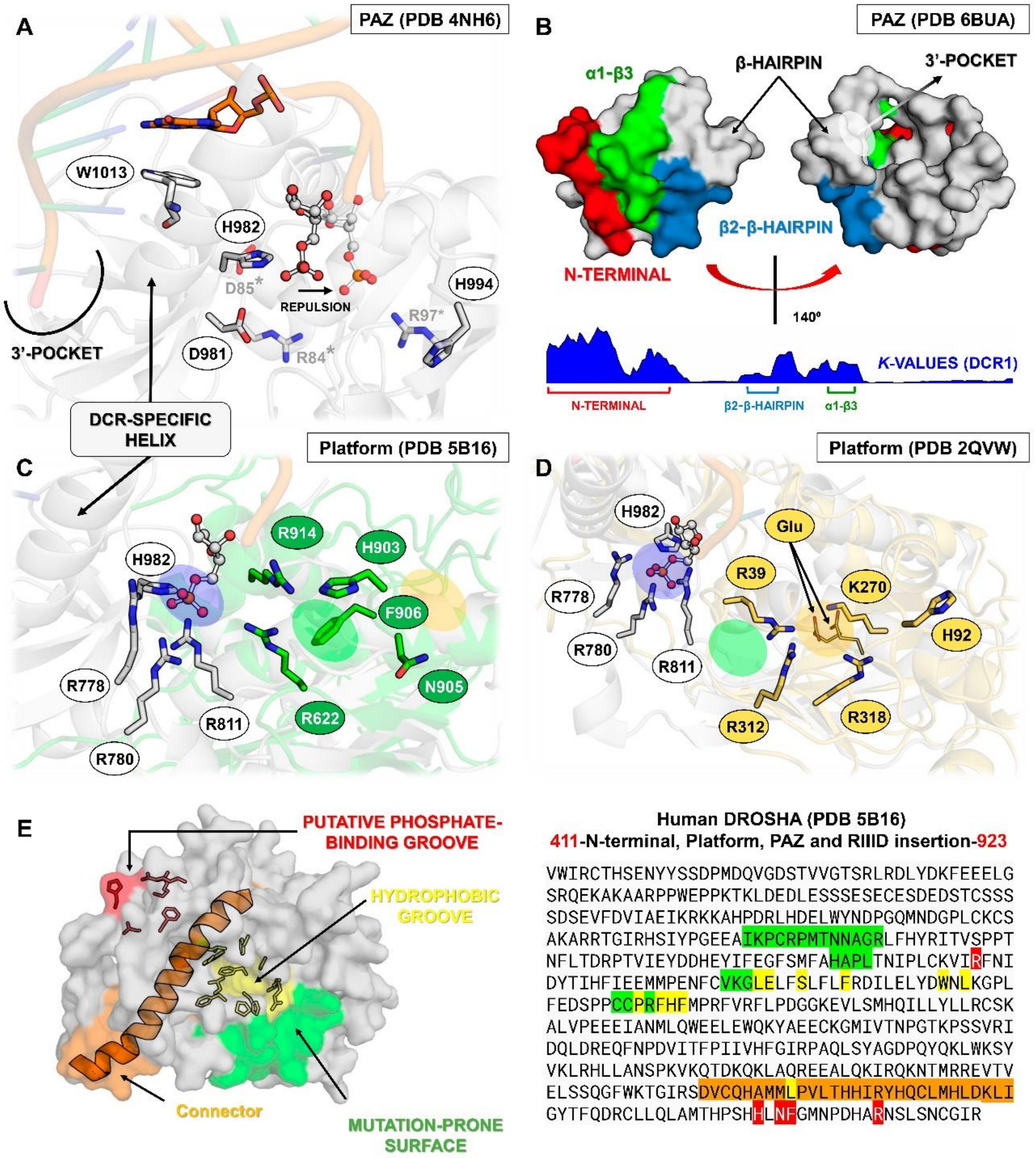
A
PAZ (PDB 4NH6)
W1013
H982
D85*
REPULSION
H994
R97*
3'-POCKET
D981
R84*
DCR-SPECIFIC HELIX
B
PAZ (PDB 6BUA)
α1-β3
β-HAIRPIN
3'-POCKET
β2-β-HAIRPIN
N-TERMINAL
140°
K-VALUES (DCR1)
N-TERMINAL
β2-β-HAIRPIN
α1-β3
C
Platform (PDB 5B16)
R914
H903
H982
F906
N905
R622
R778
R811
R780
D
Platform (PDB 2QVW)
H982
Glu
R39
K270
R778
H92
R780
R811
R312
R318
E
PUTATIVE PHOSPHATE-BINDING GROOVE
HYDROPHOBIC GROOVE
Connector
MUTATION-PRONE SURFACE
Human DROSHA (PDB 5B16)
411-N-terminal, Platform, PAZ and RIIID insertion-923
VWIRCTHSENYYSSDPMDQVGDSTVVGTSRLRDLYDKFEEELG
SRQEKAKAARPPWEPPKTKLDEDLESSSESECESDEDSTCSSS
SDSEVFDVIAEIKRKKAHPDRLHDELWYNDPGQMNDGPLCKCS
AKARRTGIRHSIYPGEEAIKPCRPMTNNAGRLFHYRITVSPPT
NFLTDRPTVIEYDDHEYIFEGFSMFAHAPLTNIPLCKVIRFNI
DYTIHFIEEMMPENFCVKGLELFSLFLFRDILELYDWNLKGPL
FEDSPPCCPRFHFMPRFVRFLPDGGKEVLSMHQILLYLLRCSK
ALVPEEEIANMLQWEELEWQKYAEECKGMIVTNPGTKPSSVRI
DQLDREQFNPDVITFPIIVHFGIRPAQLSYAGDPQYQKLWKSY
VKLRHLLANSPKVKQTDKQKLAQREEALQKIRQKNTMRREVTV
ELSSQGFWKTGIRSDVCQHAMMLPVLTHHIRYHQCLMHLDKLI
GYTFQDRCLLQLAMTHPSHHLNFGMNPDHARNSLSNCGIR

Figure 8. Variabilities within the PAZ and Platform domains. (**A**) Model for 5'-phosphate recognition in the DCR2 PAZ domain of *D. melanogaster*. Three residues were mutated in the template structure (PDB ID: 4NH6) to simulate the Drosophila PAZ domain's ability to recognize 5'-phosphate *in vitro* in DCR2. Drosophila species lack W1013 in DCR2; we speculate that substituting H982 for either Asp or Glu will repel the phosphate towards a putative phosphate-binding pocket formed by the Arthropod-specific and Drosophila-specific mutations D981R and H994R, respectively. We labelled with asterisk (*) the mutations according to their positions in the DCR2 PAZ domain alignment, shown in Figure S20. W1013 was only identified in DCR1 proteins and can be found at position 116 of Figure S19. (**B**) Our analyses of *K* values revealed that PAZ domains typically accumulate mutations in three segments that form a solvent-exposed flat surface on the three-dimensional structure of AGO, DCR and DROSHA proteins. A distinctive groove at the opposite face of this surface was observed, adjacent to the canonical 3'-overhang binding site of PAZ domains. Plants and lepidopterans display a distinctive positively-charged insertion in the N-terminal segment, suggesting their PAZ domains may bind RNA in a different orientation. (**C**) Comparison between the canonical phosphate-binding pocket of human DCR (blue ellipsis; PDB ID: 4NH6) and the putative phosphate-binding pocket we found in human DROSHA (green ellipsis; PDB ID: 5B16);

Figure 8. (cont.) this feature is also present in insects. Except for H982 (PAZ domain), all residues displayed in white color refer to the Platform domain of human DCR. The insect equivalents to R778, R780 and R811 can be found at positions 21, 23 and 54 in Figure S23, while the equivalent to H982 can be found at position 85 in Figure S20. Except for R622 (Platform domain), all residues displayed in green color refer to the DROSHA-specific insertion within the α2-α3 loop of the first Ribonuclease-III (RIIID) subunit of human DROSHA. The insect equivalents to R903, N905, F906 and R914 can be found at positions 15, 17, 18 and 26 in Figure S27, while the equivalent to R622 can be found at position 62 in Figure S24. The yellow ellipsis depicts the estimated location of *Giardia lamblia*'s putative phosphate-binding pocket. (**D**) Comparison between the canonical phosphate-binding pocket of human DCR (blue ellipsis; PDB ID: 4NH6) and the putative phosphate-binding pocket we found in *G. lamblia* DCR (glDCR; yellow ellipsis; PDB ID: 2QVW). The cavity forming the putative binding pocket is extremely well structured: two glutamate residues (E94 and E267 in glDCR) maintain four positively-charged residues coordinated around a central negatively-charged nucleus (R39, K270, R312 and R318). An additional histidine (H92 in glDCR) can potentially participate in the pocket insofar as E94 is repelled by an incoming phosphate. Except for R312 and R318 (RIIID-I subunit), all residues displayed in yellow color refer to the Platform domain of glDCR. The green ellipsis depicts the estimated location of human DROSHA's putative phosphate-binding pocket. Information regarding white-colored residues is described in **C**. (**E**) Depiction of important features we identified in DROSHA proteins. The hydrophobic residues that comprise most of the hydrophobic groove are clustered into a single segment (residues 645-681), which is also conserved in insect DCR1 and DCR2 proteins (positions 81-112 in Figures S22 and S23); however, lepidopteran DCR1 and plant Dicer-like (DCL) proteins differ by displaying distinctive positively-charged residues in this region. Similar to what we observed for the PAZ domain, several mutation-prone segments of the Platform domain sequence are common to the DCR1, DCR2 and DROSHA proteins. Furthermore, we observed that these common mutation-prone segments cluster on the three-dimensional structure of the Platform domain to form a contiguous surface. The nature of this mutation-prone surface is unclear.

DCR1 PAZ domain. We believe that these mutations might explain how 5'-phosphate recognition takes place *in vitro* in Drosophila DCR2. The aforementioned sites bearing these mutations can be seen in the sequence alignment of the DCR2 PAZ domain at position 84, which is conserved in all arthropods and adjacent to H85 of the 4NGD sequence, and position 97, which is conserved only in Drosophila and it is adjacent to R96 (Figure S20). Mutating both of these residues to alanine in DCR2 have been shown to block the *in vitro* cleavage of small dsRNAs (30-bp) bearing a 5'-monophophorilated end (Fukunaga et al., 2014); *in vivo*, however, this activity is inhibited by R2D2 and by physiological concentrations (25 mM) of inorganic phosphate (Cenik et al., 2011). Nevertheless, DCR2 from Drosophila species appear to be an exception rather than a rule with regard to 5'-phosphate recognition; first, only drosopholids display an arginine at position 97 (Figure S20); second, the ability to cleave small pre-miRNAs *in vitro* necessarily requires a phosphate at the 5' end, which differs from the activity of Drosophila DCR1 that can cleave both 5'-monophosphate and 5'-hydroxyl pre-miRNA substrates (containing 2 nt 3'-overhangs) (Fukunaga et al., 2014). We speculate that the mandatory requirement for 5'-monophosphate is likely the result of another Drosophila-specific mutation, (E/D)85 (Figure S20), which we argue is needed to repel the negatively-charged phosphate group and redirect it towards the slightly relocated phosphate pocket formed by R97 in Drosophila DCR2 (Figure 8A); in human and insect DCR1, the role of redirecting the 5' end towards the phosphate pocket is performed by a tryptophan or arginine residue present in the DCR-specific insertion within the PAZ domain (see position 116 in Figure S19), which stacks with one of the terminal nitrogenous bases via their indole or guanidino group and causes a bifurcation of the RNA double helix (Figure 8A) (Tian et al., 2014). We found that insect DCR2 lacks either of these residues (position 117, Figure S20). Furthermore, DCR1 requires a flexible (thermodynamically unstable) 5' terminus to efficiently bifurcate the dsRNA and recognize its 5' end (Park et al., 2011; Tian et al., 2014). Accordingly, the repulsion of 5'-monophosphate by E or D at position 85 could simulate a thermodynamically unstable terminus and allow the substrate to be accommodated in the 5' pocket (Figure 8A). Hence, novel structural mechanisms that nevertheless resemble the canonical 5'-phosphate-binding pocket of DCR1 may allow other arthropods to regain the ability of DCR2 to recognize 5'-phosphate specifically.

A general trend revealed by our analyses of evolutionary rates in PAZ domains is that its N-terminal region is highly variable independently of the protein, including DROSHA (Figures S17-S21); the N-terminal region ends at the first structural element (3_{10}-helix) in DCR proteins (equivalent to α1 in AGO1-2). Curiously, this region maps to a solvent-exposed flat surface composed by two other evolutionary-prone sequence segments (Figure 8B): the region between β2 and the β-hairpin (β3-β4 loop in AGO1-2) and the loop between α1 and β3 (α3-β6 loop in AGO1-2) (Figures S17-S21). Therefore, mutations appear to have accumulated within the same surface patch, indicating that this might be a variability hotspot for positive selection. Moreover, the PAZ-like domain from DROSHA harbors almost all of its variability in this surface region, although the putative α1-β3 loop is conserved (thereby creating a central conserved patch within the surface; *see* positions 90-98 in Figure S21). While the function of this surface is unclear, we observed it forms a distinctive groove at its opposite face, which suggests that PAZ-like domain can bind to specific moieties; this groove is also adjacent to the 3'-overhang binding site of the PAZ domain (Figure 8B). In accordance with our hypothesis, it has been shown that Dicer-like (DCL) proteins from plants harbor a lineage-specific insertion in the N-terminal region, which was responsible for an evolutionary increase in the affinity of the PAZ domain for RNA molecules (Jia et al., 2017); in DCL1, this insertion is longer and contains several positively-charged residues. Because of these observations, it has been proposed that plant DCLs may bind RNA in a different orientation than animal DCRs (Jia et al., 2017). This hypothesis is corroborated by the fact that DCL1 performs both pri-miRNA and pre-miRNA processing in plants, functions that are carried out separately in animals by DROSHA and

DCR1, respectively (Zhu, 2008). Curiously, we observed that lepidopteran species differ from all other insect orders by displaying a positively-charged insertion at the N-terminal region of their DCR1 PAZ domain, similar to the one found in plants (Figure S19). This raises the question of whether lepidopteran DCR1 may also binds to dsRNA in a different orientation, which might explain the different sensitivities to gene silencing mechanisms exhibited by this order of insects (Terenius et al., 2011). Alternatively, we hypothesize that the high evolutionary rates at the flat surface opposing the groove might allow the continuous selection of new potential species-specific partner proteins that reduce the free energy of the microprocessor complex (Figure 8B).

Variability within Platform domain

The Platform domain in insect DCR1 is important for the production of 22-nucleotide RNAs from double-stranded RNA precursors (miRNAs) by establishing the distance of the cleavage site from the 5’ end. In hsDCR, the interaction with the 5’ end of RNA molecules is mediated by a phosphate-binding pocket present in the region known as the Platform-PAZ-Connector cassette. Mutations in this pocket prevent correct miRNA biogenesis (Park et al., 2011). In accordance with our previous observation that the PAZ domain from DCR2 does not retain the canonical 5'-phosphate-binding residues, we also confirmed that the insect DCR2 Platform domain has a modified phosphate-binding pocket displaying sequence variability (Figure 7; Figures S22-S24; compare positions R21, R23, and R54 from the 5ZAK sequence in Figure S23). This further corroborates that the initial recognition of 5' end in dsRNA substrates is not performed by the Platform and PAZ domains in DCR2. Accordingly, DCR2 initially recognizes the dsRNA substrate via its Helicase domain, which threads the polynucleotide double-helix until it "hits" the PAZ and Platform domains at the opposite extremity of the microprocessor, thereby allowing the catalytic RIIIDs to proceed with processive cleavages in the transiently stabilized substrate (Lau et al., 2012). It should be noted that this model also predicts the possibility that the RIIID intradimer may cleave the substrate before it reaches the PAZ domain (generating fragments < 20 nt), which has indeed been demonstrated for DCR2 in *D. melanogaster* (Sinha et al., 2018). As for DROSHA, we found its 5'-pocket has been slightly relocated (~ 8.7 Å) in the template structure PDB ID: 5B16). While it bears in common with DCR1's phosphate-binding pocket the arginine residue between strands β4 and β5 (R62; in Figure S24, 5B16 sequence), the two arginine residues from loop β1-β2 have been relegated in preference of H15 and R26 from the DROSHA-specific insertion within the α2-α3 loop of the first RIIID subunit (Figures 8C and S27) (Kwon et al., 2016). The latter arginine residue is located in the so-called "Bump helix" and is conserved in all insect species investigated, while the histidine has been substituted by either an arginine or lysine residue (Figure S27). Additionally, a conserved asparagine and a phenylalanine are also found in the putative 5'-phosphate pocket (Figure 8C; NF-18 in Figure S27, 5B16 sequence). Until now, the recognition of the 5'-phosphate by DCR1 proteins has been regarded as a lineage-specific acquisition by metazoans (animals), largely due to the belief that DCR from *Giardia lamblia* (which is basal to metazoan DCRs) lacks much of the Platform and Connector domains and appears to only bind the 3' end of its RNA target (Jia et al., 2017; MacRae and Doudna, 2007). Contrary to this notion, we found that *G. lamblia* DCR (glDCR) displays most of the structural elements present in animal DCRs. Therefore, we hypothesized that the 5'-phospate pocket had not been identified previously because it could also be slightly relocated, resembling the one we found in DROSHA proteins. To investigate this issue, we have extracted the Platform domain from human DROSHA and superposed it onto the Platform domain of the full-length glDCR structure (PDB ID: 2QVW). Strikingly, we found a protuberant cavity in glDCR at approximately 7.1 Å from where we found the putative 5'-phosphate pocket in DROSHA (and at ~ 15.1 Å from the canonical DCR1 pocket; Figure 8D). Furthermore, we found this cavity to be extremely well structured: two glutamate residues (E94 and E267 in glDCR) maintain four positively-charged residues coordinated around a central negatively-charged nucleus (Figure 8D; R39, K270, R312 and R318). An additional histidine (H92 in glDCR) can potentially participate in the pocket insofar as E94 is repelled by an incoming phosphate. Interestingly, R39 is located between β4 and β5 strands of the Platform domain in glDCR, just like the conserved arginine residues within the 5'-phosphate pocket of human and insect DCR1 and DROSHA. Thus, our analyses suggest that this region's role in binding phosphate is likely more ancestral than previously reported (Jia et al., 2017). Noteworthy, we also found unique similarities between the putative glDCR and metazoan DROSHA 5'-phosphate-binding pockets, such as the participation of residues from the α2-α3 loop of the first RIIID subunit (R312 and R318); the RIIID loop in glDCR is intermediate in length to the DROSHA-specific insertion and the short loop found in metazoan RIIIDs. This implies that either DROSHA is evolutionarily closer to the ancestral eukaryote DCR than both DCR1 and DCR2 or that DROSHA acquired this characteristic independently and represents a potential case of molecular-evolutionary convergence. It should be noted that we also looked for an alternative 5'-phosphate-binding pocket in the Platform domain of DCR2 proteins by plotting conserved residues onto the structure of *D. melanogaster* DCR2 (PDB ID: 6BUA) and analyzing its surface. However, we did not find any alternative cluster of positively-charged residues and our investigation indicates that insect DCR2 has a degenerate 5'-phophate-binding pocket arranged in similar position to the one found in DCR1 proteins (Figures S22 and S23). In agreement with our observation, it has been shown that mutating DCR2 by reintroducing residues present in the 5' pocket of DCR1 Platform domain (*e.g.*, R21, R23, and R54 of 5ZAK sequence; Figure S23) can rescue high-affinity binding of DCR2 to 5'-phosphate (Jia et al., 2017).

A general trend we identified in the Platform domains from DCR and DROSHA is the presence of four common variability hotspots, which form an extensive surface adjacent to a pronounced hydrophobic groove (Figure 8E). Considering the structure of DROSHA, the regions that comprise this surface are the following: the N-terminal tail (first 12 residues of the domain), the β3-β4 loop (equivalent to β2-β3 loop in DCR1-2), the N-terminal half of α1 helix, and the loop preceding β6 (loop pre-β6) (Figure S22-S24). The loop pre-β6 is very flexible and it is located nearest to the hydrophobic groove, which is formed by residues LE-86, S89, F93, W102, L104, P117, FHF-121, and L863 (*see* 5B16 sequence in Figure S24; L863 is not depicted in the alignment and it is part of the Connector helix in the same PDB 5B16). The nature of this hydrophobic groove is unclear, but it is positioned symmetrically opposite to the 5'-phosphate pocket in the long axis of the Connector helix, resembling a mirror image (Figure 8E). All of the residues forming the hydrophobic groove, except for L863, are concentrated on the segment straddling the C-terminal half of α1 to the N-terminal half of β6 (Figures 8E and S24). The hydrophobic residues in this segment are also conserved in insect DCR1-2 proteins (positions 81-112 in Figures S22 and S23). Intriguingly, this region contains a unique insertion in plant DCL proteins and has been specifically pinpointed, alongside an insertion in the PAZ domain, as primarily responsible for increasing the affinity of the Platform domain for RNA molecules in DCLs. In particular, the plant-specific insertion in the Platform domain is rich in positively-charged residues and has been proposed to bind to the 5'-phosphate (Jia et al., 2017). Thus, the hydrophobic groove that we found in animal DROSHA and DCR proteins may turn out to be completely remodeled with positive charges in plant DCL proteins. Additionally, the remodeled groove is positioned on the same face as the plant-specific insertion in the PAZ domain, which also forms a distinctive groove. We previously mentioned that lepidopteran species also harbor a positively-charged insertion in the DCR1 PAZ domain, similar to the insertion found in plant DCL1. While the same is not true regarding the presence of a Platform insertion in the α1-β5 segment (which forms the hydrophobic groove), we found that the DCR1 Platform domain from lepidopteran species also displays distinctive positively-charged residues in this region, which largely contrasts with what we observed in species from all other insect orders (Figure S22). Altogether, it is tempting to speculate that DCR1 from lepidopterans is capable of binding RNA substrates in a similar fashion as plant DCL1, which may involve recognizing nucleic acids in a different orientation than that found in other animal DCR1 proteins. The implications of this idiosyncrasy, however, are unclear, especially since DCR1 from lepidopterans also retains the conserved residues that form the canonical 5'-phosphate and 3'-overhang pockets in the Platform and PAZ domains, respectively. Since plant DCL1 can process both pri-miRNA and pre-miRNA substrates (Zhu, 2008), it is perhaps the case that Lepidoptera DCR1s can also bind to two different substrates. This matter requires further investigation.

Variability within RIIID and RIIID-like domains

Two copies of the RIIID domain (RIIID-I and RIIID-II) have been identified in DCR1-2 and DROSHA proteins, wherein each one acts as a different subunit capable of cleaving one of the dsRNA strands (Figures 4 and 9; Figures S25-S30). Analyses of crystallographic structures have revealed that the canonical topology is composed of 7 α-helices (Figures S25-S30). In DCR1-2 and DROSHA, the second RIIID subunit displays the canonical 7-helix structure, while the first subunit lacks the α1 helix, which is instead surrogated by the C-terminal end of the Connector helix (Figure 10A). Apart from this peculiarity, all other secondary structural elements of RIIIDs from DCR1-2 and DROSHA superimpose well to each other and maintain a well-defined hydrophobic core (RMSD = 0.58 Å; Figure 10A). Conversely, the loops between helices α2-α3 and α5-α6 show remarkable variation in size and sequence identity (Figures S25-S30); for example, DROSHA displays a distinct RIIID-I subunit (known as the RIIID-like domain), which bears a large insertion between the α2 and α3 helices (Figure 9D; Figure S27). Both loops are located less than six residues from the first catalytic residues of helices α3 (positions E514 and D55 in Figure S29) and α6 (positions D156 and E159 in Figure S29). Thus, it appears that these regions may play pivotal roles in the catalytic mechanism of proteins harboring RIIID domains. In *Homo sapiens* DCR (hsDCR), the α5-α6 loop from RIIID-I has been identified as a minimal binding site for the interaction with human AGO proteins, *i.e.*, the polypeptide comprising only α5-α6 loop from hsDCR was able to interact with all members of human Argonaute proteins (Sasaki and Shimizu, 2007). Furthermore, the α5-α6 loop sequence was shown to be highly conserved among vertebrate DCR proteins but appears to have significantly changed during the evolution of their non-vertebrate orthologues (Sasaki and Shimizu, 2007). In agreement with these findings, we observed that the insect loops are shorter

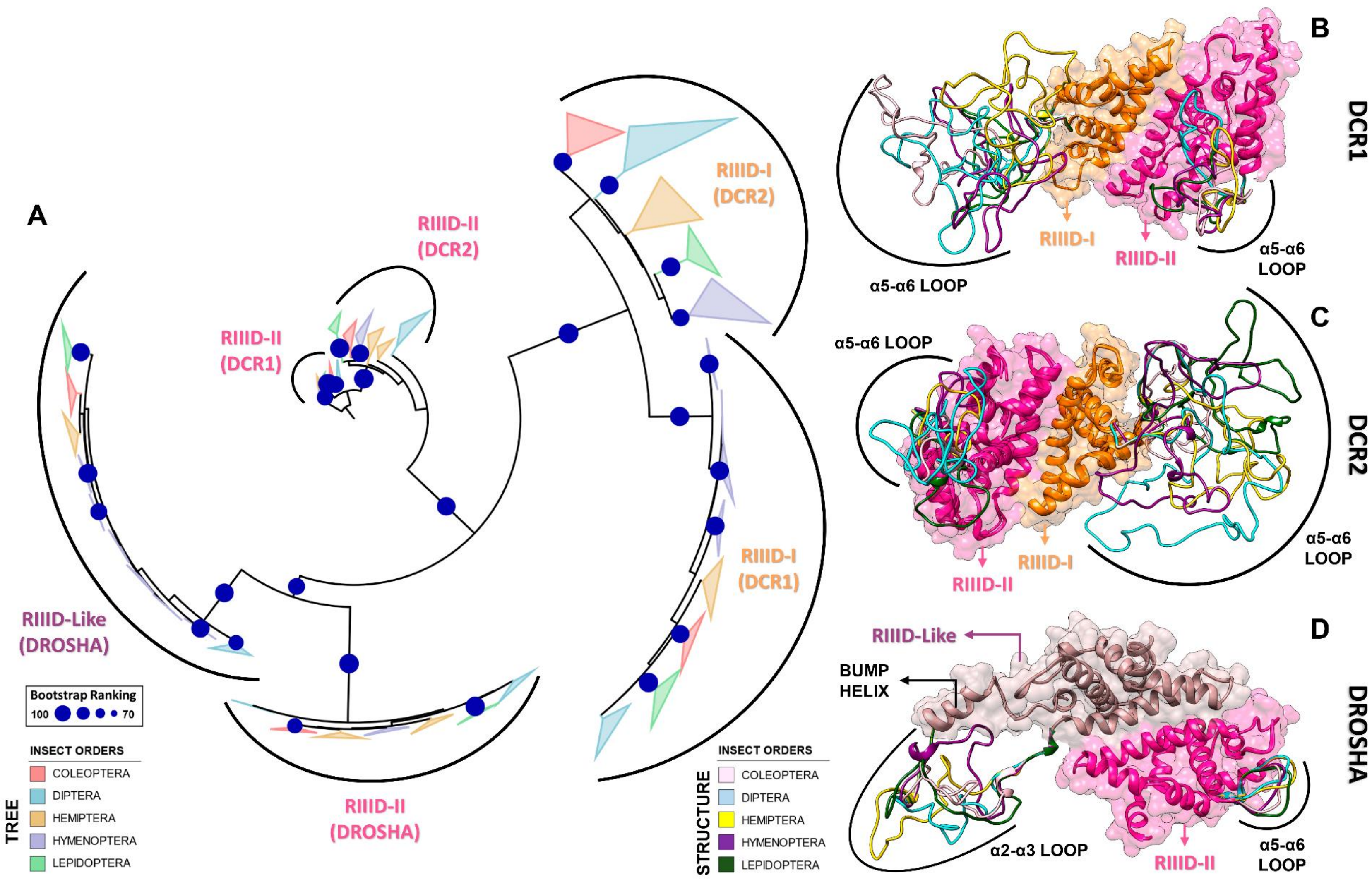

Figure 9. Structural and phylogenetic analysis of Ribonuclease III domain. (**A**) Maximum likelihood analysis of the two subunits (I and II) of Ribonuclease III domain (RIIID) present in the proteins DCR1, DCR2 and DROSHA from species belonging to the five insect orders (Coleoptera,

Figure 9. (cont.) Diptera, Hemiptera, Hymenoptera and Lepidoptera). The first subunit found in the DROSHA protein differs from the others, being then called RIIID-like. Each triangle represents an insect order, according to the color legend presented, and it is proportional to the number of branches present. The outgroup (hidden) used was the RIIID domain from human DCR1 (PDB ID: 5ZAK) and the bootstrap values are represented by dark blue circles (minimum 70). (**B-D**) Superposition of the RIIID and RIIID-like domains from DCRs (**B** and **C**) and DROSHA (**D**) proteins, highlighting the main variability spots (α5-α6 loop in both RIIID-I and RIIID-II from DCR1-2, and RIIID-II from DROSHA, as well as α2-α3 loop in RIIID-like from DROSHA; *see also* Figures S25-S30). In (**B**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC001750); **Diptera:** *D. melanogaster* (FBpp0083717); **Hemiptera:** *B. tabaci* (Bta12886); **Hymenoptera:** *A. melífera* (GB44595); and **Lepidoptera:** *M. sexta* (Msex2.10734). In (**C**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC001108); **Diptera:** *D. melanogaster* (FBpp0086061); **Hemiptera:** *B. tabaci* (Bta10685); **Hymenoptera:** *A. melífera* (GB48923); and **Lepidoptera:** *M. sexta* (Msex2.04462). In (**D**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC016208); **Diptera:** *D. melanogaster* (FBpp0087926); **Hemiptera:** *B. tabaci* (Bta10972); **Hymenoptera:** *A. melífera* (GB49096); and **Lepidoptera:** *M. sexta* (Msex2.00504).

than those from vertebrates and display low sequence identity between different orders. One explanation for the evolutionary divergence of the α5-α6 loop in insects is the existence of DCR proteins which interact with different AGO proteins, something that is not observed in vertebrates (Maillard et al., 2019). It has also been suggested that the α5-α6 loop of RIIID-I helps to align or direct the dsRNA substrates into the enzyme’s active sites, reason for which it was named the "Positioning loop" in Giardia DCR (MacRae et al., 2007). Nonetheless, the function of α5-α6 loop remains to be assessed in insects, and further investigation is needed to confirm whether it mirrors the roles described for human or Giardia DCRs (MacRae et al., 2007; Sasaki and Shimizu, 2007). A general trend we observed concerning this loop region is that the RIIID-II subunit exhibits shorter loops (45-52 residues) than the RIIID-I subunit (70-118 residues), which accounts for the majority of the second subunit's reduced length. The only exceptions to this are DCR2 from dipterans, suborder Brachycera (*e.g.,* Drosophila species), wherein the RIIID-II subunits have α5-α6 loops as large as those from RIIID-I (on average 80 and 97 residues, respectively), and DCR1 from lepidopterans, in which the RIIID-I subunits have α5-α6 loops as small as those from RIIID-II (on average 54 and 47 residues, respectively). A second general trend we observed is the strictly conserved amino acid composition of α5-α6 loops in RIIID-II from all DCR1 proteins, wherein 25-28 % of the residues are negatively charged (particularly Asp). Interestingly, this conservation occurs even in dipterans of suborder Nematocera (*e.g.,* Aedes and Anopheles genera) and ticks (Ixodidae family; Arthropoda outgroup), in which the α5-α6 loops are larger (61-65 residues) than the average length of those observed for RIIID-II subunits (~ 50 residues). In human DCR, we found that the α5-α6 Loop helix from RIIID-II (position 100-150 in Figures S27 and S28; 5ZAK sequence) interacts with the DEAD/ResIII (Hel1) and Dicer Dimer domains (Figure 10B). Furthermore, we identified that the N-terminal region flanking the Loop helix makes extensive contact with the α2-α3 loop of RIIID-II and that the flanking C-terminal region can potentially interact with the Helicase C subdomain when DCR is in the ATP-bound conformation, or with dsRNA being threaded through the Helicase domain (Figure 10B). The details how these interactions may influence the DCR mechanism deserves more attention than we can give here, but it is important to point out that regions enriched in negatively charged residues play special biological roles: they may regulate gene expression (Hsu et al., 2015; Kumar et al., 2014; Oliver et al., 2010), mimic the phosphate backbone of nucleic acids (Putnam and Tainer, 2005; Wang et al., 2014), and bind metal ions (Kinoshita et al., 2011) or specific domains (Scartezzini et al., 1997). While most D/E-rich repeats are predicted to be unstructured, as was observed for both α5-α6 loops in the RIIIDs of human DCR (PDB ID: 5ZAK), peptides composed solely of either Asp or Glu residues have been shown to adopt the structure of a polyproline-II helix; this suggests that a local structure can be attributed to unfolded or disordered D/E-rich regions. Polyproline-II helices, like β strands, exhibit an extended conformation that facilitates binding to partner molecules (Kumar and Bansal, 2016). Although the presence of proline residues are not necessary for the formation of polyproline-II helices, they are the most preferred residues within the composition of this secondary structure; in their absence, glycine, polar and charged residues are preferred (Adzhubei et al., 2013; Kumar and Bansal, 2016; Morgan and Rubenstein, 2013). We observed that, in addition to displaying larger-than-average D/E-rich loops, DCR1 RIIID-II subunits from Nematocera dipterans also present the highest Gly content among all α5-α6 loops, further suggesting that this region can adopt the structure of a polyproline-II helix.

We investigated the regions displaying higher variability by mapping the sequences and evolutionary coefficients of insect RIIIDs to their homologous domains within the structure of human DCR and DROSHA proteins (PDB IDs: 5ZAK and 5B16; Figures S25-S30). As in our previous analyses of other domains, we found that regions accumulating more mutations are generally clustered on the three-dimensional structure and form contiguous solvent-exposed surfaces. For example, the C-terminal regions of α3, α5 and α7 form a contiguous solvent-exposed surface in both RIIID subunits of DCR1 and DCR2 (Figure 10A). In DROSHA, this surface has been shown to interact with the C-terminal tail of PASHA (Figure 10A) (Kwon et al., 2016). Furthermore, the Loop helix and subsequent unresolved region extending towards α6 (Figures S25-S30) are also adjacent to this solvent-exposed surface (Figure 10A). In RIIID-I of DCR2, an additional mutation-prone, solvent-exposed surface is formed by the C-terminal region of α2 and the unresolved region between α5 and the Loop helix (Figures 10B and S26).

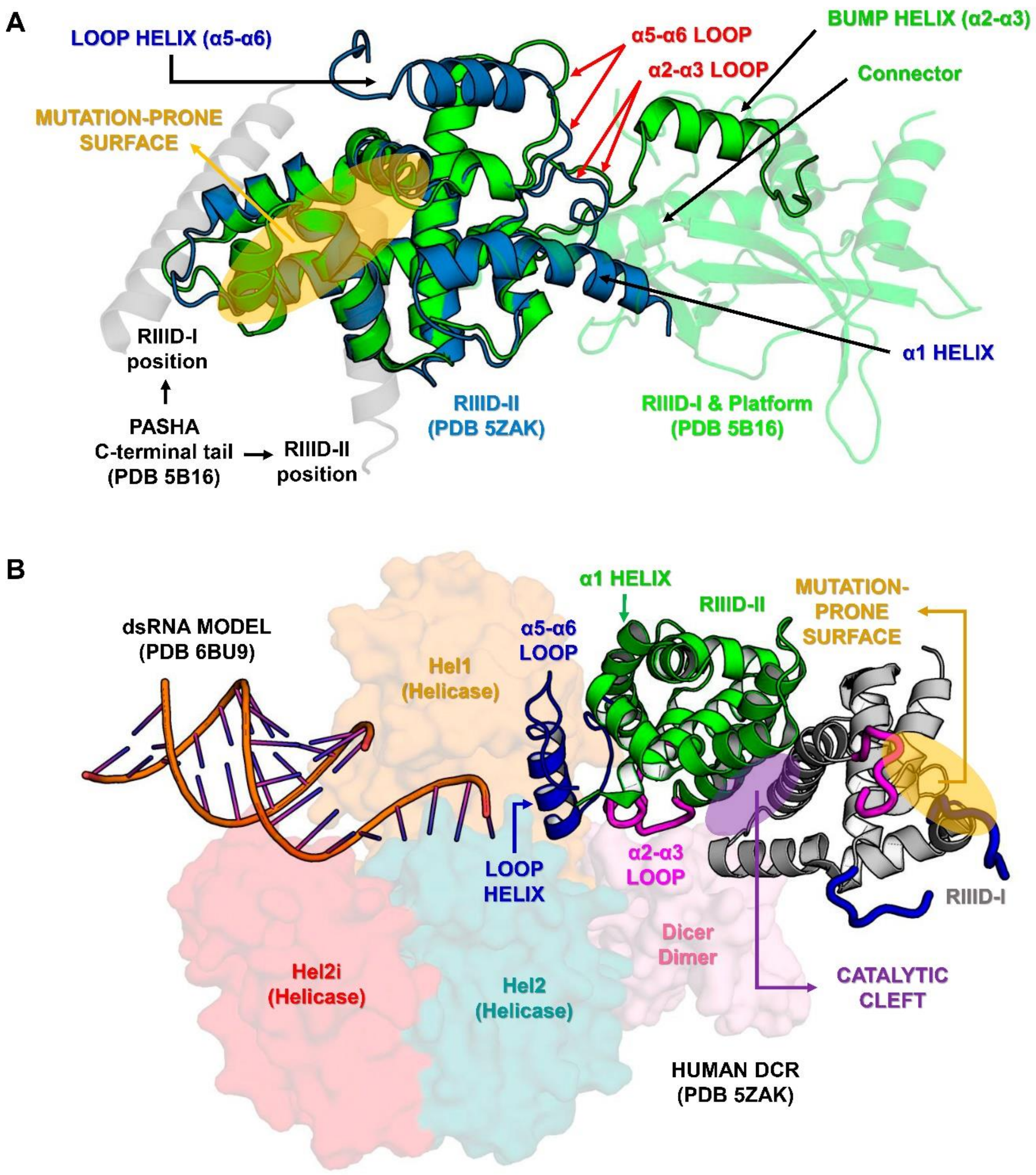

Figure 10. Variabilities within the Ribonuclease-III domain (RIIID). (**A**) Depiction of all of the different features we found in insect RIIIDs; this was achieved by superposing the second RIIID subunit (in blue) of human DCR (PDB ID: 5ZAK) onto the first RIIID subunit (in green) of human DROSHA (PDB ID: 5B16). The Platform domain of human DROSHA was kept in the image (green transparency) to show how the Connector helix acts as surrogate for helix α1 in the first RIIID subunit of DCR and DROSHA proteins. The Bump helix is a unique feature of DROSHA proteins, which display a long insertion in the α2-α3 loop. The Loop helix is typically found in the α5-α6 loops of RIIIDs belonging to DCR proteins. The mutation-prone surface was identified in insects and is composed by the C-terminal regions of helices α3, α5 and α7. In human DROSHA, this region has been shown to bind the C-terminal tail of PASHA at two different positions, depending on which of the two RIIID subunits the binding event occurs. (**B**) Overview of RIIID features in the context of DCR proteins. The Loop helix from RIIID-II interacts with the Hel1 and Dicer Dimer domains. The N-terminal region flanking the Loop helix makes extensive contact with the α2-α3 loop of RIIID-II, while the flanking C-terminal region can potentially interact with the Hel2 subdomain when DCR is

Figure 10. (cont.) in the ATP-bound conformation, or with dsRNA being threaded through the Helicase domain. The α1 helix of RIIID-II is prone to accumulate mutations and located opposite to the catalytic sites; this region forms a solvent-exposed surface in-between the Hel1 domain and the rest of RIIID-II. In RIIID-I, a mutation-prone, solvent-exposed surface is formed by the C-terminal region of α2 and the unresolved region between α5 and the "Loop helix". Just for illustrative purposes, a dsRNA molecule was modeled onto the structure of human DCR using the dsRNA from PDB 6BU9 as template.

Interestingly, the same two regions are also prone to mutations in RIIID-II of DCR1and DCR2, but they do not form solvent-exposed surfaces; rather, they co-participate in intradomain interactions with the Helicase and Dicer Dimer domains (Figures S28 and S29). Finally, we found that the α1 helix of RIIID-II is prone to accumulate mutations; this region forms a solvent-exposed surface in-between the RIIID-II and DEAD/ResIII (Hel1) domains, located opposite to the catalytic cleft (Figure 10B). While this surface has no known or apparent function, the α1 helix appears to be important for maintaining the DEAD/ResIII domain in a relatively fixed position relative to the catalytic domains (Figure 10B).

Variability within the Helicase domain

Dicers can be classified as RIG-I-like proteins due to their harboring an RNA Helicase domain at the N-terminus; in particular, RIG-I-like proteins differ from other RNA helicases because they exhibit a large insertion between the two canonical Helicase subdomains, DEAD/ResIII and Helicase C (aka RecA-like domains) (Jankowsky and Fairman-Williams, 2010). According to Sinha and coworkers (2018), the structure of the Helicase domain from *D. melanogaster* DCR2 (dmDCR2) is composed by four functional subdomains: DEAD/ResIII (aka Hel1), Hel2i (the large insertion found in RIG-I-like proteins), Helicase C (aka Hel2) and Pincer (Figure 11; Figures S31-S32) (Sinha et al., 2018). With respect to the cryo-EM structure of dmDCR2, the Hel1 and Hel2 domains, along with Pincer, could be fitted into the electron density map as a single rigid body. On the other hand, the Hel2i domain had to be fitted as a separate rigid body. In most RIG-I-like helicases, the functional domains perform activities that are intrinsic to ATP-driven translocases (Jankowsky and Fairman-Williams, 2010). Whether translocation on the dsRNA substrate is also coupled with unwinding of the helix is still unclear for most RIG-I-like proteins. According to Jankowsky & Fairman-Williams (2010), six conserved-sequence motifs of RIG-I-like helicases are important for ATP binding and hydrolysis (Q, I, II, III, Va and VI) and five are important for RNA binding (Ia, Ic, IV, IVa and V) (Jankowsky and Fairman-Williams, 2010). Among the conserved-sequence motifs that we identified in the DEAD/ResIII subdomain of DCR1, those related to RNA binding are

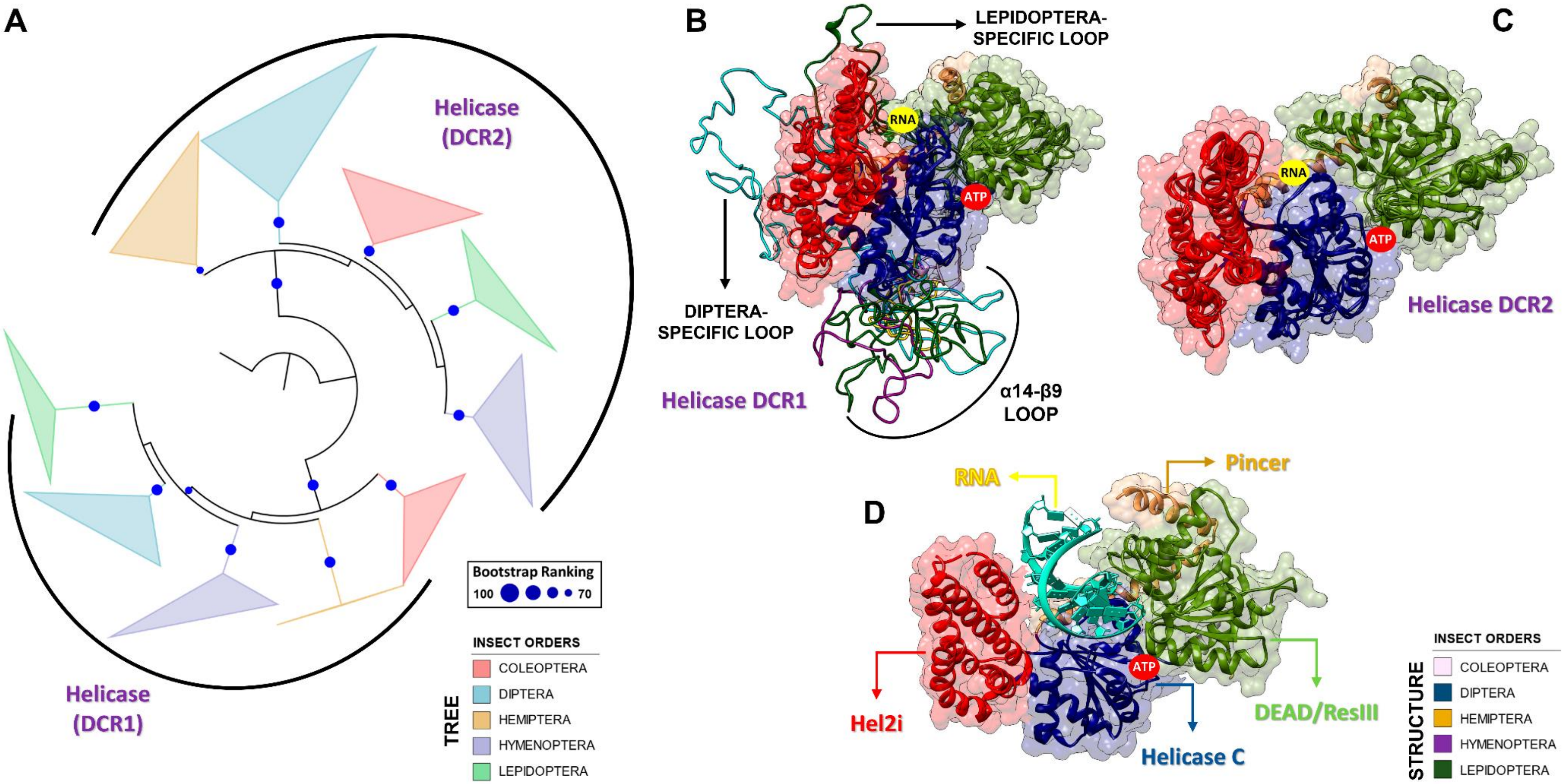

Figure 11. **Structural and phylogenetic analysis of Helicase domain.** (**A**) Maximum likelihood analysis of the complete Helicase domain present in the proteins DCR1 and DCR2 from species belonging to the five insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera). Each triangle represents an insect order, according to the color legend presented, and it is proportional to the number of branches present. The outgroup (hidden) used was the Helicase domain from human DCR1 (PDB ID: 5ZAK) and the bootstrap values are represented by dark blue circles (minimum 70). (**B** and **C**) Superposition of the models from DCR Helicase domains, highlighting the main variability spots. Specifically in the DCR1 Helicase models (**B**), lepidopteran and dipteran-specific loops (β6-α7 and β13-α18 regions, respectively), as well as α14- β9 loop (identified in all insect orders) were highlighted (*see also* Figure S31). In (**B**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC001750); **Diptera:** *D. melanogaster* (FBpp0083717);

Figure 11. (cont.) Hemiptera: *B. tabaci* (Bta12886); **Hymenoptera:** *A. melífera* (GB44595); and **Lepidoptera:** *M. sexta* (Msex2.10734). In (**C**), the species that represented each insect order were: **Coleoptera:** *T. castaneum* (TC001108); **Diptera:** *D. melanogaster* (FBpp0086061); **Hemiptera:** *B. tabaci* (Bta10685); **Hymenoptera:** *A. melífera* (GB48923); and **Lepidoptera:** *M. sexta* (Msex2.04462). (**D**) Illustrative representation of Helicase domain from human RIG-I (PDB ID: 5E3H), where its four functional subdomains were highlighted: *olive green* - DEAD/ResIII (Hel1); *red* - Hel2i; *dark blue* - Helicase C (Hel2); and *light brown* - Pincer. RNA molecule is represented in *cyan blue* color. The recognition sites of ATP hydrolysis and binding as well as RNA binding are represented by red and yellow circles, respectively.

degenerate compared to those related to ATP binding and hydrolysis. For example, motif Ia, which typically harbors conserved residues that establish side-chain contacts with RNA, is almost completely disfigured, and motif Ic displays variations in the canonical RNA-binding residue that characterizes RIG-I-like helicases (Figures S31) (Jankowsky and Fairman-Williams, 2010). We also noticed that the Lepidoptera order does not display the canonical glutamine residue in motif Q (Figures S31); as such, the Helicase domain of species within this order is likely able to hydrolyze any of the four NTPs (in contrast, glutamine introduces specific contacts that select for the adenine base). On the other hand, all of the conserved-sequence motifs that we found in the DCR2 Helicase domain displayed the canonical ATP- and RNA-binding residues (Figure S32). For translocation and/or unwinding to occur on the dsRNA substrate, the ATP-binding event must communicate with the RNA-binding event (and vice-versa). However, the ATP- and RNA-binding sites are separated by ~30 Å and it is still unclear how this communication is established between them (Mastrangelo et al., 2012). Recent evidence has identified two positions within motif V that are critical for communication between the ATP-binding pocket and the RNA-binding cleft in the closely related family of viral DExH helicases (aka NS3/NPH-II family) (Du Pont et al., 2020). Interestingly, these positions, which predominantly display a threonine and serine (T407 and S411) that interact with each other, displayed the highest residue variability across motif V of all flavivirus NS3 helicases. Overall, Du Pont and coworkers (2020) showed that removing the polar groups with H-bonding potential from positions T407 and S411 (*see* blue circles in Figures S31 and S32) increases the helicase turnover rate, especially in the latter position, but have opposite effects by either improving (T407) or reducing (S411) the affinity for dsRNA substrates in the presence of ATP. In particular, we found that the presence of non-polar group at position T407 (such as methyl or thiol) is important for coordination of four hydrophobic residues that influence the ATP- and RNA-binding residues in NS3 helicases (Figure 12A). We observed that the hydrophobic nature of these residues, as well as the presence of a non-polar group at the T407-equivalent position, are also conserved in the Helicase domains of insect DCR proteins (*see* black circles in Figures S31 and S32). This suggests that a similar mechanism for the communication between the ATP- and RNA-binding sites may apply to viral and RNAi-related helicases. The four hydrophobic residues coordinated around the insect T407 and S411 counterparts, henceforth denominated iT407 and iS411 for the sake of simplicity, are distributed across motifs IV, IVa and V, but we found they further coordinate a second layer of eleven conserved hydrophobic residues in the structure of RIG-I-like helicases (PDB ID: 5E3H). These residues span motifs Va and VI in insect DCR proteins, as well as a non-motif region between motifs IVa and V (Figure 12A; *see* grey circles in Figures S31 and S32). In RID-I-like helicases, this non-motif region is conserved and also harbors important RNA-binding residues (PDB IDs: 5E3H and 4A36). Hence, we have designated this region as motif IVb. We found that this second layer of hydrophobic residues can directly influence the positions of the ATP- and RNA-binding residues (red and yellow circles in Figures S31 and S32, respectively); thus, the central position occupied by iT407 in this network of hydrophobic contacts appears to play an important role in regulating the translocation and/or unwinding activity of DCR helicases by indirectly coordinating residues at both binding sites (Figure 12A). In particular, we found that the ATP-binding residues regulated by iT407 and iS411 (motifs Va and VI) are all conserved in DCR1 and DCR2, but the RNA-binding residues (motifs IV and IVa) are somewhat degenerate in DCR1. Thus, at least where the translocation and/or unwinding mechanisms are concerned, DCR2 binds to dsRNA in a more conserved manner.

We also noticed that, while present in DCR1, the canonical ATP-binding residues of motifs I (Walker A) and II (Walker B) display some variability and might render ATP hydrolysis less efficient in this protein, especially in lepidopteran species (Figures S31). In addition, the Lepidoptera order displays a large insertion that extends motif III in the DEAD/ResIII subdomain of DCR1 (Figures 11 and S31); motif III has been implicated in sensing both the ATP-hydrolysis state and nucleic acid-binding event in some SF1 and SF2 helicases (Caruthers and McKay, 2002; Papanikou et al., 2004). We also identified a dipteran-specific insertion between the Helicase C and Pincer subdomains of DCR1 (Figures 11 and S31). While the function of this insertion is elusive, it is placed in a privileged position to interact with or block any dsRNA molecule binding to the DCR1 Helicase domain (Figure 11B). This peculiarity of dipterans indicates that *D. melanogaster* might not be the best model for studying RNAi in insects. With regards to with DCR2, all five insect orders studied here display a large insertion between helix α14 and strand β9 of the DCR1 Helicase C subdomain (Figures 11B and S31). Again, the function of this insertion remains elusive, but we noticed that it is located near the Dicer Dimer domain in the structure of human DCR structure and in a privileged position to interact with the stem loop of pre-miRNAs in both the open and closed states of this enzyme (PDB IDs: 5ZAL, 5ZAM, and 5ZAK) (Liu et al., 2018). Furthermore, this insertion abuts the ATP-binding site and may interfere with the helicase turnover activity (Figure 11B). Overall, our data indicate that the DCR1 Helicase domain of insects is capable of hydrolyzing ATP efficiently but binds to dsRNA through a less conserved mechanism, which may explain the lower affinity of this domain for siRNA precursors. The large insertions we observed in the DCR1 Helicase domain could have originated by recombination of a long DNA fragment into the locus that encodes an ancestral DCR1 ortholog, thereby leading this enzyme to specialize in the processing of pre-miRNAs molecules (Deddouche et al., 2008).

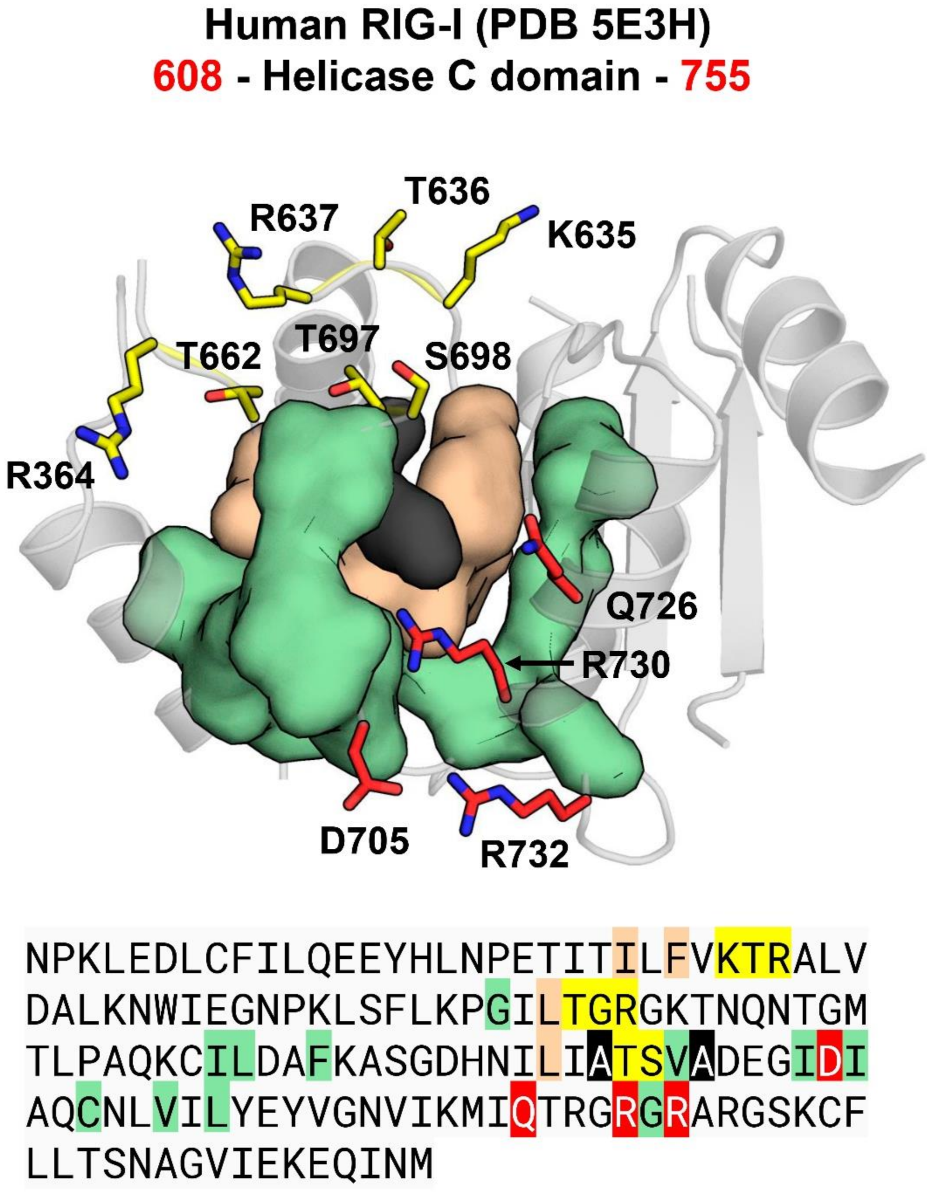
Human RIG-I (PDB 5E3H)
608 - Helicase C domain - 755
R637
T636
K635
T662
T697
S698
R364
Q726
R730
D705
R732
NPKLEDLCFILQEEYHLNPETITILFVKTRALV
DALKNWIEGNPKLSFLKPGILTGRGKTNQNTGM
TLPAQKCILDAFKASGDHNILIATSVADEGIDI
AQCNLVILYEYVGNVIKMIQTRGRGRARGSKCF
LLTSNAGVIEKEQINM

Figure 12. __Communication hub for the ATP- and RNA-binding site in RIG-I-like helicases.__ A network of hydrophobic interactions is arranged around two main amino acid residues (in black). The first layer

Figure 12. (cont.) of hydrophobic residues to interact with the core residues is composed by four residues (in beige) that span motifs IV, IVa and V in insect DCR proteins (*see* Figures S31 and S32). The second layer is composed by eleven residues (in olive) that span motifs Va and VI, as well as a hitherto undescribed region which we designated as motif IVb. Together, these two layers coordinate the positioning of the ATP- and RNA-binding residues (in red and yellow, respectively). This coordination is important because for translocation and/or unwinding to occur on the dsRNA substrate, the ATP-binding event must communicate with the RNA-binding event (and vice-versa). In insect DCR proteins, the residues participating in this hub are also conserved, which suggests that a similar mechanism for the communication between the ATP- and RNA-binding sites may apply to viral and RNAi-related helicases (*see* blue, black, grey, red and yellow circles in Figures S31 and S32).

CONCLUSIONS & FINAL REMARKS

The bioinformatics integration of the data presented in this study sheds light on the variability of domains within the RNAi machinery of five insect orders. We confirmed the universality of the RNAi mechanism in insects, as orthologues of the eight core proteins were identified in species of all five orders. All species are expected to have the basic elements of both the miRNA and siRNA machinery, but due to the fragmentation and incompleteness of a large number of publicly available genomes, as well as limitations in the methodologies for detection of divergent orthologues, some elements were not detected in several of the selected species. Thus, it is essential that future analyses be performed using curated databases harboring well assembled genomes/transcriptomes and using more than one method for ortholog detection. In this regard, we have now established well-defined sequence limits and better HMM profiles for annotating functional domains of the RNAi machinery in insects, which should greatly facilitate the identification of homologous proteins in both new and old genomes/transcriptomes. The structure-based sequence alignments that were generated using our methodology provide better inputs for phylogenetic inference and structure-function analyses of RNAi-related proteins. Unfortunately, the available structural data for insect proteins, especially those belonging to the RNAi machinery, are mostly limited to model species, such as *D. melanogaster*. Thus, further studies with non-model insect species are needed to allow for ample functional analyses of insect proteins. In particular, considering the RNAi pathway, it is imperative that more structural models with atomic resolution be solved in order for us to answer questions about the intricacies of this mechanism in insects. Nonetheless, our results show that considerable variability exists in elements of the RNAi machinery, all of which can potentially affect the efficiency of gene silencing triggered by exogenous RNA.

Regulation mechanisms of the siRNA pathway have coevolved with viral infections, and among the insect orders studied here, lepidopterans have been shown to be the most susceptible to viral attacks (approximately 80 % of the species), followed by the dipterans (9 %), coleopterans (5 %), hymenopterans (4 %) and hemipterans (1 %) (Swevers et al., 2013). One can argue that this observation correlates with the efficiency of a given order in controlling viral infections through RNAi-mediated mechanisms; if true, lepidopterans would be expected to show the lowest efficiency. Intriguingly, our phylogenetic analyses have clearly shown that, in practically all domains analyzed, the Lepidoptera order has the greatest evolutionary distance compared to the other orders. This corroborates previous reports that underscore the different efficiencies displayed by lepidopteran species during exogenous dsRNA-mediated gene knockdown. The variability and phylogenetic distance that we observed may be evidence that sufficient idiosyncrasies exist in the RNAi machinery of Lepidoptera to set them apart from other insects. Coleopterans are generally susceptible to RNAi and display higher silencing efficiency than lepidopterans, which are generally recalcitrant to RNAi. This has led to the recently-approved commercialization of a new genetically modified crop event (MON87411) wherein the heterologous production of *Bt* toxins was coupled with the expression of dsRNA molecules in order to control the western corn rootworm (*Diabrotica virgifera virgifera*, LeConte; Coleoptera: Chrysomelidae) (Dias et al., 2020). Our analyses highlighted several variability hotspots within the core elements of the RNAi machinery, thereby enabling us to compare the data between non-efficient lepidopterans and those coleopteran species that exhibit acceptable silencing efficiencies. Four of the five domains we analyzed displayed differences which could explain the contrasting gene silencing efficiency between Coleoptera and Lepidoptera species: (i) dsrm; (ii) Helicase; (iii) PAZ; and (iv) RIIID. While these differences are readily apparent, most of them were found in proteins pertaining to the miRNA pathway, which, in theory, should not cause major disturbances in RNAi-mediated gene knockdown. Nevertheless, core RNAi enzymes from the miRNA and piRNA pathways have also been shown to participate in the exogenous RNAi responses of *Bombyx mori* (Lepidoptera), *Leptinotarsa decemlineata* (Coleoptera), and *D. melanogaster* (Diptera) (Cooper et al., 2019). Additionally, the miRNA pathway has been shown to play a role in the modulation of gene expression in response to viral infection in mammals (Claycomb, 2014), as well as to produce miRNAs that target specific sites of the viral genome (Trobaugh and Klimstra, 2017). It was even demonstrated that DROSHA, which acts upon pri-miRNAs in the nucleus, can be recruited to the cytoplasm in response to virus infections, where it has been proposed to cleave viral RNA secondary structures or host cytoplasmatic RNA hairpins (Shapiro, 2013). Therefore, we cannot exclude the possibility that differences in proteins of the miRNA pathway may somehow influence RNAi-mediated gene silencing sensitivity in lepidopterans. With that said, most of the variability displayed across insects were present in the loop regions of domains. The structure of large flexible loops is difficult to resolve; accordingly, most of them are not represented in the publicly available structural data, thereby limiting the quality of the homology models that can be generated. Nevertheless, these regions can significantly influence the activity of the proteins whereupon they are inserted; for example, they can modify substrate affinity, block catalytic sites, or even interact with other proteins. Hence, both *in vitro* and *in silico* studies aiming to characterize these regions are essential to completely elucidate the mechanism of action of the core RNAi proteins we analyzed.

A marked difference was found in the dsrm-II domain of LOQS-PB, wherein lepidopterans display an insertion, V(N/A)RR, in the β1-β2 loop region (Figure S13). As previously mentioned in previous sections, the dsrm β1-β2 loop binds to the minor groove of dsRNA and greatly affects the affinity for this substrate. In particular, the lepidopteran-specific insertion adds positive charges to this loop, which may increase the number of contacts made with the phosphate backbone and thereby improve the affinity of the DCR1 microprocessor for pre-miRNA. Alternatively, the insertion can extend the distance between the guanine-binding histidine in loop β1-β2 and the sequence-specific binding residue from helix α1, which will affect the size of the dsRNA regions that are specifically recognized by the dsrm domain (Masliah et al., 2013).

Compared to their coleopteran orthologues, the DCR1 Helicase domains of lepidopterans display a large insertion between β6 (motif III) and α7 in the DEAD/ResIII subdomain (Figure S31). Insertions in this region are common in other families of SF1 and SF2 helicases and have been implicated in the communication between the ATP- and RNA-binding sites (Caruthers and McKay, 2002; Papanikou et al., 2004). In addition, we showed that lepidopterans lack the canonical Q residue in the eponymous Q motif (Figure S31), giving rise to the intriguing possibility that the DCR1 Helicase domain of Lepidoptera may hydrolyze NTPs other than ATP. In parallel, the insertion between α14 and β9 in the Helicase C subdomain of DCR1, which protrudes towards the ATP-binding site (Figure 11B), displays many order-specific sequence segments that suggest the existence of a convoluted mechanism underlying the DCR1 helicase activity (Figures S31). This insertion can be considered the major difference between the Helicase domains of DCR1 and DCR2 and likely plays an important role in how this domain engages substrates in both proteins. While coleopteran species display the shortest α14-β9 insertions among all insect DCR1 proteins that we evaluated, lepidopterans display the longest; however, the role of this region in the processing of pre-miRNAs, or even siRNA precursors, remains to be explained.

Lepidopterans display an insertion of 3-5 amino acids in the β4-β5 loop (β-hairpin module) of the PAZ domain from AGO2 proteins (Figure S18). The β-hairpin module recognizes the 3' end of dsRNA molecules that are loaded onto AGO proteins. Therefore, this insertion can modify how lepidopterans interact with and load dsRNA during formation of the RISC complex (Song et al., 2003). With respect to the DCR1 PAZ domain, lepidopterans have acquired a positively-charged insertion at the N-terminal region; intriguingly, this insertion is similar to the N-terminal region of plant DCL1 proteins (Figure S19). As we have previously mentioned, this insertion could lead to lepidopteran DCR1 interacting with dsRNA in a different orientation compared to coleopteran DCR1, thereby triggering downstream variations in the gene knockdown efficiency. In parallel, the regions interacting with dsRNA in the PAZ domain of DCR2 proteins can also be considered an important source of variability between coleopterans and lepidopterans. Unlike AGO PAZ domains, the DCR counterpart harbors an insertion between β3-β4 (β6-β7 in AGO proteins) that is rich in polar and positively-charged residues (Figure S20). In the X-ray structure of the Platform-PAZ cassette of human DCR (PDB ID: 4NGD), this insertion is important for stabilizing the DCR-dsRNA complex and forms a helical structure (α2 in the PAZ domain of DCR1-2; Figures S19 and S20) that is associated with the release and transfer of the cleaved dsRNA molecule onto AGO proteins (Tian et al., 2014). In coleopterans, this insertion is shorter and has a more positive residual charge than its lepidopteran orthologues, which might result in a stronger interaction of this domain with the dsRNA backbone. Consequently, the PAZ domain of coleopterans might confer higher thermodynamic stability to the DCR2 microprocessor, allowing higher delivery rates of siRNAs to AGO proteins.

The most relevant regions of variability between the endonuclease domains of DCR proteins were found in the RIIID-I domain, more precisely in the α5-α6 loop (Figures S25 and S26). As mentioned before, this loop is responsible for the interaction of human DCR with AGO proteins and may also be involved in the catalytic mechanism (Sasaki and Shimizu, 2007). DCR1 RIIID-I domains from lepidopterans exhibit the most divergent α5-α6 loops among all species analyzed, displaying a large deletion after the Loop helix (Figure S25). This deletion may beget divergent DCR1-AGO1 interactions in lepidopterans compared to insects from other orders. Similarly, the DCR2 RIIID-I domains of lepidopterans maintain a conserved 4-residue signature in the α5-α6 loop, EXE(P/K), that differentiate them from all other analyzed species. The importance of this signature in the DCR2 mechanism is unclear, but its potential involvement in Lepidoptera RNAi efficiency should be investigated nonetheless (Figure S26).

It is also known that viral infections may leave "scars" in the host insect genome, the so-called endogenous viral elements (EVEs). Accordingly, EVEs related to transposons, baculoviruses and bracoviruses (viruses of parasitic wasps) can be found integrated in lepidopteran genomes (Drezen et al., 2017). As previously mentioned, (Supplementary Text ST1), defective viral genomes (DVGs) can be retro-transcribed into viral DNA (vDNA) and incorporated into the host genome as an EVE; these will then act as an immunological memory by providing additional substrate to help boost the RNA interference response through the siRNA pathway, potentially promoting viral persistence in insects (Vignuzzi and López, 2019). Moreover, in addition to giving rise to endogenous viral siRNAs (vsiRNAs) via DCR2-LOQS-PD processing (Figure 1; step 16), EVEs also produce viral piRNAs (vpiRNAs) that contribute to the antiviral response via the piRNA pathway (Guo et al., 2019; Kolliopoulou et al., 2019). In this regard, EVEs are widespread in arthropod genomes and commonly give rise to PIWI-interacting RNAs that can potentially play a role in the antiviral response (Ter-Horst et al., 2019). Interestingly, Cui and Holmes (2012) have also presented evidence that EVEs with high similarity to plant viruses are integrated in the genomes of mosquitoes, fruit flies, bees, ants, silkworm, pea aphid, Monarch butterfly and wasps (Cui and Holmes, 2012). We have found that lepidopterans carry a plant-like, positively-charged insertion at the N-terminal region of the DCR1 PAZ domain, suggesting that RNA recognition by DCR1 in this order may function similarly to that related with plant DCL1 (Figure S19). Furthermore, the Platform domain from lepidopterans, like those from plants, also display a cluster of positively-charged residues that are positioned adjacent to the PAZ N-terminal insertion. Why does lepidopteran DCR1 harbor similar characteristics to those of plant DCL1? These observations are particularly interesting given that more than 70 % of all agricultural pests are insects in the order Lepidoptera (Guan et al., 2018a). Indeed, much of the Lepidoptera diversity can be attributed to the radiation of species in association with flowering plants: they represent the single most diverse lineage of organisms to have primarily evolved dependent upon angiosperm plants, and their numbers exceed those of the other major plant-feeding insects, such as those belonging to the orders Heteroptera, Homoptera, and Coleoptera (Chrysomeloidea and Curculionoidea) (Powell, 2009). One hypothesis for the similarities between plants and Lepidoptera is that lepidopteran DCR1 can recognize plant pri- or pre-miRNAs that are ingested during feeding and then further process them to regulate the expression of their own genes, particularly those associated with countering the plant's defense mechanisms. This may provide a way for the insect to fine tune the expression of certain genes in accordance with the plant's miRNA-mediated response to predation. To test this hypothesis, one would need to compare the complementarity of the 5' and 3'-UTR regions of plant and lepidopteran mRNAs to the sequence of plant miRNAs that are overexpressed during insect feeding. This hypothesis also raises the question of whether lepidopterans have also evolved to take advantage of plant-produced vsiRNAs or vpiRNAs to defend themselves from plant viruses that can be ingested. If similar EVEs associated with plant viruses are present in the genomes of both plants and lepidopterans, then the plant-produced piRNAs or endo-siRNAs related to those EVEs, which are potentially being used to modulate a viral infection or transposable element, may also be used to trigger specific responses in the insect. What is clear is that lepidopterans engage different RNAi-related mechanisms in response to viral infections, and these mechanisms appear to differ from those involved with the responses found in other insects (Zografidis et al., 2015). For example, while DCR2 predominantly targets viral dsRNA during the infection of *B. mori* with its eponymous Cytoplasmic Polyhedrosis Virus (BmCPV), an unknown RNAse has also been linked to the origins of vsiRNA biogenesis and distribution, and an additional pathway is triggered in response to viral mRNA derived from a specific segment of the viral genome (Zografidis et al., 2015). Irrespective of the reason, these similarities between plant DCL1 and Lepidoptera DCR1 certainly merit further investigation.

While EVEs can encode functional proteins, for the most part, they become inactive over the course of evolution (Lavialle et al., 2013; Ryabov, 2017). Nevertheless, these elements can retain some advantageous characteristics, which, among other functions, can act to suppress other viral infections (some viruses produce antivirals proteins to overcome competition) Antivirals proteins encoded in endogenous vDNA can therefore equip the host with tools capable of turning a fatal viral infection into a latent infection. Alternatively, endogenous vDNA may also encode viral suppressors of RNAi (VSRs), which can weaken the host antiviral defense to turn an otherwise acute infection (in which the host eliminates the virus) into a persistent infection. The main modes of action for viral suppressors of RNAi are: (i) binding to the dsRNA substrate, which prevents cleavage by DCR2; (ii) binding to siRNA, which prevents loading into RISC; (iii) degrading the siRNA molecule; and (iv) direct interaction with DCR2 or AGO2, which prevents their actions (Mongelli and Saleh, 2016). Thus, both the antivirals and VSRs encoded in endogenous vDNA may influence the sensitivity of insects to RNAi-mediated gene silencing. In *D. melanogaster*, the expression of two insect VSRs and three out of six plant VSRs inhibited siRNA responses associated with viral RNA and injected dsRNA, suggesting that some viral suppressors can negatively impact the RNAi efficiency in some systems (Berry et al., 2009). Given the large number of viruses that infect Lepidoptera species, it is reasonable to speculate that EVEs derived from DVGs may generate molecules capable of, for example, binding to DCR2 or siRNAs and preventing their loading into RISC (Cooper et al., 2019). In sum, we found clear distinctions between domains from coleopterans and lepidopterans. While these variations alone cannot irrefutably explain the differences that have been observed in RNAi-mediated gene silencing efficiency between these orders, they underscore specific regions that should be addressed to better understand the RNAi mechanism in these insects.

Our results also highlight an important factor to be considered when evaluating the efficiency of RNAi-mediated gene silencing in insects: the structural stability of the DROSHA-pri-miRNA, DCR1-pre-miRNA and DCR2-dsRNA complexes. It is important to note that structural stability (*i.e.*, persistence of interactions, or robustness) is fundamentally different from thermodynamic stability (*i.e.*, binding free energy, or ΔG_{bind}). In the case of enzymes, such as DCR, structural stability speaks about the need of keeping the substrate in place for efficient catalysis, while thermodynamic stability refers to the affinity of the enzyme for its substrate. Consequently, the higher the structural stability of the aforementioned complexes (*i.e.*, the longer the substrate remains correctly positioned in the binding site), the higher the turnover rate of miRNA and siRNA produced. In this regard, the presence of elements that increase the structural stability of these microprocessor complexes is vital for an effective response of the RNAi machinery. Studies have shown that he Staufen C protein, unique to members of the Coleoptera order, is an important factor in the development of insect resistance to RNAi (Yoon et al., 2018). This protein contains multiple domains harboring the dsRBD fold, some of which have been shown to bind to dsRNA. Due to this structural characteristic, as well as the involvement of this protein in the DCR2-mediated processing of dsRNA into siRNAs, one can hypothesize that Staufen C confers structural stability to the DCR2 microprocessor in coleopterans. Therefore, it is important to identify other dsRNA-binding proteins that may also contribute positively to increasing the efficiency of dsRNA processing in insects, which should provide a better understanding of the RNAi silencing mechanism or even be used as a biotechnological tool.

Another important factor to be considered is how insects detect the presence of viruses since viral dsRNA (as well as exogenous dsRNAs) can be considered an important pathogen-associated molecular pattern (PAMP) (Kingsolver et al., 2013). In addition to the viral control mediated by RNAi, there are several other signaling pathways capable of controlling viral infections, mainly by triggering insect innate immune responses, among which we can highlight: (i) JAK-STAT, which regulates the downstream production of effector molecules, such as antimicrobial peptides (AMPs) (Brown et al., 2001; Kingsolver et al., 2013); and (ii) IMD and TOLL, which are NF-κB-related pathways in which the final transcriptional factors responsible for signal transduction are Relish (Rel1 and Rel2) and Dorsal/Dif, respectively (Gottar et al., 2002; Kingsolver et al., 2013). Not surprisingly, these three signal transduction pathways display crosstalk between each other, wherein the signal is transduced by protein kinases and culminates in the regulation of several target genes/proteins. In this context, DCR proteins, specifically DCR2, can be considered pathogen recognition receptors (PRRs) involved in the detection of viral infections in insects (Kingsolver et al., 2013). A study involving *D. melanogaster* infected with *Drosophila C Virus* (DCV) showed the participation of DCR2 Helicase domain in viral dsRNA recognition, which in turn stimulated the expression of antiviral genes through the upregulation of a cysteine-rich peptide, Vago, which acts in a similar way to mammalian RIG-I-like sensors (Deddouche et al., 2008; Paradkar et al., 2014). This mechanism involving DCR2 was also characterized in the *Culex quinquefasciatus* mosquito in response to the *West Nile Virus* (WNV), but some differences were observed when compared to the response displayed by *D. melanogaster* (Cheng et al., 2016; Paradkar et al., 2014). The presence of viral dsRNA is detected by the DCR2 Helicase domain, and the Rel2 transcription factor of *C. quinquefasciatus* induces the expression of the *vago* gene via TNF receptor-associated factor (TRAF). Thereafter, similar to what occurs in Drosophila, the secreted *Cx*Vago peptide induces the JAK-STAT-mediated antiviral response (Paradkar et al., 2012, 2014). In short, this mosquito's immune response can be considered a crosstalk between the RNAi, JAK-STAT and IMD pathways (Sim et al., 2014). The central role played by the DCR2 Helicase domain in activating molecular signaling during antiviral responses, including exogenous dsRNA, highlights the importance of identifying variability within this "hub" domain (Figure S32). We hypothesize that some of the variabilities we identified in DCR2 may produce yet unknown consequences in the Vago-mediated activation of the JAK-STAT pathway, or even in the biogenesis of DVGs (Poirier et al., 2018). No studies have yet reported the characterization of the JAK-STAT pathway in lepidopterans. It is also possible that other uncharacterized pathways may operate during the antiviral response of lepidopterans (Zografidis et al., 2015).

In parallel, studies have shown that the low efficiency of RNAi-mediated gene silencing in some insect species can be directly associated with the expression levels of miRNA/siRNA elements, which may provide a partial explanation for the differences in RNAi efficiency observed in Lepidoptera. For example, it is known that the expression levels of the *translin* gene (a component of the C3PO complex) are very low in *B. mori* and *M. sexta* cells, and in addition, some lepidopterans exhibit almost undetectable levels of the R2D2 transcript, even during viral infections (Cooper et al., 2019). Studies that overexpressed elements of insect RNAi machinery (AGO2 and DCR2) in lepidopteran cells reinforce these observations since they considerably increased the RNAi-mediated antiviral response (Santos et al., 2018). However, why is there such variation in the expression of insect RNAi-related genes? How does this regulation occur? It is known that in *D. melanogaster*, the transcription factor Forkhead box O (dFOXO) upregulates the expression of important genes in the RNAi pathway, such as AGO2 and DCR2 (Spellberg and Marr, 2015). Following on the participation of dFOXO in responses related to metabolic changes and its relationship with multiple stress responses, a recent study has identified the participation of insulin in the antiviral response of insect vectors (Ahlers et al., 2019). Insulin-mediated dFOXO repression inhibits the RNAi response (by suppressing the transcription of genes encoding the AGO2 and DCR2 proteins) and, in parallel, activates the JAK-STAT pathway (Ahlers et al., 2019). Could the insulin-mediated response be predominant in lepidopterans, thus culminating in the repression of genes related to the RNAi pathway? Considering that the signaling pathways mediated by the Vago peptide and insulin are distinct, even though both converge to achieve an antiviral response mediated by JAK-STAT, and the fact that all these findings have also been validated in lepidopterans, we hypothesize that mutations in the receptors that sense viral infections and/or exogenous dsRNA, such as the Helicase domain, may be related to the predominance of an insulin-mediated response in some species of this insect order. Although speculative at this point, this hypothesis, associated with the data presented here, may help explain the low efficiency of RNAi-mediated gene silencing in Lepidoptera.

Considering the application of RNAi as a biotechnological tool, one question lingers: is it possible to universally apply RNAi-mediated gene silencing to control insect pest populations? The data presented here show that we are likely to fail if we generalize the application of RNAi-mediated gene silencing based on the restricted studies of a few model organisms. We have pinpointed some intriguing peculiarities within the functional domains of the RNAi machinery that must be addressed using a more species-specific approach in order to understand the nuances of differences associated with RNAi mechanisms in insects. For example, dipterans of suborders Brachycera and Nematocera show markedly different characteristics across all of the domains we analyzed, implying that studies on *D. melanogaster* may not provide a solid framework for understanding RNAi in *Aedes aegypti*, and vice-versa. Besides, small modifications to the experimental design can considerably increase the efficiency of exogenous dsRNA-mediated gene silencing in specific species. Recent studies have shown that for two lepidopteran species (*Helicoverpa armigera* and *Ostrinia furnicalis*), the presence of GGU nucleotides in exogenously administered dsRNA considerably increases siRNA production due to cleavage by DCR2, downstream of this motif. On the other hand, the same study showed that in *T. castaneum*, a member of the Coleoptera order, dsRNA was cut downstream of more diverse sites, such as AAG, GUG, and GUU (Guan et al., 2018b). In light of these reports, it is crucial to decipher how DCR2 recognizes the motifs upstream of the cleavage sites, as this would significantly improve the design of exogenous dsRNAs and considerably increase the efficiency of gene knockdown, especially in lepidopteran species.

Overall, it can be concluded that studies focusing on the genetic and structural variability of the core RNAi proteins are crucial to better understand how insects fine tune their RNAi-mediated development and antiviral response, which will ultimately drive how we design adapted biotechnological tools for the control of insect pest populations.

REFERENCES

Adzhubei, A.A., Sternberg, M.J.E., Makarov, A.A., 2013. Polyproline-II helix in proteins: structure and function. J. Mol. Biol. 425, 2100–2132. doi:10.1016/j.jmb.2013.03.018

Agrawal, A., Rajamani, V., Reddy, V.S., Mukherjee, S.K., Bhatnagar, R.K., 2015. Transgenic plants over-expressing insect-specific microRNA acquire insecticidal activity against *Helicoverpa armigera*: an alternative to *Bt*-toxin technology. Transgenic Res 24, 791–801. doi:10.1007/s11248-015-9880-x

Ahlers, L.R.H., Trammell, C.E., Carrell, G.F., Mackinnon, S., Torrevillas, B.K., Chow, C.Y., Luckhart, S., Goodman, A.G., 2019. Insulin potentiates JAK/STAT signaling to broadly inhibit flavivirus replication in insect vectors. Cell Rep. 29, 1946–1960.e5. doi:10.1016/j.celrep.2019.10.029

Airs, P.M., Bartholomay, L.C., 2017. RNA interference for mosquito and mosquito-borne disease control. Insects 8. doi:10.3390/insects8010004

Almeida-Garcia, R., Lima Pepino Macedo, L., Cabral do Nascimento, D., Gillet, F.X., Moreira-Pinto, C.E., Faheem, M., Moreschi Basso, A.M., Mattar Silva, M.C., Grossi-de-Sa, M.F., 2017. Nucleases as a barrier to gene silencing in the cotton boll weevil, *Anthonomus grandis*. PLoS One 12, e0189600. doi:10.1371/journal.pone.0189600

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. doi:10.1093/nar/25.17.3389

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., Notredame, C., 2006. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. 34, W604–8. doi:10.1093/nar/gkl092

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–8. doi:10.1093/nar/gkp335

Bally, J., Fishilevich, E., Bowling, A.J., Pence, H.E., Narva, K.E., Waterhouse, P.M., 2018. Improved insect-proofing: expressing double-stranded RNA in chloroplasts. Pest Manag Sci 74, 1751–1758. doi:10.1002/ps.4870

Berry, B., Deddouche, S., Kirschner, D., Imler, J.-L., Antoniewski, C., 2009. Viral suppressors of RNA silencing hinder exogenous and endogenous small RNA pathways in Drosophila. PLoS One 4, e5866. doi:10.1371/journal.pone.0005866

Blaszczyk, J., Gan, J., Tropea, J.E., Court, D.L., Waugh, D.S., Ji, X., 2004. Non-catalytic assembly of ribonuclease III with double-stranded RNA. Structure 12, 457–466. doi:10.1016/j.str.2004.02.004

Blaszczyk, J., Tropea, J.E., Bubunenko, M., Routzahn, K.M., Waugh, D.S., Court, D.L., Ji, X., 2001. Crystallographic and modeling studies of RNase III suggest a mechanism for double-stranded RNA cleavage. Structure 9, 1225–1236.

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., Hannon, G.J., 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128, 1089–1103. doi:10.1016/j.cell.2007.01.043

Brown, S., Hu, N., Hombría, J.C., 2001. Identification of the first invertebrate interleukin JAK/STAT receptor, the Drosophila gene domeless. Curr. Biol. 11, 1700–1705. doi:10.1016/s0960-9822(01)00524-3

Burd, C.G., Dreyfuss, G., 1994. Conserved structures and diversity of functions of RNA-binding proteins. Science 265, 615–621. doi:10.1126/science.8036511

Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J.M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D.S., Green, R.K., Guranovic, V., Guzenko, D., Hudson, B.P., Kalro, T., Liang, Y., Lowe, R., Namkoong, H., Peisach, E., Periskova, I., Prlic, A., Randle, C., Rose, A., Rose, P., Sala, R., Sekharan, M., Shao, C., Tan, L., Tao, Y.-P., Valasatava, Y., Voigt, M., Westbrook, J., Woo, J., Yang, H., Young, J., Zhuravleva, M., Zardecki, C., 2019. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology, and energy. Nucleic Acids Res. 47, D464–D474. doi:10.1093/nar/gky1004

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. doi:10.1093/bioinformatics/btp348

Caruthers, J.M., McKay, D.B., 2002. Helicase structure and mechanism. Curr. Opin. Struct. Biol. 12, 123–133. doi:10.1016/s0959-440x(02)00298-1

Cenik, E.S., Fukunaga, R., Lu, G., Dutcher, R., Wang, Y., Tanaka Hall, T.M., Zamore, P.D., 2011. Phosphate and R2D2 restrict the substrate specificity of Dicer-2, an ATP-driven ribonuclease. Mol. Cell 42, 172–184. doi:10.1016/j.molcel.2011.03.002

Cerutti, H., Casas-Mollano, J.A., 2006. On the origin and functions of RNA-mediated silencing: from protists to man. Curr. Genet. 50, 81–99. doi:10.1007/s00294-006-0078-x

Cerutti, L., Mian, N., Bateman, A., 2000. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. Trends Biochem. Sci. 25, 481–482. doi:10.1016/S0968-0004(00)01641-8

Cheng, G., Liu, Y., Wang, P., Xiao, X., 2016. Mosquito defense strategies against viral infection. Trends Parasitol. 32, 177–186. doi:10.1016/j.pt.2015.09.009

Chow, J., Kagan, J.C., 2018. The fly way of antiviral resistance and disease tolerance. Adv Immunol 140, 59–93. doi:10.1016/bs.ai.2018.08.002

Claycomb, J.M., 2014. Ancient endo-siRNA pathways reveal new tricks. Curr. Biol. 24, R703–15. doi:10.1016/j.cub.2014.06.009

Conrad, C., Rauhut, R., 2002. Ribonuclease III: new sense from nuisance. Int. J. Biochem. Cell Biol. 34, 116–129. doi:10.1016/S1357-2725(01)00112-1

Cooper, A.M., Silver, K., Zhang, J., Park, Y., Zhu, K.Y., 2019. Molecular mechanisms influencing efficiency of RNA interference in insects. Pest Manag Sci 75, 18–28. doi:10.1002/ps.5126

Cui, J., Holmes, E.C., 2012. Endogenous RNA viruses of plants in insect genomes. Virology 427, 77–79. doi:10.1016/j.virol.2012.02.014

Dalquen, D.A., Dessimoz, C., 2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol. Evol. 5, 1800–1806. doi:10.1093/gbe/evt132

Davis-Vogel, C., Van Allen, B., Van Hemert, J.L., Sethi, A., Nelson, M.E., Sashital, D.G., 2018. Identification and comparison of key RNA interference machinery from western corn rootworm, fall armyworm, and southern green stink bug. PLoS One 13, e0203160. doi:10.1371/journal.pone.0203160

de Jong, D., Eitel, M., Jakob, W., Osigus, H.-J., Hadrys, H., Desalle, R., Schierwater, B., 2009. Multiple dicer genes in the early-diverging metazoa. Mol. Biol. Evol. 26, 1333–1340. doi:10.1093/molbev/msp042

Deddouche, S., Matt, N., Budd, A., Mueller, S., Kemp, C., Galiana-Arnoux, D., Dostert, C., Antoniewski, C., Hoffmann, J.A., Imler, J.-L., 2008. The DExD/H-box helicase Dicer-2 mediates the induction of antiviral activity in drosophila. Nat. Immunol. 9, 1425–1432. doi:10.1038/ni.1664

Dias, N.P., Cagliari, D., Dos Santos, E.A., Smagghe, G., Jurat-Fuentes, J.L., Mishra, S., Nava, D.E., Zotti, M.J., 2020. Insecticidal gene silencing by RNAi in the neotropical region. Neotrop. Entomol. 49, 1–11. doi:10.1007/s13744-019-00722-4

Dias, R., Manny, A., Kolaczkowski, O., Kolaczkowski, B., 2017. Convergence of domain architecture, structure, and ligand affinity in animal and plant RNA-binding proteins. Mol. Biol. Evol. 34, 1429–1444. doi:10.1093/molbev/msx090

Dominska, M., Dykxhoorn, D.M., 2010. Breaking down the barriers: siRNA delivery and endosome escape. J. Cell Sci. 123, 1183–1189. doi:10.1242/jcs.066399

Dowling, D., Pauli, T., Donath, A., Meusemann, K., Podsiadlowski, L., Petersen, M., Peters, R.S., Mayer, C., Liu, S., Zhou, X., Misof, B., Niehuis, O., 2016. Phylogenetic origin and diversification of RNAi pathway genes in insects. Genome Biol. Evol. 8, 3784–3793. doi:10.1093/gbe/evw281

Drezen, J.-M., Josse, T., Bézier, A., Gauthier, J., Huguet, E., Herniou, E.A., 2017. Impact of lateral transfers on the genomes of Lepidoptera. Genes (Basel) 8. doi:10.3390/genes8110315

Du Pont, K.E., Davidson, R.B., McCullagh, M., Geiss, B.J., 2020. Motif V regulates energy transduction between the flavivirus NS3 ATPase and RNA-binding cleft. J. Biol. Chem. 295, 1551–1564. doi:10.1074/jbc.RA119.011922

Eddy, S.R., 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform 23, 205–211. doi:10.1142/9781848165632_0019

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., Mello, C.C., 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature 391, 806–811. doi:10.1038/35888

Fukunaga, R., Colpan, C., Han, B.W., Zamore, P.D., 2014. Inorganic phosphate blocks binding of pre-miRNA to Dicer 2 via its PAZ domain. EMBO J. 33, 371–384. doi:10.1002/embj.201387176

Gan, J., Tropea, J.E., Austin, B.P., Court, D.L., Waugh, D.S., Ji, X., 2006. Structural insight into the mechanism of double-stranded RNA processing by ribonuclease III. Cell 124, 355–366. doi:10.1016/j.cell.2005.11.034

Gertz, E.M., Yu, Y.-K., Agarwala, R., Schäffer, A.A., Altschul, S.F., 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 4, 41. doi:10.1186/1741-7007-4-41

Gottar, M., Gobert, V., Michel, T., Belvin, M., Duyk, G., Hoffmann, J.A., Ferrandon, D., Royet, J., 2002. The Drosophila immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. Nature 416, 640–644. doi:10.1038/nature734

Guan, R., Li, H.-C., Fan, Y.-J., Hu, S.-R., Christiaens, O., Smagghe, G., Miao, X.-X., 2018a. A nuclease specific to lepidopteran insects suppresses RNAi. J. Biol. Chem. 293, 6011–6021. doi:10.1074/jbc.RA117.001553

Guan, R., Hu, S., Li, H., Shi, Z., Miao, X., 2018b. The *in vivo* dsRNA cleavage has sequence preference in insects. Front. Physiol. 9, 1768. doi:10.3389/fphys.2018.01768

Guo, Z., Li, Y., Ding, S.-W., 2019. Small RNA-based antimicrobial immunity. Nat. Rev. Immunol. 19, 31–44. doi:10.1038/s41577-018-0071-x

Hall, T.M.T., 2005. Structure and function of argonaute proteins. Structure 13, 1403–1408. doi:10.1016/j.str.2005.08.005

Hsu, T.I., Lin, S.C., Lu, P.S., Chang, W.C., Hung, C.Y., Yeh, Y.M., Su, W.C., Liao, P.C., Hung, J.J., 2015. MMP7-mediated cleavage of nucleolin at Asp255 induces MMP9 expression to promote tumor malignancy. Oncogene 34, 826–837. doi:10.1038/onc.2014.22

Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26, 680–682. doi:10.1093/bioinformatics/btq003

Jankowsky, E., Fairman-Williams, M.E., 2010. Chapter 1. an introduction to RNA helicases: superfamilies, families, and major themes, in: Jankowsky, E. (Ed.), RNA Helicases. Royal Society of Chemistry, Cambridge, pp. 1–31. doi:10.1039/9781849732215-00001

Jaroszewski, L., Li, Z., Cai, X., Weber, C., Godzik, A., 2011. FFAS server: novel features and applications. Nucleic Acids Res. 39, W38–44. doi:10.1093/nar/gkr441

Jia, H., Kolaczkowski, O., Rolland, J., Kolaczkowski, B., 2017. Increased affinity for RNA targets evolved early in animal and plant Dicer lineages through different structural mechanisms. Mol. Biol. Evol. 34, 3047–3063. doi:10.1093/molbev/msx187

Jin, S., Singh, N.D., Li, L., Zhang, X., Daniell, H., 2015. Engineered chloroplast dsRNA silences cytochrome p450 monooxygenase, V-ATPase and chitin synthase genes in the insect gut and disrupts *Helicoverpa zea* larval development and pupation. Plant Biotechnol. J. 13, 435–446. doi:10.1111/pbi.12355

Joga, M.R., Zotti, M.J., Smagghe, G., Christiaens, O., 2016. RNAi efficiency, systemic properties, and novel delivery methods for pest insect control: what we know so far. Front. Physiol. 7, 553. doi:10.3389/fphys.2016.00553

Kandasamy, S.K., Fukunaga, R., 2016. Phosphate-binding pocket in Dicer 2 PAZ domain for high-fidelity siRNA production. Proc. Natl. Acad. Sci. USA 113, 14031–14036. doi:10.1073/pnas.1612393113

Katoh, K., Frith, M.C., 2012. Adding unaligned sequences into an existing alignment using MAFFT and LAST. Bioinformatics 28, 3144–3146. doi:10.1093/bioinformatics/bts578

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. doi:10.1093/molbev/mst010

Kavi, H.H., Fernandez, H., Xie, W., Birchler, J.A., 2008. Genetics and biochemistry of RNAi in Drosophila. Curr. Top. Microbiol. Immunol. 320, 37–75.

Kingsolver, M.B., Huang, Z., Hardy, R.W., 2013. Insect antiviral innate immunity: pathways, effectors, and connections. J. Mol. Biol. 425, 4921–4936. doi:10.1016/j.jmb.2013.10.006

Kinoshita, S., Katsumi, E., Yamamoto, H., Takeuchi, K., Watabe, S., 2011. Molecular and functional analyses of aspolin, a fish-specific protein extremely rich in aspartic acid. Mar. Biotechnol. 13, 517–526. doi:10.1007/s10126-010-9322-y

Kolliopoulou, A., Santos, D., Taning, C.N.T., Wynant, N., Vanden Broeck, J., Smagghe, G., Swevers, L., 2019. PIWI pathway against viruses in insects. Wiley Interdiscip Rev RNA 10, e1555. doi:10.1002/wrna.1555

Kosik, K.S., 2010. MicroRNAs and cellular phenotypy. Cell 143, 21–26. doi:10.1016/j.cell.2010.09.008

Kumar, A., Lualdi, M., Loncarek, J., Cho, Y.-W., Lee, J.-E., Ge, K., Kuehn, M.R., 2014. Loss of function of mouse Pax-Interacting Protein 1-associated glutamate rich protein 1a (Pagr1a) leads to reduced Bmp2 expression and defects in chorion and amnion development. Dev. Dyn. 243, 937–947. doi:10.1002/dvdy.24125

Kumar, P., Bansal, M., 2016. Structural and functional analyses of PolyProline-II helices in globular proteins. J. Struct. Biol. 196, 414–425. doi:10.1016/j.jsb.2016.09.006

Kurihara, Y., Takashi, Y., Watanabe, Y., 2006. The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. RNA 12, 206–212. doi:10.1261/rna.2146906

Kurowski, M.A., Bujnicki, J.M., 2003. GeneSilico protein structure prediction meta-server. Nucleic Acids Res. 31, 3305–3307. doi:10.1093/nar/gkg557

Kurzynska-Kokorniak, A., Pokornowska, M., Koralewska, N., Hoffmann, W., Bienkowska-Szewczyk, K., Figlerowicz, M., 2016. Revealing a new activity of the human Dicer DUF283 domain *in vitro*. Sci. Rep. 6, 23989. doi:10.1038/srep23989

Kwon, S.C., Nguyen, T.A., Choi, Y.-G., Jo, M.H., Hohng, S., Kim, V.N., Woo, J.-S., 2016. Structure of human DROSHA. Cell 164, 81–90. doi:10.1016/j.cell.2015.12.019

Laraki, G., Clerzius, G., Daher, A., Melendez-Peña, C., Daniels, S., Gatignol, A., 2008. Interactions between the double-stranded RNA-binding proteins TRBP and PACT define the Medipal domain that mediates protein-protein interactions. RNA Biol. 5, 92–103. doi:10.4161/rna.5.2.6069

Lau, P.-W., Guiley, K.Z., De, N., Potter, C.S., Carragher, B., MacRae, I.J., 2012. The molecular architecture of human Dicer. Nat. Struct. Mol. Biol. 19, 436–440. doi:10.1038/nsmb.2268

Lavialle, C., Cornelis, G., Dupressoir, A., Esnault, C., Heidmann, O., Vernochet, C., Heidmann, T., 2013. Paleovirology of "syncytins", retroviral *env* genes exapted for a role in placentation. Philos. Trans. R. Soc. Lond. B, Biol. Sci. 368, 20120507. doi:10.1098/rstb.2012.0507

Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307–1320. doi:10.1093/molbev/msn067

Leggewie, M., Schnettler, E., 2018. RNAi-mediated antiviral immunity in insects and their possible application. Curr Opin Virol 32, 108–114. doi:10.1016/j.coviro.2018.10.004

Letunic, I., Bork, P., 2018. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 46, D493–D496. doi:10.1093/nar/gkx922

Letunic, I., Bork, P., 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47, W256–W259. doi:10.1093/nar/gkz239

Li, H., Li, W.X., Ding, S.W., 2002. Induction and suppression of RNA silencing by an animal virus. Science 296, 1319–1321. doi:10.1126/science.1070948

Li, Z., Natarajan, P., Ye, Y., Hrabe, T., Godzik, A., 2014. POSA: a user-driven, interactive multiple protein structure alignment server. Nucleic Acids Res. 42, W240–5. doi:10.1093/nar/gku394

Li, Z., Tiley, G.P., Galuska, S.R., Reardon, C.R., Kidder, T.I., Rundell, R.J., Barker, M.S., 2018. Multiple large-scale gene and genome duplications during the evolution of hexapods. Proc. Natl. Acad. Sci. USA 115, 4713–4718. doi:10.1073/pnas.1710791115

Lin, H., Spradling, A.C., 1997. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. Development 124, 2463–2476.

Liu, J., Swevers, L., Iatrou, K., Huvenne, H., Smagghe, G., 2012. *Bombyx mori* DNA/RNA non-specific nuclease: expression of isoforms in insect culture cells, subcellular localization and functional assays. J. Insect Physiol. 58, 1166–1176. doi:10.1016/j.jinsphys.2012.05.016

Liu, Z., Wang, J., Cheng, H., Ke, X., Sun, L., Zhang, Q.C., Wang, H.-W., 2018. Cryo-EM Structure of Human Dicer and Its Complexes with a Pre-miRNA Substrate. Cell 173, 1191–1203.e12. doi:10.1016/j.cell.2018.03.080

MacRae, I.J., Doudna, J.A., 2007. An unusual case of pseudo-merohedral twinning in orthorhombic crystals of Dicer. Acta Crystallogr. Sect. D, Biol. Crystallogr. 63, 993–999. doi:10.1107/S0907444907036128

MacRae, I.J., Zhou, K., Doudna, J.A., 2007. Structural determinants of RNA recognition and cleavage by Dicer. Nat. Struct. Mol. Biol. 14, 934–940. doi:10.1038/nsmb1293

Maillard, P.V., van der Veen, A.G., Poirier, E.Z., Reis e Sousa, C., 2019. Slicing and dicing viruses: antiviral RNA interference in mammals. EMBO J. 38. doi:10.15252/embj.2018100941

Mallory, A., Vaucheret, H., 2010. Form, function, and regulation of Argonaute proteins. Plant Cell 22, 3879–3889. doi:10.1105/tpc.110.080671

Mamta, B., Rajam, M.V., 2017. RNAi technology: a new platform for crop pest control. Physiol. Mol. Biol. Plants 23, 487–501. doi:10.1007/s12298-017-0443-x

Masliah, G., Barraud, P., Allain, F.H.-T., 2013. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. Cell Mol. Life Sci. 70, 1875–1895. doi:10.1007/s00018-012-1119-x

Mastrangelo, E., Bolognesi, M., Milani, M., 2012. Flaviviral helicase: insights into the mechanism of action of a motor protein. Biochem. Biophys. Res. Commun. 417, 84–87. doi:10.1016/j.bbrc.2011.11.060

McEwan, D.L., Weisman, A.S., Hunter, C.P., 2012. Uptake of extracellular double-stranded RNA by SID-2. Mol. Cell 47, 746–754. doi:10.1016/j.molcel.2012.07.014

Méndez-Acevedo, K.M., Valdes, V.J., Asanov, A., Vaca, L., 2017. A novel family of mammalian transmembrane proteins involved in cholesterol transport. Sci. Rep. 7, 7450. doi:10.1038/s41598-017-07077-z

Milburn, D., Laskowski, R.A., Thornton, J.M., 1998. Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. Protein Eng 11, 855–859. doi:10.1093/protein/11.10.855

Mirdita, M., Steinegger, M., Söding, J., 2019. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics 35, 2856–2858. doi:10.1093/bioinformatics/bty1057

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., et al., 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346, 763–767. doi:10.1126/science.1257570

Mongelli, V., Saleh, M.-C., 2016. Bugs are not to be silenced: small RNA pathways and antiviral responses in insects. Annu. Rev. Virol. 3, 573–589. doi:10.1146/annurev-virology-110615-042447

Moran, Y., Agron, M., Praher, D., Technau, U., 2017. The evolutionary origin of plant and animal microRNAs. Nat. Ecol. Evol. 1, 27. doi:10.1038/s41559-016-0027

Morgan, A.A., Rubenstein, E., 2013. Proline: the distribution, frequency, positioning, and common functional roles of proline and polyproline sequences in the human proteome. PLoS One 8, e53785. doi:10.1371/journal.pone.0053785

Mukherjee, K., Campos, H., Kolaczkowski, B., 2013. Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants. Mol. Biol. Evol. 30, 627–641. doi:10.1093/molbev/mss263

Murphy, D., Dancis, B., Brown, J.R., 2008. The evolution of core proteins involved in microRNA biogenesis. BMC Evol. Biol. 8, 92. doi:10.1186/1471-2148-8-92

Napoli, C., Lemieux, C., Jorgensen, R., 1990. Introduction of a chimeric chalcone synthase gene into Petunia results in reversible co-suppression of homologous genes *in trans*. Plant Cell 2, 279–289. doi:10.1105/tpc.2.4.279

Okamura, K., Lai, E.C., 2008. Endogenous small interfering RNAs in animals. Nat. Rev. Mol. Cell Biol. 9, 673–678. doi:10.1038/nrm2479

Okonechnikov, K., Golosova, O., Fursov, M., UGENE team, 2012. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28, 1166–1167. doi:10.1093/bioinformatics/bts091

Oliver, D., Sheehan, B., South, H., Akbari, O., Pai, C.-Y., 2010. The chromosomal association/dissociation of the chromatin insulator protein Cp190 of *Drosophila melanogaster* is mediated by the BTB/POZ domain and two acidic regions. BMC Cell Biol. 11, 101. doi:10.1186/1471-2121-11-101

Olsen, P.H., Ambros, V., 1999. The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. Dev. Biol. 216, 671–680. doi:10.1006/dbio.1999.9523

Papanikou, E., Karamanou, S., Baud, C., Sianidis, G., Frank, M., Economou, A., 2004. Helicase Motif III in SecA is essential for coupling preprotein binding to translocation ATPase. EMBO Rep. 5, 807–811. doi:10.1038/sj.embor.7400206

Paradkar, P.N., Duchemin, J.-B., Voysey, R., Walker, P.J., 2014. Dicer-2-dependent activation of Culex Vago occurs via the TRAF-Rel2 signaling pathway. PLoS Negl. Trop. Dis. 8, e2823. doi:10.1371/journal.pntd.0002823

Paradkar, P.N., Trinidad, L., Voysey, R., Duchemin, J.-B., Walker, P.J., 2012. Secreted Vago restricts West Nile virus infection in Culex mosquito cells by activating the Jak-STAT pathway. Proc. Natl. Acad. Sci. USA 109, 18915–18920. doi:10.1073/pnas.1205231109

Park, J.-E., Heo, I., Tian, Y., Simanshu, D.K., Chang, H., Jee, D., Patel, D.J., Kim, V.N., 2011. Dicer recognizes the 5’ end of RNA for efficient and accurate processing. Nature 475, 201–205. doi:10.1038/nature10198

Peng, Y., Wang, K., Fu, W., Sheng, C., Han, Z., 2018. Biochemical comparison of dsRNA degrading nucleases in four different insects. Front. Physiol. 9, 624. doi:10.3389/fphys.2018.00624

Poirier, E.Z., Goic, B., Tomé-Poderti, L., Frangeul, L., Boussier, J., Gausson, V., Blanc, H., Vallet, T., Loyd, H., Levi, L.I., Lanciano, S., Baron, C., Merkling, S.H., Lambrechts, L., Mirouze, M., Carpenter, S., Vignuzzi, M., Saleh, M.-C., 2018. Dicer-2-dependent generation of viral DNA from defective genomes of RNA viruses modulates antiviral immunity in insects. Cell Host Microbe 23, 353–365.e8. doi:10.1016/j.chom.2018.02.001

Powell, J.A., 2009. Lepidoptera: moths, butterflies, in: Resh, V., Cardé, R. (Eds.), Encyclopedia of Insects. Elsevier, USA, pp. 559–587. doi:10.1016/B978-0-12-374144-8.00160-0

Prentice, K., Smagghe, G., Gheysen, G., Christiaens, O., 2019. Nuclease activity decreases the RNAi response in the sweetpotato weevil *Cylas puncticollis*. Insect Biochem. Mol. Biol. 110, 80–89. doi:10.1016/j.ibmb.2019.04.001

Putnam, C.D., Tainer, J.A., 2005. Protein mimicry of DNA and pathway regulation. DNA Repair (Amst.) 4, 1410–1420. doi:10.1016/j.dnarep.2005.08.007

Rombel, I.T., Sykes, K.F., Rayner, S., Johnston, S.A., 2002. ORF-FINDER: a vector for high-throughput gene identification. Gene 282, 33–41. doi:10.1016/S0378-1119(01)00819-8

Rozewicki, J., Li, S., Amada, K.M., Standley, D.M., Katoh, K., 2019. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res. 47, W5–W10. doi:10.1093/nar/gkz342

Rubio, M., Maestro, J.L., Piulachs, M.-D., Belles, X., 2018. Conserved association of Argonaute 1 and 2 proteins with miRNA and siRNA pathways throughout insect evolution, from cockroaches to flies. Biochim. Biophys. Acta Gene Regul. Mech. 1861, 554–560. doi:10.1016/j.bbagrm.2018.04.001

Ryabov, E.V., 2017. Invertebrate RNA virus diversity from a taxonomic point of view. J. Invertebr. Pathol. 147, 37–50. doi:10.1016/j.jip.2016.10.002

Ryter, J.M., Schultz, S.C., 1998. Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. EMBO J. 17, 7505–7513. doi:10.1093/emboj/17.24.7505

Saini, R.P., Raman, V., Dhandapani, G., Malhotra, E.V., Sreevathsa, R., Kumar, P.A., Sharma, T.R., Pattanayak, D., 2018. Silencing of HaAce1 gene by host-delivered artificial microRNA disrupts growth and development of *Helicoverpa armigera*. PLoS One 13, e0194150. doi:10.1371/journal.pone.0194150

Santos, D., Wynant, N., Van den Brande, S., Verdonckt, T.-W., Mingels, L., Peeters, P., Kolliopoulou, A., Swevers, L., Vanden Broeck, J., 2018. Insights into RNAi-based antiviral immunity in Lepidoptera: acute and persistent infections in *Bombyx mori* and Trichoplusia ni cell lines. Sci. Rep. 8, 2423. doi:10.1038/s41598-018-20848-6

Sasaki, T., Shimizu, N., 2007. Evolutionary conservation of a unique amino acid sequence in human DICER protein essential for binding to Argonaute family proteins. Gene 396, 312–320. doi:10.1016/j.gene.2007.04.001

Scartezzini, P., Egeo, A., Colella, S., Fumagalli, P., Arrigo, P., Nizetic, D., Taramelli, R., Rasore-Quartino, A., 1997. Cloning a new human gene from chromosome 21q22.3 encoding a glutamic acid-rich protein expressed in heart and skeletal muscle. Hum. Genet. 99, 387–392. doi:10.1007/s004390050377

Sempere, L.F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E., Ambros, V., 2004. Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. Genome Biol. 5, R13. doi:10.1186/gb-2004-5-3-r13

Senturia, R., Faller, M., Yin, S., Loo, J.A., Cascio, D., Sawaya, M.R., Hwang, D., Clubb, R.T., Guo, F., 2010. Structure of the dimerization domain of DiGeorge critical region 8. Protein Sci. 19, 1354–1365. doi:10.1002/pro.414

Shapiro, J.S., 2013. Processing of virus-derived cytoplasmic primary-microRNAs. Wiley Interdiscip Rev RNA 4, 463–471. doi:10.1002/wrna.1169

Sharath-Chandra, G., Asokan, R., Manamohan, M., Krishna Kumar, N., 2019. Enhancing RNAi by using concatemerized double-stranded RNA. Pest Manag Sci 75, 506–514. doi:10.1002/ps.5149

Sharma, C., Mohanty, D., 2018. Sequence- and structure-based analysis of proteins involved in miRNA biogenesis. J Biomol Struct Dyn 36, 139–151. doi:10.1080/07391102.2016.1269687

Shatsky, M., Nussinov, R., Wolfson, H.J., 2006. Optimization of multiple-sequence alignment based on multiple-structure alignment. Proteins 62, 209–217. doi:10.1002/prot.20665

Sim, S., Jupatanakul, N., Dimopoulos, G., 2014. Mosquito immunity against arboviruses. Viruses 6, 4479–4504. doi:10.3390/v6114479

Sinha, N.K., Iwasa, J., Shen, P.S., Bass, B.L., 2018. Dicer uses distinct modules for recognizing dsRNA termini. Science 359, 329–334. doi:10.1126/science.aaq0921

Song, H., Fan, Y., Zhang, J., Cooper, A.M., Silver, K., Li, D., Li, T., Ma, E., Zhu, K.Y., Zhang, J., 2019. Contributions of dsRNAses to differential RNAi efficiencies between the injection and oral delivery of dsRNA in *Locusta migratoria*. Pest Manag Sci 75, 1707–1717. doi:10.1002/ps.5291

Song, J.-J., Liu, J., Tolia, N.H., Schneiderman, J., Smith, S.K., Martienssen, R.A., Hannon, G.J., Joshua-Tor, L., 2003. The crystal structure of the Argonaute 2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. Nat. Struct. Biol. 10, 1026–1032. doi:10.1038/nsb1016

Spellberg, M.J., Marr, M.T., 2015. FOXO regulates RNA interference in Drosophila and protects from RNA virus infection. Proc. Natl. Acad. Sci. USA 112, 14587–14592. doi:10.1073/pnas.1517124112

Spielman, S.J., Kosakovsky Pond, S.L., 2018. Relative evolutionary rate inference in HyPhy with LEISR. PeerJ 6, e4339. doi:10.7717/peerj.4339

Spit, J., Philips, A., Wynant, N., Santos, D., Plaetinck, G., Vanden Broeck, J., 2017. Knockdown of nuclease activity in the gut enhances RNAi efficiency in the Colorado potato beetle, *Leptinotarsa decemlineata*, but not in the desert locust, *Schistocerca gregaria*. Insect Biochem. Mol. Biol. 81, 103–116. doi:10.1016/j.ibmb.2017.01.004

St Johnston, D., Brown, N.H., Gall, J.G., Jantsch, M., 1992. A conserved double-stranded RNA-binding domain. Proc. Natl. Acad. Sci. USA 89, 10979–10983. doi:10.1073/pnas.89.22.10979

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. doi:10.1093/bioinformatics/btu033

Swevers, L., Liu, J., Smagghe, G., 2018. Defense mechanisms against viral infection in Drosophila: RNAi and non-RNAi. Viruses 10. doi:10.3390/v10050230

Swevers, L., Vanden Broeck, J., Smagghe, G., 2013. The possible impact of persistent virus infection on the function of the RNAi machinery in insects: a hypothesis. Front. Physiol. 4, 319. doi:10.3389/fphys.2013.00319

Sydykova, D.K., Jack, B.R., Spielman, S.J., Wilke, C.O., 2017. Measuring evolutionary rates of proteins in a structural context. [version 2; peer review: 4 approved]. F1000Res. 6, 1845. doi:10.12688/f1000research.12874.2

Ter-Horst, A.M., Nigg, J.C., Dekker, F.M., Falk, B.W., 2019. Endogenous viral elements are widespread in arthropod genomes and commonly give rise to PIWI-interacting RNAs. J. Virol. 93. doi:10.1128/JVI.02124-18

Terenius, O., Papanicolaou, A., Garbutt, J.S., Eleftherianos, I., Huvenne, H., et al., 2011. RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. J. Insect Physiol. 57, 231–245. doi:10.1016/j.jinsphys.2010.11.006

Tian, Y., Simanshu, D.K., Ma, J.-B., Park, J.-E., Heo, I., Kim, V.N., Patel, D.J., 2014. A phosphate-binding pocket within the platform-PAZ-connector helix cassette of human Dicer. Mol. Cell 53, 606–616. doi:10.1016/j.molcel.2014.01.003

Tomoyasu, Y., Miller, S.C., Tomita, S., Schoppmeier, M., Grossmann, D., Bucher, G., 2008. Exploring systemic RNA interference in insects: a genome-wide survey for RNAi genes in Tribolium. Genome Biol. 9, R10. doi:10.1186/gb-2008-9-1-r10

Trettin, K.D., Sinha, N.K., Eckert, D.M., Apple, S.E., Bass, B.L., 2017. Loquacious-PD facilitates Drosophila Dicer-2 cleavage through interactions with the helicase domain and dsRNA. Proc. Natl. Acad. Sci. USA 114, E7939–E7948. doi:10.1073/pnas.1707063114

Trobaugh, D.W., Klimstra, W.B., 2017. MicroRNA regulation of RNA virus replication and pathogenesis. Trends Mol. Med. 23, 80–93. doi:10.1016/j.molmed.2016.11.003

Ulvila, J., Parikka, M., Kleino, A., Sormunen, R., Ezekowitz, R.A., Kocks, C., Rämet, M., 2006. Double-stranded RNA is internalized by scavenger receptor-mediated endocytosis in Drosophila S2 cells. J. Biol. Chem. 281, 14370–14375. doi:10.1074/jbc.M513868200

Vélez, A.M., Fishilevich, E., 2018. The mysteries of insect RNAi: a focus on dsRNA uptake and transport. Pestic Biochem Physiol 151, 25–31. doi:10.1016/j.pestbp.2018.08.005

Vignuzzi, M., López, C.B., 2019. Defective viral genomes are key drivers of the virus-host interaction. Nat. Microbiol. 4, 1075–1087. doi:10.1038/s41564-019-0465-y

Vuković, L., Koh, H.R., Myong, S., Schulten, K., 2014. Substrate recognition and specificity of double-stranded RNA binding proteins. Biochemistry 53, 3457–3466. doi:10.1021/bi500352s

Wallot, S., Leonardi, G., 2018. Deriving inferential statistics from recurrence plots: a recurrence-based test of differences between sample distributions and its comparison to the two-sample Kolmogorov-Smirnov test. Chaos 28, 085712. doi:10.1063/1.5024915

Wang, H.-C., Ho, C.-H., Hsu, K.-C., Yang, J.-M., Wang, A.H.-J., 2014. DNA mimic proteins: functions, structures, and bioinformatic analysis. Biochemistry 53, 2865–2874. doi:10.1021/bi5002689

Wang, K., Peng, Y., Fu, W., Shen, Z., Han, Z., 2019. Key factors determining variations in RNA interference efficacy mediated by different double-stranded RNA lengths in *Tribolium castaneum*. Insect Mol. Biol. 28, 235–245. doi:10.1111/imb.12546

Ward, N., Moreno-Hagelsieb, G., 2014. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? PLoS One 9, e101850. doi:10.1371/journal.pone.0101850

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., Zdobnov, E.M., 2018a. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35, 543–548. doi:10.1093/molbev/msx319

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T., 2018b. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296–W303. doi:10.1093/nar/gky427

Whitten, M.M., 2019. Novel RNAi delivery systems in the control of medical and veterinary pests. Curr. Opin. Insect Sci. 34, 1–6. doi:10.1016/j.cois.2019.02.001

Wickham, L., Duchaîne, T., Luo, M., Nabi, I.R., DesGroseillers, L., 1999. Mammalian staufen is a double-stranded-RNA- and tubulin-binding protein which localizes to the rough endoplasmic reticulum. Mol. Cell. Biol. 19, 2220–2230.

Wilson, K.A., Holland, D.J., Wetmore, S.D., 2016. Topology of RNA-protein nucleobase-amino acid π-π interactions and comparison to analogous DNA-protein π-π contacts. RNA 22, 696–708. doi:10.1261/rna.054924.115

Wilson, R.C., Tambe, A., Kidwell, M.A., Noland, C.L., Schneider, C.P., Doudna, J.A., 2015. Dicer-TRBP complex formation ensures accurate mammalian microRNA biogenesis. Mol. Cell 57, 397–407. doi:10.1016/j.molcel.2014.11.030

Winston, W.M., Molodowitch, C., Hunter, C.P., 2002. Systemic RNAi in *C. elegans* requires the putative transmembrane protein SID-1. Science 295, 2456–2459. doi:10.1126/science.1068836

Wu, S., Zhang, Y., 2007. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. 35, 3375–3382. doi:10.1093/nar/gkm251

Wynant, N., Santos, D., Vanden Broeck, J., 2017. The evolution of animal Argonautes: evidence for the absence of antiviral AGO Argonautes in vertebrates. Sci. Rep. 7, 9230. doi:10.1038/s41598-017-08043-5

Wynant, N., Santos, D., Verdonck, R., Spit, J., Van Wielendaele, P., Broeck, J.V., 2014. Identification, functional characterization and phylogenetic analysis of double stranded RNA degrading enzymes present in the gut of the desert locust, *Schistocerca gregaria*. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2013.12.008

Xiao, D., Gao, X., Xu, J., Liang, X., Li, Q., Yao, J., Zhu, K.Y., 2015. Clathrin-dependent endocytosis plays a predominant role in cellular uptake of double-stranded RNA in the red flour beetle. Insect Biochem. Mol. Biol. 60, 68–77. doi:10.1016/j.ibmb.2015.03.009

Yang, S.W., Chen, H.-Y., Yang, J., Machida, S., Chua, N.-H., Yuan, Y.A., 2010. Structure of Arabidopsis Hyponastic Leaves 1 and its molecular implications for miRNA processing. Structure 18, 594–605. doi:10.1016/j.str.2010.02.006

Ye, X., Paroo, Z., Liu, Q., 2007. Functional anatomy of the Drosophila microRNA-generating enzyme. J. Biol. Chem. 282, 28373–28378. doi:10.1074/jbc.M705208200

Yogindran, S., Rajam, M.V., 2016. Artificial miRNA-mediated silencing of ecdysone receptor (EcR) affects larval development and oogenesis in *Helicoverpa armigera*. Insect Biochem. Mol. Biol. 77, 21–30. doi:10.1016/j.ibmb.2016.07.009

Yoon, J.S., Gurusamy, D., Palli, S.R., 2017. Accumulation of dsRNA in endosomes contributes to inefficient RNA interference in the fall armyworm, *Spodoptera frugiperda*. Insect Biochem. Mol. Biol. 90, 53–60. doi:10.1016/j.ibmb.2017.09.011

Yoon, J.-S., Mogilicherla, K., Gurusamy, D., Chen, X., Chereddy, S.C.R.R., Palli, S.R., 2018. Double-stranded RNA binding protein, Staufen, is required for the initiation of RNAi in coleopteran insects. Proc. Natl. Acad. Sci. USA 115, 8334–8339. doi:10.1073/pnas.1809381115

Yu, N., Christiaens, O., Liu, J., Niu, J., Cappelle, K., Caccia, S., Huvenne, H., Smagghe, G., 2013. Delivery of dsRNA for RNAi in insects: an overview and future directions. Insect Sci 20, 4–14. doi:10.1111/j.1744-7917.2012.01534.x

Zhang, J., Khan, S.A., Heckel, D.G., Bock, R., 2017a. Next-generation insect-resistant plants: RNAi-mediated crop protection. Trends Biotechnol. 35, 871–882. doi:10.1016/j.tibtech.2017.04.009

Zhang, X., Li, P., Lin, J., Huang, H., Yin, B., Zeng, Y., 2017b. The insertion in the double-stranded RNA binding domain of human Drosha is important for its function. Biochim. Biophys. Acta Gene Regul. Mech. 1860, 1179–1188. doi:10.1016/j.bbagrm.2017.10.004

Zhang, M., Leong, H.W., 2010. Bidirectional best hit r-window gene clusters. BMC Bioinformatics 11 Suppl 1, S63. doi:10.1186/1471-2105-11-S1-S63

Zhu, J.-K., 2008. Reconstituting plant miRNA biogenesis. Proc. Natl. Acad. Sci. USA 105, 9851–9852. doi:10.1073/pnas.0805207105

Zografidis, A., Van Nieuwerburgh, F., Kolliopoulou, A., Apostolou-Karampelis, K., Head, S.R., Deforce, D., Smagghe, G., Swevers, L., 2015. Viral small-RNA analysis of *Bombyx mori* larval midgut during persistent and pathogenic cytoplasmic polyhedrosis virus infection. J. Virol. 89, 11473–11486. doi:10.1128/JVI.01695-15

SUPPLEMENTARY MATERIAL

Due to their large size, the figures and other supplementary files in this Chapter are permanently deposited at the following link. Therefore, only the subtitles/titles of the files are provided here:

https://www.tandfonline.com/doi/suppl/10.1080/15476286.2020.1861816/suppl_file/krnb_a_1861816_sm0917.zip

Figure S1. Structural representation of AGO-specific domains. Identification of each AGO-specific domain in human AGO2 (PDB ID: 4Z4K), where each one was highlighted: *red* - ArgoN; *yellow* - ArgoL1; *forest green* - ArgoL2; *dark blue* - ArgoMid; *light green* - PAZ (not AGO-specific domain); and *pink* - Piwi.

Figure S2. Phylogenetic analysis of AGO-specific domains. (**A-E**) Maximum likelihood analysis of the ArgoN (**A**), ArgoL1 (**B**), ArgoL2 (**C**), ArgoMid (**D**) and Piwi (**E**) present in the proteins AGO1 and AGO2 from species belonging to the five insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera). Each triangle represents an insect order, according to the color legend presented, and it is proportional to the number of branches present. The outgroups (hidden) used come from human AGO1 and AGO2 (PDB IDs: 4W5N; 4Z4F; 4Z4H; 5T7B and 4ZAF) and the bootstrap values are represented by dark blue circles (minimum 70).

Figure S3. Phylogenetic analysis of Connector domain. Maximum likelihood analysis of the Connector domain presents in the proteins DCR1, DCR2 and DROSHA from species belonging to the five insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera). Each triangle represents an insect order, according to the color legend presented, and it is proportional to the number of branches present. The outgroup (hidden) was human DCR (PDB ID: 5ZAK). The bootstrap values are represented by dark blue circles (minimum 70).

Figure S4. Structural representation and phylogenetic analysis of PASHA-specific domains. Maximum likelihood analysis of the Rhed (**A**) and C-terminal domain (CTD) domains (**B**) present only in PASHA proteins from species belonging to the five insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera). Each triangle represents an insect order, according to the color legend presented, and is proportional to the number of branches present. The outgroup (hidden) used in (**A**) was the Rhed domain from human PASHA (Q8WIQ5), and in (**B**) was the CTD domain from human PASHA (PDB ID: 5B16). The bootstrap values are represented by dark blue circles (minimum 70). Still in (**A**), it is possible to observe the structural model of the dimerization domain of human DGCR8 (PDB ID: 3LE4), which is inserted in the Rhed domain, and (**B**) the structural model of CTD from human PASHA (PDB ID: 5B16). There is no structural model available of complete Rhed.

Figure S5. Alignment of dsrm domain from DROSHA proteins. Structural protein sequence alignment of dsrm domain from DROSHA belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in

red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the dsrm domain presented here come from human DROSHA (PDB ID: 5B16). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S6. **Alignment of Dicer Dimer domain from DCR1 proteins.** Structural protein sequence alignment of Dicer Dimer domain from DCR1 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the Dicer Dimer domain presented here come from *A. thaliana* DCL protein (PDB ID: 2KOU). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S7. **Alignment of Dicer Dimer domain from DCR2 proteins.** Structural protein sequence alignment of Dicer Dimer domain from DCR2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera, and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the Dicer Dimer domain presented here come from *A. thaliana* DCL protein (PDB ID: 2KOU). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S8. **Alignment of first dsrm domain (dsrm-I) from LOQS proteins.** Structural protein sequence alignment of first dsrm domain (dsrm-I) from LOQS belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high

evolutionary rate values. The secondary structure of the dsrm domain presented here come from *D. melanogaster* LOQS (PDB ID: 5NPG). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S9. **Alignment of first dsrm domain (dsrm-I) from PASHA proteins.** Structural protein sequence alignment of first dsrm domain (dsrm-I) from PASHA belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the dsrm domain presented here come from human DGCR8 (PDB ID: 1X47). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S10. **Alignment of first dsrm domain (dsrm-I) from R2D2 proteins.** Structural protein sequence alignment of first dsrm domain (dsrm-I) from R2D2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the dsrm domain presented here come from *D. melanogaster* LOQS (PDB ID: 5NPG). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S11. **Alignment of dsrm domain from DCR1 proteins.** Structural protein sequence alignment of dsrm domain from DCR1 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the dsrm domain presented here come from mouse Dicer (PDB ID:

3C4B). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S12. **Alignment of dsrm domain from DCR2 proteins.** Structural protein sequence alignment of dsrm domain from DCR2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the dsrm domain presented here come from mouse Dicer (PDB ID: 3C4B). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S13. **Alignment of second dsrm domain (dsrm-II) from LOQS proteins.** Structural protein sequence alignment of second dsrm domain (dsrm-II) from LOQS belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the dsrm domain presented here come from human TAR (PDB ID: 2CPN). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S14. **Alignment of second dsrm domain (dsrm-II) from PASHA proteins.** Structural protein sequence alignment of second dsrm domain (dsrm-II) from PASHA belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the dsrm domain presented here is come from human DGCR8 (PDB ID: 2YT4). The numbering at the top of the

alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S15. Alignment of second dsrm domain (dsrm-II) from R2D2 proteins. Structural protein sequence alignment of second dsrm domain (dsrm-II) from R2D2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the dsrm domain presented here come from human TAR (PDB ID: 2CPN). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S16. Alignment of Staufen C-terminal domain from LOQS proteins. Structural protein sequence alignment of Staufen C-terminal domain from LOQS belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure presented here come from *D. melanogaster* LOQS (PDB ID: 4X8W). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S17. Alignment of PAZ domain from AGO1 proteins. Structural protein sequence alignment of PAZ domain from AGO1 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the PAZ domain presented here come from *D. melanogaster* AGO1 (PDB ID: 1R4K). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S18. Alignment of PAZ domain from AGO2 proteins. Structural protein sequence alignment of PAZ domain from AGO2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the PAZ domain presented here come from *D. melanogaster* AGO2 (PDB ID: 1T2R). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S19. Alignment of PAZ domain from DCR1 proteins. Structural protein sequence alignment of PAZ domain from DCR1 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the PAZ domain presented here come from human DCR (PDB ID: 4NGD). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S20. Alignment of PAZ domain from DCR2 proteins. Structural protein sequence alignment of PAZ domain from DCR2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the PAZ domain presented here come from human DCR (PDB ID: 4NGD). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S21. Alignment of PAZ domain (PAZ-like) from DROSHA proteins. Structural protein sequence alignment of PAZ-like domain from DROSHA belonging to species of five

different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. There is no PDB model available to this domain. The numbering at the top of the alignment refers to the position of each amino acid residue in the first sequence presented.

Figure S22. **Alignment of Platform domain from DCR1 proteins.** Structural protein sequence alignment of Platform domain from DCR1 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the Platform domain presented here come from human DCR (PDB ID: 5ZAK). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S23. **Alignment of Platform domain from DCR2 proteins.** Structural protein sequence alignment of Platform domain from DCR2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the Platform domain presented here come from human DCR (PDB ID: 5ZAK). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S24. **Alignment of Platform domain from DROSHA proteins.** Structural protein sequence alignment of Platform domain from DROSHA belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the

similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the Platform domain presented here come from human DROSHA (PDB ID: 5B16). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S25. **Alignment of first Ribonuclease III domain (RIIID-I) from DCR1 proteins.** Structural protein sequence alignment of RIIID-I domain from DCR1 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted with a red background and the similarity only with red letter. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the RIIID-I domain presented here come from human DCR (PDB ID: 5ZAK). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used. *LH* - loop helix.

Figure S26. **Alignment of first Ribonuclease III domain (RIIID-I) from DCR2 proteins.** Structural protein sequence alignment of RIIID-I domain from DCR2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted with a red background and the similarity only with red letter. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the RIIID-I domain presented here come from human DCR (PDB ID: 5ZAK). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used. *LH* - loop helix.

Figure S27. **Alignment of first Ribonuclease III domain (RIIID-like) from DROSHA proteins.** Structural protein sequence alignment of RIIID-like domain from DROSHA belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment,

the identity is highlighted with a red background and the similarity only with red letter. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the RIIID-like domain presented here come from human DROSHA (PDB ID: 5B16). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used. *BH* - bump helix.

Figure S28. **Alignment of second Ribonuclease III domain (RIIID-II) from DCR1 proteins.** Structural protein sequence alignment of RIIID-II domain from DCR1 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted with a red background and the similarity only with red letter. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the RIIID-II domain presented here come from human DCR1 (PDB ID: 5ZAK). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used. *LH* - loop helix.

Figure S29. **Alignment of second Ribonuclease III domain (RIIID-II) from DCR2 proteins.** Structural protein sequence alignment of RIIID-II domain from DCR2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted with a red background and the similarity only with red letter. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the RIIID-II domain presented here come from human DCR1 (PDB ID: 5ZAK). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used. *LH* - loop helix.

Figure S30. **Alignment of second Ribonuclease III domain (RIIID-II) from DROSHA proteins.** Structural protein sequence alignment of RIIID-II domain from DROSHA belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and

Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted in yellow and the similarity in red. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the RIIID-II domain presented here come from human DROSHA (PDB ID: 5B16). The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S31. <u>Alignment of Helicase domain from DCR1 proteins.</u> Structural protein sequence alignment of Helicase domain from DCR1 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted with a red background and the similarity only with red letter. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the Helicase domain presented here come from human RIG-I (PDB ID: 5E3H), and its four functional subdomains are highlighted: *olive green* - DEAD/ResIII (Hel1); *red* - Hel2i; dark blue - Helicase C (Hel2); and *brown* - Pincer. All the canonical (Q, I, Ia, Ib, Ic, II, III, IV, IVa, V, Va and VI) and non-canonical (IVb) conserved-sequence motifs, important to ATP binding and hydrolysis (*red* circles), RNA binding (*yellow* circles), and in the communication between ATP and RNA binding sites (*blue* circles) were identified. In addition, it is important to highlight: *black* circles - first layer of hydrophobic residues around those important in the communication between ATP and RNA binding sites (blue circles); *gray* circles - first layer of hydrophobic waste around the same waste represented by *blue* circles. The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Figure S32. <u>Alignment of Helicase domain from DCR2 proteins.</u> Structural protein sequence alignment of Helicase domain from DCR2 belonging to species of five different insect orders (Coleoptera, Diptera, Hemiptera, Hymenoptera and Lepidoptera), represented by colors according to legend. In the alignment, the identity is highlighted with a red background and the similarity only with red letter. At the top, the area of the trend curve of evolutionary rate is represented in dark blue (with highest value highlighted). The most variable regions have high evolutionary rate values. The secondary structure of the Helicase domain presented here come from human RIG-I (PDB ID: 5E3H), and its four functional subdomains are highlighted: *olive green* - DEAD/ResIII (Hel1); *red* - Hel2i; dark blue - Helicase C (Hel2); and *brown* - Pincer. All the canonical (Q, I, Ia, Ib, Ic, II, III, IV, IVa, V, Va and VI) and non-canonical (IVb) conserved-sequence motifs, important to ATP binding and hydrolysis (*red* circles), RNA binding (*yellow* circles), and in the communication between ATP and RNA binding sites (*blue* circles) were identified. In addition, it is important to highlight: *black* circles - first layer of hydrophobic residues around those important in the communication between ATP and RNA binding sites (blue circles); *gray* circles - first layer of hydrophobic waste around the same waste represented by *blue* circles. The numbering at the top of the alignment refers to the position of each amino acid residue in the sequence of the PDB model used.

Supplementary Alignments (SA; FASTA files). Structural-based alignment of each analyzed domain presented in Supplementary Figures: ***SA1*** - dsrm (DROSHA) (Figure S5); ***SA2*** - Dicer Dimer (DCR1) (Figure S6); ***SA3*** - Dicer Dimer (DCR2) (Figure S7); ***SA4*** - dsrm-I (LOQS) (Figure S8); ***SA5*** - dsrm-I (PASHA) (Figure S9); ***SA6*** - dsrm-I (R2D2) (Figure S10); ***SA7*** - dsrm (DCR1) (Figure S11); ***SA8*** - dsrm (DCR2) (Figure S12); ***SA9*** - dsrm-II (LOQS) (Figure S13); ***SA10*** - dsrm-II (PASHA) (Figure S14); ***SA11*** - dsrm-II (R2D2) (Figure S15); ***SA12*** - Staufen C (LOQS) (Figure S16); ***SA13*** - PAZ (AGO1) (Figure S17); ***SA14*** - PAZ (AGO2) (Figure S18); ***SA15*** - PAZ (DCR1) (Figure S19); ***SA16*** - PAZ (DCR2) (Figure S20); ***SA17*** - PAZ (DROSHA) (Figure S21); ***SA18*** - Platform (DCR1) (Figure S22); ***SA19*** - Platform (DCR2) (Figure S23); ***SA20*** - Platform (DROSHA) (Figure S24); ***SA21*** - RIIID-I (DCR1) (Figure S25); ***SA22*** - RIIID-I (DCR2) (Figure S26); ***SA23*** - RIIID-Like (DROSHA) (Figure S27); ***SA24*** - RIIID-II (DCR1) (Figure S28); ***SA25*** - RIIID-II (DCR2) (Figure S29); ***SA26*** - RIIID-II (DROSHA) (Figure S30); ***SA27*** - Helicase (DCR1) (Figure S31); and ***SA28*** - Helicase (DCR2) (Figure S32).

Supplementary Phylogenetic Trees (SP; TRE files). Maximum likelihood phylogenetic trees presented in Figure 1 in a format used in almost phylogenetic tree viewers: ***SP1*** - AGO (Figure 1A); ***SP2*** – RIIID (DCR1-2 and DROSHA) (Figure 1B); ***SP3*** - LOQS and R2D2 (Figure 1C); and ***SP4*** - PASHA (Figure 1D).

Supplementary Sequences (SS; PDF file). ***SS1*** - Supplementary insect protein sequences obtained from *in house* assembled transcriptomes (*in silico* transcriptional translation).

Supplementary Tables (XLSX files): ***S1*** - Summary of insect databases analyzed; ***S2*** - Summary of other Metazoa selected proteins.

Supplementary Text (ST; PDF file): ***ST1*** - The miRNA and siRNA pathways in insects: An overview.

5. CHAPTER 02: IMPLICATIONS OF ETHYLENE BIOSYNTHESIS AND SIGNALING IN SOYBEAN DROUGHT STRESS TOLERANCE

Published in	**:**	BMC Plant Biology
Impact Factor	**:**	3.49
DOI	**:**	10.1186/s12870-015-0597-z

Arraes *et al. BMC Plant Biology* (2015) 15:213
DOI 10.1186/s12870-015-0597-z

BMC Plant Biology

RESEARCH ARTICLE Open Access

Implications of ethylene biosynthesis and signaling in soybean drought stress tolerance

Fabricio Barbosa Monteiro Arraes[1,2], Magda Aparecida Beneventi[1,2], Maria Eugenia Lisei de Sa[2,4], Joaquin Felipe Roca Paixao[2,3], Erika Valeria Saliba Albuquerque[2], Silvana Regina Rockenbach Marin[5], Eduardo Purgatto[6], Alexandre Lima Nepomuceno[5] and Maria Fatima Grossi-de-Sa[2,7*]

Abstract

Background: Ethylene is a phytohormone known for inducing a triple response in seedlings, leaf abscission and other responses to various stresses. Several studies in model plants have evaluated the importance of this hormone in crosstalk signaling with different metabolic pathways, in addition to responses to biotic stresses. However, the mechanism of action in plants of agricultural interest, such as soybean, and its participation in abiotic stresses remain unclear.

Results: The studies presented in this work allowed for the identification of 176 soybean genes described elsewhere for ethylene biosynthesis (108 genes) and signal transduction (68 genes). A model to predict these routes in soybean was proposed, and it had great representability compared to those described for *Arabidopsis thaliana* and *Oryza sativa*. Furthermore, analysis of putative gene promoters from soybean gene orthologs permitted the identification of 29 families of *cis*-acting elements. These elements are essential for ethylene-mediated regulation and its possible crosstalk with other signaling pathways mediated by other plant hormones. From genes that are differentially expressed in the transcriptome database, we analyzed the relative expression of some selected genes in resistant and tolerant soybean plants subjected to water deficit. The differential expression of a set of five soybean ethylene-related genes (*MAT*, *ACS*, *ACO*, *ETR* and *CTR*) was validated with RT-qPCR experiments, which confirmed variations in the expression of these soybean target genes, as identified in the transcriptome database. In particular, two families of ethylene biosynthesis genes (*ACS* and *ACO*) were upregulated under these experimental conditions, whereas *CTR* (involved in ethylene signal transduction) was downregulated. In the same samples, high levels of ethylene production were detected and were directly correlated with the free fraction levels of ethylene's precursor. Thus, the combination of these data indicated the involvement of ethylene biosynthesis and signaling in soybean responses to water stress.

(Continued on next page)

* Correspondence: fatima.grossi@embrapa.br
[2]Embrapa Genetic Resources and Biotechnology, PqEB, Av. W5-Norte, Postal Code 02372, CEP 70770–910, Brasilia, DF, Brazil
[7]Catholic University of Brasilia, SGAN 916, Modulo B, Av. W5, Asa Norte, CEP 70790–160, Brasilia, DF, Brazil
Full list of author information is available at the end of the article

RESUMO

O etileno é um fitormônio conhecido pela indução da resposta tripla em plântulas (inibição do alongamento caulinar, espessamento do caule e hábito de crescimento horizontal ou perda de sensibilidade gravitrópica), além de ser associado com abscisão foliar e respostas a vários estresses. Diversos estudos com plantas modelo avaliaram a importância desse hormônio no *crosstalk* com diferentes vias metabólicas, principalmente em respostas à estresses bióticos. No entanto, o mecanismo de ação em plantas de interesse agrícola, como a soja, e sua participação em estresses abióticos ainda não foram completamente elucidados.

Os estudos aqui apresentados permitiram identificar 176 genes de soja possivelmente envolvidos tanto com a biossíntese de etileno (108 genes) e a transdução de sinal por ele mediada (68 genes). Foi proposto um modelo para essas rotas na soja, que apresentava grande representatividade em comparação com o que já foi descrito para *Arabidopsis thaliana* e *Oryza sativa*. Associado a isso, foram analisadas as possíveis regiões promotoras dos ortólogos identificados, onde foram identificadas 29 famílias de elementos de *cis*-elementos, dentre os quais se destacam aqueles responsivos à seca. Esses elementos são essenciais para a regulação mediada por etileno e sua possível interferência com outras vias de sinalização mediadas por outros hormônios vegetais.

A partir da análise de bancos de dados de transcritômica, foi possível identificar *in silico* genes diferencialmente expressos, cuja a expressão relativa foi avaliada posteriormente por RT-qPCR em cultivares de soja tolerante (Embrapa 48) e suscetível (BR16) à seca, ambas submetidas à déficit hídrico (sistema hidropônico). Desta forma, a expressão diferencial *in silico* de um conjunto de cinco genes relacionados com a biossíntese de etileno e a transdução de sinal por ele mediada (MAT, ACS, ACO, ETR e CTR) foi validada, confirmando o perfil de expressão predito para eles nas condições analisadas. Em particular, duas famílias de genes de biossíntese de etileno (ACS e ACO) foram reguladas positivamente sob essas condições experimentais, enquanto a CTR (envolvida na transdução de sinal de etileno) foi regulada negativamente. Nas mesmas amostras, altos níveis de produção de etileno foram detectados, o que foi diretamente correlacionado com os níveis de fração livre do principal precursor deste fitormônio (ACC).

Assim, a análise *in silico*, combinada com a quantificação da produção de etileno (e seu precursor), associadas com experimentos de quantificação dos níveis de expressão gênica, permitiu uma melhor compreensão da importância do etileno em nível molecular em soja, bem

como de seu papel na resposta a abióticos. Em resumo, todos os dados apresentados sugerem que as respostas da soja ao estresse hídrico podem ser reguladas pelo *crosstalk* entre diferentes vias de sinalização, que podem envolver diferentes fitormônios, como auxinas, ABA e ácido jasmônico. A integração destes dados também pode contribuir para o desenvolvimento de estratégias e novos ativos biotecnológicos que vissem o aumento da tolerância à seca.

ABSTRACT

Ethylene is a phytohormone known for inducing a triple response in seedlings, leaf abscission and other responses to various stresses. Several studies in model plants have evaluated the importance of this hormone in crosstalk signaling with different metabolic pathways, in addition to responses to biotic stresses. However, the mechanism of action in plants of agricultural interest, such as soybean, and its participation in abiotic stresses remain unclear.

The studies presented in this work allowed for the identification of 176 soybean genes described elsewhere for ethylene biosynthesis (108 genes) and signal transduction (68 genes). A model to predict these routes in soybean was proposed, and it had great representability compared to those described for *Arabidopsis thaliana* and *Oryza sativa*. Furthermore, analysis of putative gene promoters from soybean gene orthologs permitted the identification of 29 families of *cis*-acting elements. These elements are essential for ethylene-mediated regulation and its possible crosstalk with other signaling pathways mediated by other plant hormones.

From genes that are differentially expressed in the transcriptome database, we analyzed the relative expression of some selected genes in resistant and tolerant soybean plants subjected to water deficit. The differential expression of a set of five soybean ethylene-related genes (*MAT*, *ACS*, *ACO*, *ETR* and *CTR*) was validated with RT-qPCR experiments, which confirmed variations in the expression of these soybean target genes, as identified in the transcriptome database. In particular, two families of ethylene biosynthesis genes (*ACS* and *ACO*) were upregulated under these experimental conditions, whereas *CTR* (involved in ethylene signal transduction) was downregulated. In the same samples, high levels of ethylene production were detected and were directly correlated with the free fraction levels of ethylene’s precursor. Thus, the combination of these data indicated the involvement of ethylene biosynthesis and signaling in soybean responses to water stress.

The *in silico* analysis, combined with the quantification of ethylene production (and its precursor) and RT-qPCR experiments, allowed for a better understanding of the importance of ethylene at a molecular level in this crop as well as its role in the response to abiotic stresses. In summary, all of the data presented here suggested that soybean responses to water stress could be regulated by a crosstalk network among different signaling pathways, which might involve various phytohormones, such as auxins, ABA and jasmonic acid. The integration of *in silico* and physiological data could also contribute to the application of biotechnological strategies to the development of improved cultivars with regard to different stresses, such as the isolation of stress-specific plant promoters.

Keywords: Soybean, ethylene, biosynthesis, signaling, drought, crosstalk, stress, transcriptome

INTRODUCTION

Phytohormones are organic compounds that exist naturally in plants and that even in low concentrations, orchestrate a broad range of physiological processes, including growth and development, as well as responses to abiotic and biotic stresses (Gerashchenkov and Rozhnova, 2013). These hormones overlap signal transduction pathways or gene expression profiles by rapid induction or by preventing the degradation of transcriptional regulators (Bari and Jones, 2009; Forcat et al., 2008; Kaya et al., 2009; Santner and Estelle, 2010).

Among all of the described phytohormones, ethylene, a naturally occurring triple response growth regulator (shoot elongation, stem thickening and horizontal growth habit) in seedlings, has been studied since ancient times (Doubt, 1917). Ethylene is also involved in leaf abscission, fruit ripening and senescence (Doubt, 1917; Nath et al., 2006) as well as seed germination, growth of adventitious roots under flooding conditions, epinasty stimulation, inhibition of shoot growth and stomatal closing and flowering (Trusov and Botella, 2006; Wilmowicz et al., 2008). Moreover, it is involved in a wide variety of stresses, including wounding, pathogen attack, flooding, drought, hypoxia, and temperature shifts (Bleecker, 1999; Yang and Hoffman, 1984).

Ethylene biosynthesis is derived from the amino acid methionine provided by the Yang cycle (Miyazaki and Yang, 1987), in which the precursor S-adenosylmethionine (AdoMet or SAM) is synthesized from ATP and methionine by S-adenosylmethionine synthetase (SAMS; EC 2.5.1.6) (Roje, 2006). AdoMet is then converted into 1-aminocyclopropane-1-carboxylic acid (ACC) and 5-methylthioadenosine (MTA) by the enzyme 1-aminocyclopropane-1-carboxylase synthase (ACS, EC 4.4.1.14) (Roje, 2006). MTA is recycled through a series of Yang cycle reactions back to methionine (Argueso et al., 2007).

Active ACSs are encoded by eight genes in *Arabidopsis thaliana*, and at least one encodes a catalytically inactive ACS (AtACS1) (Liang et al., 1992; Yamagami et al., 2003). Based on the sequence present in its C-terminal region, these proteins can be divided into three main groups: *type I* proteins, which are the targets for phosphorylation by mitogen-activated protein kinase 3 and/or 6 (AtMPK3-6; EC 2.7.11.24) (Liu and Zhang, 2004) as well as by calcium-dependent protein kinase (AtCDPK2; CDPK or CPK; EC 2.7.11.1); *type II* proteins, which show phosphorylation sites for only CPK (Tatsuki and Mori, 2001); and *type III* proteins, which have the C-terminal portion greatly reduced and do not present phosphorylation sites for either kinase. Furthermore, the ACSs can be regulated by putative endogenous signal receptors (*i.e.*, phytohormones) and/or intracellular accumulation of secondary metabolites, such as calcium. In the absence of an endogen signal, type II ACSs are degraded by 26S proteasome. This degradation is mediated by ETO proteins (ethylene overproducer*)* and EOL (ETO-like), which are members of specific plant proteins with E3 ubiquitin ligase domain (Wang et al., 2004). This process activates kinase protein signaling, which culminates in the stabilization of type II ACSs. Furthermore, MPK3-6 kinases are able to phosphorylate the C-terminal of type I ACSs, which preserve and stabilize their degradation via the 26S proteasome pathway, thereby increasing the production of ethylene and inducing other ethylene-dependent signaling pathways (Hahn and Harter, 2009).

The enzyme directly responsible for the ethylene biosynthesis is 1-aminociclopropane-1-acid carboxylic oxidase (ACO or EFE - ethylene forming enzyme; EC 1.14.17.4), which converts ACC into this plant hormone (Hegg and Que, 1997).

Several reports have suggested that the ACC metabolite could combine with other organic molecules. Different studies have demonstrated that the ACC N-malonyzation pathway in various plant tissues is involved in the regulation of ethylene production, wherein the conjugate 1-malonyl-ACC (MACC) is formed by 1-aminocyclopropane-1-acid carboxylic acid-N-malonyltransferase, an enzyme that has been purified from plant protein extracts but without reference to its respective gene (Amrhein et al., 1984; Hoffman et al., 1982). In addition to MACC formation through a metabolic route, ACC can also be conjugated in the form of 1-glutamyl-ACC (GACC) in a reaction that is catalyzed by γ-glutamyl transpeptidase (GGT; EC 2.3.2.2) (Martin et al., 1995).

Another possible ACC metabolic pathway is the reaction catalyzed by the enzyme ACC deaminase (ACD; EC 3.5.99.7), a protein that degrades ACC into oxobutyrate (or OXB; 2-oxobutanoate) and ammonia (NH_3), thus decreasing the levels of ACC that are available for ethylene production (Glick, 2005; Klee et al., 1991). The *ACD* gene was first identified in *A. thaliana* and *Populus,* and studies of tomato plants have shown that ACD activity varies during fruit ripening and that its peak activity coincides with the reduction in ethylene synthesis (McDonnell et al., 2009).

The classic routes of ethylene intracellular signal transduction, initially described in *A. thaliana*, are triggered by the gas interaction with membrane receptors (encoded by *ETR* genes - *ethylene receptor*) and the modulation of CTR1 (constitutive triple response – MKKK; EC 2.7.11.1) activity to regulate the expression of several genes, such as *EIN3/EIL* (*ethylene* *insensitive 3*; *EIN3-like*). Both receptors and CTR1 function as negative regulators of the signal transduction pathway in the absence of ethylene. The kinase CTR1 phosphorylates the EIN2 (ethylene insensitive 2) C-terminal domain, allowing for the degradation of this protein. ETP1 and ETP2 (EIN2 targeting protein) play important roles in EIN2 proteolysis. These proteins, which have F-box domains, interact with the conserved EIN2 C-terminal domain that was previously phosphorylated by CTR1. Thus, in the absence of ethylene, the phosphorylated EIN2 C-terminal domain is ubiquitinated and then degraded by the 26S proteasome (Qiao et al., 2009). However, in the presence of ethylene, instead of being phosphorylated, the EIN2 domain is cleaved and transported to the nucleus to stimulate EIN3/EIL activity by repressing EBF (EIN3 binding F-box protein). Thus, EIN3/EINL induce the transcription of target genes, mainly the AP2/ERF transcription factor superfamily (Ju et al., 2012). Earlier studies have also suggested an EIN3/EIL activation route independent of EIN2 and CTR via a phosphorylation cascade of kinase proteins, MKK4-5-9 (EC 2.7.12.2) → MPK3-6, which is mitogen activated (Hahn and Harter, 2009; Stepanova and Alonso, 2009; Yoo et al., 2008). In the presence of a signal, EIN3/EIL transcription factors are phosphorylated by MPK3-6 and do not interact with the F-box protein EBF (EIN3 binding F-box protein), preventing their degradation through the 26S proteasome. Thus, these factors that accumulate in the nucleus interact with target gene promoters and trigger different ethylene responses (Stepanova and Alonso, 2009). In addition, the exoribonuclease 5'-3' EIN5 (EC 3.1.1.3.-), another positive regulator, promotes EBF mRNA decrease and thereby increases EIN3/EIL protein levels in the nucleus (Olmedo et al., 2006).

Ethylene signal transduction triggers substantial changes in the gene expression of plant cells. Promoter region analyses of the genes induced by ethylene led to the identification of *cis*-acting elements as well as the *trans*-acting protein EREBP (ethylene responsive element binding protein) family, which interacts with DNA and ERFs (ethylene response factors) (Deikman et al., 1998; Leubner-Metzger et al., 1998; Ohme-Takagi and Shinshi, 1995). Recent studies have demonstrated that EIN3/EIL are ERF1 (ethylene response factor 1) gene activators, constituting an ERF family member that establishes a hierarchy of ethylene-mediated signaling (Solano et al., 1998). The homodimers EIN3/EIL interact with *cis*-acting elements in the *ERF1* promoter region that once transcribed and translated, interact with other *cis*-acting elements present in the promoter regions of target genes (Solano et al., 1998). EIN3 can induce transcription not only of *ERF1* but also of other members of the AP2/ERF transcription factor superfamily (Vandenbussche et al., 2012).

The mechanism underlying environmental stress tolerance has been extensively *studied* in model plants in attempts to determine its impact on agriculture (Ding et al., 2014). The metabolic pathways induced under drought in *A. thaliana* have been associated with abscisic acid (ABA)-dependent and ABA-independent pathways governing drought-inducible gene expression (Guimarães-Dias et al., 2012; Shinozaki and Yamaguchi-Shinozaki, 2007) as well as the existence of an interconnection between both signaling pathways (Kizis and Pagès, 2002; Yamaguchi-Shinozaki and Shinozaki, 1994). Furthermore, advanced ABA and ethylene signaling research has revealed that under stress, both hormones act antagonistically among yield-impacting processes (Wilkinson et al., 2012).

Although ethylene has been extensively studied in the plant senescence process, its role during drought-induced senescence is less well known. It has been demonstrated that under drought conditions, ethylene caused leaf abscission and consequently reduced water loss (Oh et al., 1997). Under water deficit, ethylene production was paralleled by an increase and subsequent decrease in ACC, suggesting that water stress induced the *de novo* synthesis of ACC synthase, which is the rate-controlling enzyme along the pathway of ethylene biosynthesis. Moreover, ethylene and its metabolic process are important for activating plant responses to flooding and water deficit (Habben et al., 2014; Voesenek and Bailey-Serres, 2009). It activates a signal transduction network that culminates in the synthesis of several transcription factors that regulate gene activation/repression during stress, such as ERF1 (Fujimoto et al., 2000; Shinozaki and Yamaguchi-Shinozaki, 2007; Xu et al., 2007).

Despite important insights having been reported in ethylene signaling pathways, the available studies have not addressed the soybean *(Glycine max* [L.] Merrill), an economically important crop. This commodity is the second largest source of edible oil and the most important high-quality vegetable protein for feeding both humans and animals worldwide. However, deficiency in water supply can negatively impact this crop, reducing yields and posing threats to farmers and food production in several countries (Statista - *http://www.statista.com/statistics/263937/vegetable-oils-global-consumption*).

Considering the important position that soybean occupies in the Brazilian economy, the second largest world soybean producer, *the Brazilian Soybean Genome Consortium* (GENOSOJA) was created to identify the genes related to different biotic and abiotic stresses. Because there have been no reports concerning ethylene molecular mechanisms in soybean, this work described the ethylene metabolic pathway *in silico* in the soybean genome using various databases. The gene expression profile data obtained from the GENOSOJA database was validated by RT-qPCR experiments, and determinations of free ACC levels and ethylene production in susceptible and tolerant soybean genotypes under water deficit conditions were also performed. Moreover, transcriptional regulation was studied by analyzing putative *cis*-acting elements present in the possible promoters. These data allowed for the inference of the first accurate *in silico* models for soybean ethylene biosynthesis and signaling, which facilitated a better understanding of the molecular mechanisms involved in this important phytohormone.

RESULTS AND DISCUSSION

In silico reconstruction of soybean ethylene molecular models

To evaluate the influence of ethylene in soybean water stress response, it was necessary to reconstruct the metabolic pathways to improve those available in public databases. Hence, we conducted an extensive search in the crop genome for genes previously associated with ethylene biosynthesis and signal transduction. Thus, a total of 322 genes were analyzed, of which 146 corresponded to model plants (74 from *Arabidopsis thaliana* and 72 from *Oryza sativa*) and 176 to *Glycine max* (Table 1). All of the soybean genes were mapped on their respective chromosomes (Figure S1 in Additional File 1) and were functionally annotated (Figure S2 in Additional File 1). The proteins identified in model plants *A. thaliana* (Tables S1 and S2 in Additional File 2) and *O. sativa* (Tables S3 and S4 in Additional File 2) as well as in *Glycine max* (Tables S5 and S6 in Additional File 2) were thoroughly characterized *in silico*, making possible the identification of the main characteristic domains. The soybean orthologous proteins in *A. thaliana* and *O. sativa* were investigated by BBH (best bidirectional hit) analysis, comparing the three species databases (Figure S3 in Additional File 1; Tables S7 and S8 in Additional File 2). According to these data (see Additional File 3), accurate soybean models of ethylene biosynthesis and signal transduction have been proposed. The putative soybean proteins that participate in the metabolic pathways involved in ethylene biosynthesis and signaling mediated by this molecule are highly conserved, with domains that have already been described for their homologs in model organisms.

The BBH experiment suggested a higher phylogenetic proximity of soybean to *A. thaliana*, corroborating that both are classified as dicotyledonous, although significant portions of these proteins are conserved in all three species. The ontological analysis indicated the same conclusion, showing that both function and molecular processes as well as the cell localization of these proteins were similar in different species.

Table 1. Ethylene biosynthesis and signal transduction gene summary in different plants

Group	**Number of Genes**		
	Arabidopsis thaliana	***Glycine max***	***Oryza sativa***
Biosynthesis	44	108	38
Signal Transduction	30	68	34
Total	**74**	**176**	**72**

Soybean ethylene biosynthesis model

Based on the model for ethylene biosynthesis in *A. thaliana*, the 108 genes of soybean related to this metabolic route were divided into three groups: Yang cycle genes (21.3 %); ethylene biosynthesis (44.4 %); and ACC conjugation or degradation (34.3 %) (Table S5).

Pommerrenig et al. (2011) described a model for methionine recycling reactions through the Yang cycle in *Plantago* and *A. thaliana* (Pommerrenig et al., 2011). Based on this work, we proposed an *in silico* model for this route in soybean, in which the homologs for all components were identified: MTN (5-methylthioadenosine nucleosidase; EC 3.2.2.16), MTK (5-methylthioribose kinase; EC 2.7.1.100), MTI (5-methylthioribose-1-phosphate isomerase; EC 5.3.1.23), DEP (dehydratase-enolase-phosphatase complex; EC 4.2.1.109 and 3.1.3.77), ARD (acireductone dioxigenase; EC 1.13.11.53 and 1.13.11.54) and AAT (amino acid transferase) or ASP (aspartate aminotransferase) (EC 2.6.1.1) (Figure 1). Each of the identified enzymes had at least one ortholog in *A. thaliana* and/or *O. sativa* identified *in silico* through the BHH experiment, suggesting plausible conservation of the pathway in different plant species. The first enzyme in the biosynthesis pathway, MAT (methionine adenosyltransferase) or SAMS, is responsible for the production of the AdoMet used for ethylene production and also for lignin and polyamine synthesis (Amthor, 2003; Yang and Hoffman, 1984). Among the eleven MAT proteins in soybean, five were BHH-positive with possible orthologs in *A. thaliana* and/or *O. sativa*.

Subsequently, the classification of 21 soybean ACSs was proposed by Tucker et al. (2010), who reported phylogenetic relationships with similar ACSs in *A. thaliana*, suggesting that they are expressed when the plant is infected by the nematode *Heterodera glycines* (Tucker et al., 2010). In our work, we studied the phylogenetic relationships of ACS amino acids residues between *G. max* and *A. thaliana* and also with its homologues in *O. sativa.*

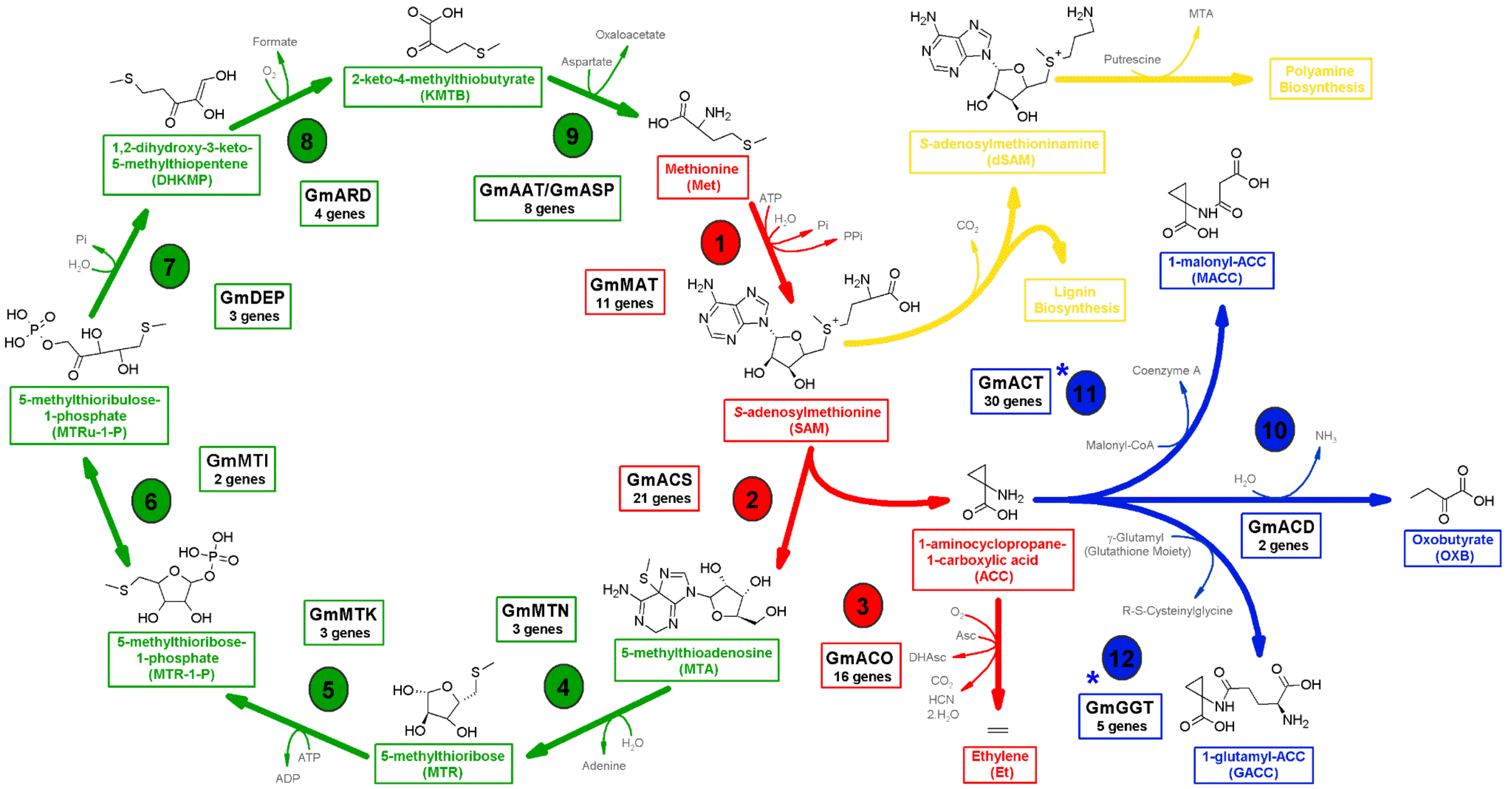

Figure 1. <u>Soybean model of ethylene biosynthesis.</u> *In silico* experiments identified 108 proteins that could be involved directly or indirectly in soybean ethylene biosynthesis. In this putative model: **green** - Yang cycle; **red** - ethylene biosynthesis; **blue** - ACC (1-aminocyclopropane-1-carboxylic acid) degradation and conjugation with other metabolites (malonyl and glutamyl groups); **yellow** - lignin and polyamine biosynthesis (example of S-adenosylmethionine production deviation for other metabolic pathways). Enzymes: **1 - MAT** (methionine adenosyltransferase) or *SAMS* (S-adenosylmethionine synthetase); **2 - ACS** (1-aminocyclopropane-1-carboxylic acid synthase); **3 - ACO** (1-aminocyclopropane-1-carboxylic acid oxidase); **4 - MTN** (5-methylthioadenosine nucleosidase); **5 - MTK** (5-methylthioribose kinase); **6 - MTI** (5-methylthioribose-1-phosphate isomerase);

Figure 1. (cont.) 7 - DEP (dehydratase-enolase-phosphatase complex); **8 - ARD** (acireductone dioxygenase); **9 - AAT** (amino acid transferase) or **ASP** (aspartate aminotransferase); **10 - ACD** (1-aminocyclopropane-1-carboxylic acid deaminase); **11 - ACT** (acyltransferase; N-malonyltransferase); **12 - GGT** (γ-glutamyltranspeptidase). Other abbreviations: *Asc* - ascorbate; *DHAsc* - dihydroxyascorbate; *HCN* - hydrogen cyanide. The **blue asterisks** (*) present in numbers **11** and **12** indicate enzymes that could be candidates to play the roles described in the model, but their functions described *in vitro* and *in vivo* are not primarily associated with these metabolic pathways. Each enzyme is represented by a generic name (see Table S5 in Additional File 2).

We also determined *in silico* the possible phosphorylation sites of the respective kinases (Figure S4 in Additional File 1). The distribution of the sequences is similar to that presented by Tucker (2010) because they are distributed uniformly, indicating high conservation between species. Moreover, although the sequences of GmACS#003, GmACS#013, GmACS#016 and GmACS#019 present high similarity with ACS, they are phylogenetically unrelated to the rest because differences were found in the catalytic domain. Therefore, these sequences were named ACS-like, i.e., belonging to the family of AATs (amino acid transferases). Among the seventeen ACS sequences identified in soybean, six were possible orthologs of *A. thaliana* and/or *O. sativa,* of which two were determined to be type I (GmACS#011 and GmACS#014), two to be type II (GmACS#017 and GmACS#020) and two to be type III (GmACS#006 and GmACS#012) (Table S7 in Additional File 2; Figure S4 in Additional File 1).

Regarding the conversion of ACC into ethylene, sixteen *ACO* genes were identified in the soybean genome, with 6 of them encoding ortholog proteins in *A. thaliana* and/or *O. sativa* (GmACO#004, GmACO#006, GmACO#007, GmACO#008, GmACO#009 and GmACO#014) (Table S7 in Additional File 2).

Furthermore, ACC can also be used in combination with malonyl and glutamyl in the synthesis of MACC (1-malonyl-ACC) and GACC (1-glutamyl-ACC) (Kende, 1993; Martin et al., 1995). We selected thirty possible candidate genes with this function in soybean, based on six *acyltransferases* (*ACT*) from *A. thaliana* and *O. sativa* (Tables S1 and S3 in Additional File 2). Five were considered BBH-positive with *A. thaliana* and/or *O. sativa* (GmACT#003, GmACT#006, GmACT#017, GmACT#020 and GmACT#023) (Table S5 in Additional File 2). It is important to emphasize that although most of the malonyltransferase enzymes play roles in fatty acids, they could also have N-malonyzation activity. Thus, it would be interesting to characterize them *in vitro* and *in vivo* after selecting them *in silico.* With regard to the formation of GACC, five γ-glutamyl transpeptidases (GGTs) were identified in soybean, and two of them (GmGGT#001 and GmGGT#003) were BBH-positive with *A. thaliana* and *O. sativa* (Table S7 in Additional File 2).

Finally, ACC could be the substrate of ACC deaminase (ACD) in soybean because we identified two genes that codified for homologous ACD enzymes in *A. thaliana* (GmACD#001 and GmACD#002), of which only one was BBH-positive (GmACD#001) (Table S7 in Additional File 2).

Model for soybean ethylene-mediated signal transduction

In this work, we identified 68 genes related to ethylene-mediated signal transduction. We found that 38.3 % of the proteins coded by these genes had orthologs in *A. thaliana* and/or *O. sativa* (Table S8 in Additional File 2). The main components of this signal route were represented because 32.4 % were specific receptors (ETR) and proteins important for receptor activity (RTE and RAN), 7.4 % were CTR, 4.4 % were EIN2 proteins, approximately 19.0 % were kinases (CPK, MKK, MPK), 7.4 % were EIN3/EINL transcription factors, 25.0 % were important in proteolysis routes (EBF and ETO), and 4.4 % were orthologs of EIN5 exoribonuclease, which is important for EIN3/EINL activity regulation (Figure 2; Table S6 in Additional File 2).

Four of the five ethylene receptors described in soybean were found to be homologs of ETR1 and ETR2 (subfamily I - GmETR#001, GmETR#003, GmETR#006 and GmETR#007) and of ERS1 and EIN4 (subfamily II - GmETR#002, GmETR#004, GmETR#005, GmETR#008, GmETR#009, GmETR#010 and GmETR#011) (Chang et al., 1993; Hua and Meyerowitz, 1998; Hua et al., 1995; Sakai et al., 1998).

The receptors in soybean have four principal domains similar to those in *A. thaliana*: (i) receptor response regulation domain (PF00072); (ii) *histidine kinase A domain* (PF00512); (iii) *GAF domain* (PF01590); and (iv) *histidine kinase*$^{-}$, *DNA girase B*$^{-}$ and *ATPase-like* (PF02518). The different combinations of these four domains comprise the different families of receptors in soybean. For example, the ETR1 homologs have the four domains in their structure because homologs to ETR2 and EIN4 have only the (i), (ii) and (iii) domains and ERS1 has the (ii), (iii) and (iv) domains.

Regarding canonical ethylene signal transduction, we identified five soybean homologs of CTR1, four of RTE genes, seven RAN transporters and three homologs of EIN2 (GmEIN#002, GmEIN#004 and GmEIN#007) (Figure 2). It is worth mentioning that homologs encoding the ETP proteins could not be found in soybean, suggesting either that other proteins are performing this role or that other mechanisms regulating EIN2 exist but have not yet been

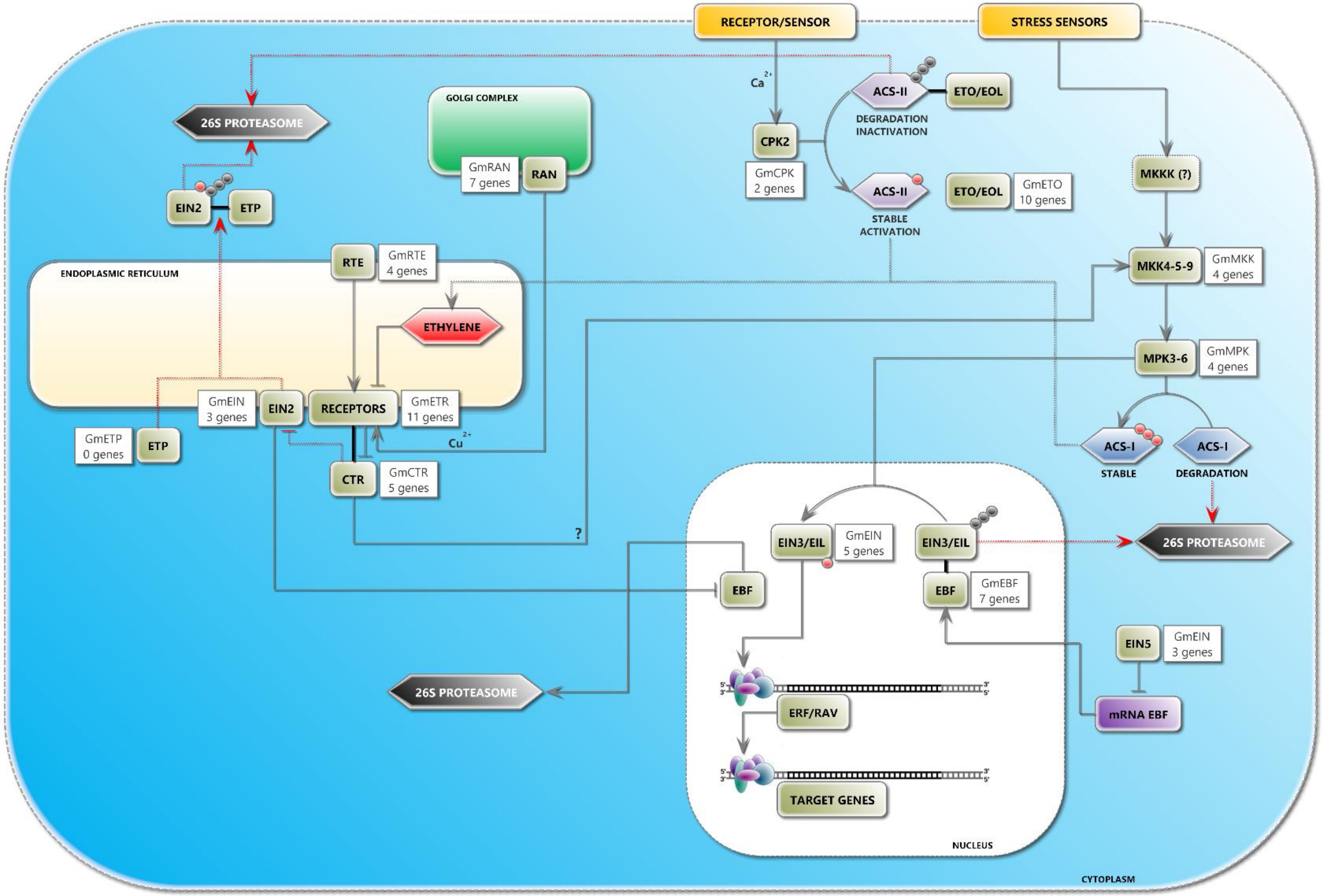

Figure 2. Soybean model of ethylene signal transduction. *In silico* experiments identified 68 proteins that could be involved directly or indirectly in soybean signal transduction initiated by ethylene. In this putative model, **brown rectangles** show the route-identified

Figure 2. (cont.) proteins in *A. thaliana*, and **white rectangles** show the soybean genes that encode proteins homologous to this plant model; **orange rectangles** illustrate membrane sensors that respond to biotic and abiotic stress in addition to receptors/sensors for endogenous signals (i.e., other phytohormones); the **purple rectangle** represents mRNAs related to ETP proteins; the **rectangle with dotted outline** (accompanied by a question mark) represents a protein in this pathway that has not been identified in the studied plants; **blue and purple hexagons** represent ACSs types I and II, respectively; **black and red circles** correspond to ubiquitin and phosphate groups, respectively; **gray arrows** correspond to routes that occur in the presence of ethylene and/or biotic/abiotic stress; **dotted arrows in red and gray** represent pathways that occur in the absence of this hormone and routes that culminate in ethylene biosynthesis, respectively; **black lines** indicate interactions among proteins. Cellular compartments represented: *endoplasmic reticulum* (**beige**), *Golgi complex* (**green**), *nucleus* (**white**) and cytoplasm (**blue**). Symbols: **ACS**: 1-aminocyclopropane-1-carboxylic acid synthase; **CPK** (or **CDPK**): calcium-dependent protein kinase; **CTR**: constitutive triple response protein; **EBF**: EIN3 binding F-Box protein; **EIL**: EIN protein like; **EIN**: ethylene insensitive; **EOL**: ETO protein like; **ERF**: ethylene response factor; **ETP**: EIN2 targeting protein; **ETO**: ethylene overproducer; **MKKK** (or **MAPKKK**): MAP kinase kinase kinase; **MKK** (or **MAPKK**): MAP kinase kinase; **MPK** (or **MAPK**): mitogen-activated protein kinase; **RAN**: responsive to antagonist; **RAV**: related to ABI3/VP1; **RTE**: reversion to ethylene sensitivity. The route of intracellular signal transduction is initiated by the interaction of ethylene with a membrane receptor (encoded by *ETR* genes) and through the modulation of CTR activity, which regulates the activity of several genes, such as EIN3. The receptors with CTR (similar to the protein kinase RAF - MKKK) work similarly to negative regulators of the pathway and, in the absence of ethylene, suppress downstream positive components of signal transduction. The hormone binding blocks the receptors in an inactive conformation, reducing the repression of metabolic pathway-positive regulators (Bleecker, 1999). In the absence of ethylene, CTR phosphorylates the EIN2 C-terminal domain, promoting its interaction with ETP F-box protein (not identified in soybean) and its subsequent degradation via proteasome 26S (Qiao et al., 2009). In the absence of EIN2 C-terminal phosphorylation (presence of the hormone), this domain is cleaved and moves to the nucleus, where it stimulates EIN3/EIL activity by EBF repression (stimulating the degradation of this F-box protein by unknown mechanisms), which in turn induces target genes transcription through some members of the AP2/ERF superfamily of transcriptional factors (Ju et al., 2012). In addition to the interaction with the C-terminus of EIN2, EIN3/EIL activity can be influenced by the MKK4-5-9 → MPK3-6 phosphorylation cascade, which is CTR/EIN2-independent. In the presence of a signal, the EIN3/EIL transcriptional factors are phosphorylated by MPK3-6, preventing the interaction with EBF and their degradation via the 26S proteasome. Thus, EIN3 and EIL accumulate in the nucleus, interact with gene target promoters and trigger ethylene responses (Stepanova and Alonso, 2009). Another positive regulator is EIN5, a 5'-3'-exoribonuclease that promotes EBF mRNA decay, increasing the levels of EIN3/EIL in the nucleus (Olmedo et al., 2006). Additionally, ethylene biosynthesis is also regulated. Possible receptors for endogenous signals (i.e., other phytohormones) can induce the secondary metabolites accumulation (i.e., calcium) in an intracellular environment and activate protein kinases (i.e., CPK2), culminating in the stabilization of type II ACSs, an important enzyme in ethylene biosynthesis. Then, type II ACSs (in *A. thaliana* AtACS5 and AtACS9) are phosphorylated by CPK2, which prevents the interaction of these enzymes with ETO/EOL and their subsequent degradation by the 26S proteasome. This event induces an increase in ethylene production and the activation of signal transduction pathways (Ecker, 2004). Moreover, various stress conditions (biotic and abiotic) induce the activation of MAPK modules (in *Arabidopsis thaliana* MKK4-5-9 and MPK3-6). The MPK3 and MPK6 kinases are able to phosphorylate the C-terminal type I ACSs (in *A. thaliana* AtACS2 and AtACS6), which stabilize and protect these enzymes against 26S proteasome degradation (Hahn and Harter, 2009). There is no consensus regarding the direct participation of CTR in a route involving MPK3-6 (Vandenbussche et al., 2012). The receptor activity is associated with two proteins: RAN, a copper carrier protein (copper is an important cofactor in receptor activity) (Binder et al., 2010); and RTE, a protein with an unknown mechanism of action that facilitates the transition among active and inactive states of one receptor, ETR1 (Dong et al., 2008; Stepanova and Alonso, 2009). Each protein is represented by a generic name: *EIN2*: GmEIN#002, GmEIN#004 and GmEIN#007; *EIN3*: GmEIN#001, GmEIN#005, GmEIN#006, GmEIN#008 and GmEIN#010; *EIN5*: GmEIN#003, GmEIN#009 and GmEIN#011;

Figure 2. (cont.) *MKK4*: GmMKK#001 and GmMKK#003; *MKK5*: without representatives identified in soybean; *MKK9*: GmMKK#002 and GmMKK#004; *MPK3*: GmMPK#003 and GmMPK#004; *MPK6*: GmMPK#001 and GmMPK#002; *Receptors*: EIN4: GmETR#002, GmETR#004, GmETR#008 and GmETR#011; ERS1: GmETR#001 and GmETR#007; ERS2: without representatives identified in soybean; ETR1: GmETR#003 and GmETR#006; ETR2: GmETR#005, GmETR#009 and GmETR#010 (see Table S6 in Additional File 2).

discovered. Furthermore, we also found five homologs of *EIN3/EIL* (*GmEIN#001*, *GmEIN#005*, *GmEIN#006*, *GmEIN#008* and *GmEIN#010*) and three of *EIN5* (*GmEIN#003*, *GmEIN#009* and *GmEIN#011*) in the *G. max* genome.

Finally, with regard to the main kinases and F-box proteins related to ethylene signal transduction, thirteen homologs of the kinases were found in the soybean genome, with four of them being homologs of *MKK4/MKK9*, four of *MPK3/MPK6* and five of *CPK2* as well as seven of *EBF* and ten homologs of *ETO/EOL* (Figure 2).

Transcriptional regulation of soybean ethylene genes

To understand better their transcriptional regulation mechanisms, we performed an *in silico* analysis of the putative promoter regions of the 176 soybean genes. We identified 14,385 elements in these putative promoters, corresponding to 29 *cis*-acting element families described in the literature for their transcriptional regulation in different plant species (Figure 3; Table S9 in Additional File 4).

As expected, all of the promoter regions contained elements from *PTPB* (plants *TATA-box*) and/or *CAAT* (*CCAAT-box*), suggesting that the analyzed sequences have a strong likelihood of being real gene promoters.

Apart from the *PTPB* and *CAAT* families, the most represented families in this analysis were those related to transcription factors MYB, MYC and NAC (Table S9 in Additional File 4; Figure 3) and to elements known for heat and light response (*LREM* and *HEAT,* respectively*)*. Interestingly, no *cis*-acting elements were found from the *RAV3* family in any of the putative promoters, indicating that there are possible variations in recognizing the sequence of the B3 domain that is representative of the RAV family in soybean. Another possibility could be that the regulation occurs because of the interaction of the AP2 domain with the *RAV5 cis*-acting element, which is broadly dispersed in the analyzed regions (Wittkopp and Kalay, 2011).

The families *EINL* (*ethylene insensitive 3-like*) and *GCCF* (*GCC-box*) of *cis*-acting elements are most likely directly related to the regulation of metabolic pathways in which ethylene plays a critical role. *EINL* and *GCCF* were present in 63.1 % and 11.4 %, respectively,

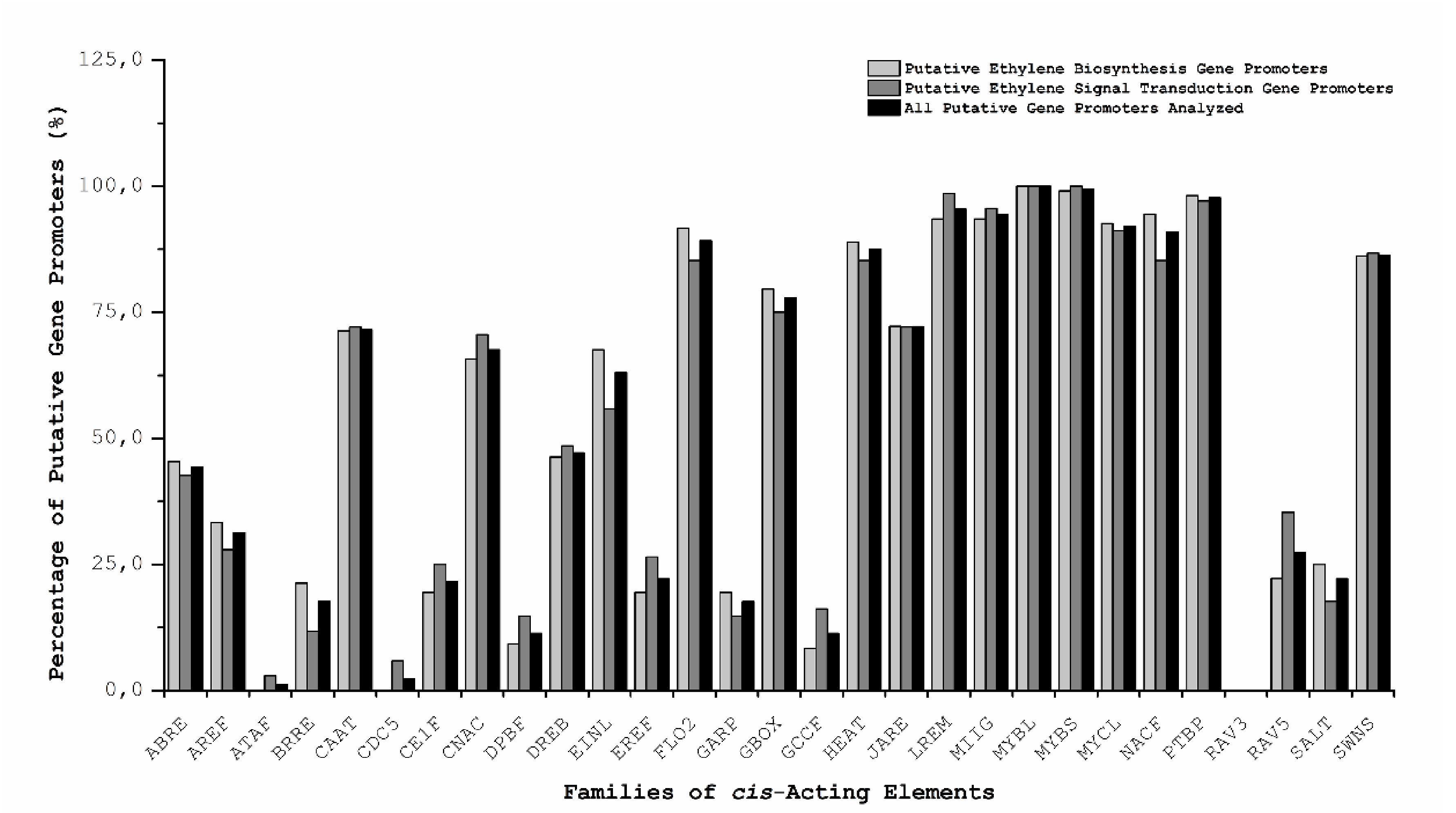

Figure 3. Distribution of *cis*-Acting elements in putative soybean gene promoters. The graph shows the distribution of *cis*-acting elements in promoter regions of soybean genes, related to ethylene biosynthesis and signal transduction. The *cis*-acting element families identified were as follows: *ABRE* (ABA response elements); *AREF* (auxin response elements); *ATAF* (ATAF-like NAC domain containing proteins); *BRRE* (brassinosteroid response elements); *CAAT* (CCAAT binding factors); *CDC5* (*A. thaliana* CDC5 homologs); *CE1F* (coupling elements 1 binding factors); *CNAC* (calcium regulated

Figure 3. (cont.) NAC-factors); *DPBF* (*Dc3* promoter binding factors); *DREB* (dehydration responsive element binding factors); *EINL* (ethylene insensitive 3 like factors); *EREF* (ethylene response element factors); *FLO2* (floral homeotic protein APETALA2); *GARP* (MYB-related DNA binding proteins - Golden2, ARR, Psr); *GBOX* (plant *G-box*/*C-box* bZIP proteins); *GCCF* (*GCC-box* family); *HEAT* (heat shock factors); *JARE* (jasmonate response elements); *LREM* (light responsive element motifs, not modulated by different light qualities); *MIIG* (MYB IIG-type binding sites); *MYBL* (MYB-like proteins); *MYBS* (MYB proteins with single DNA binding repeat); *MYCL* (MYC-like basic helix-loop-helix binding factors); *NACF* (plant specific NAC transcriptional factors); *PTBP* (plant TATA binding protein factors); *RAV3* (3'-part of bipartite RAV1 binding site); *RAV5* (5'-part of bipartite RAV1 binding site); *SALT* (salt/drought responsive elements); *SWNS* (secondary wall NACS).

of the putative promoters analyzed (Figure 3). The *DREB* (*dehydration responsive element binding factors*) and *EREF* (*ethylene response element factors*) elements are known for their involvement in the response to different stresses, and they were found in 47.2 % and 22.2 %, respectively, of the analyzed sequences.

When we analyzed the *cis*-acting elements contained in the putative promoters of the ethylene biosynthesis genes, we observed that 67.6 % had *EINL* elements and that 8.3 % had *GCCF* elements. Moreover, other *cis*-acting elements that respond to other phytohormones were detected, of which the *JARE* family (*jasmonic acid*) was present in more than 70.0 % of the putative promoters, followed by the *ABRE* and *CE1F* (*ABA response*) families, which were present in 45.4 % and 19.4 %, respectively, of the putative promoters. Moreover, 30.0 % of them have elements that respond to auxin (*AREF*) and 21.3 % to brassinosteroids (*BRRE*). Finally, the elements *DREB* and *EREF* could be detected in 46.3 % and 19.4 % of the putative promoters, respectively.

Considering the group with an ethylene-mediated transduction signal, we observed the presence of *EINL* elements in 55.9 % and *GCCF* in 16.8 % of the putative promoters. We also detected the *JARE* element in more than 70.0 % of the sequences analyzed, *ABRE* and *CE1F* in 42.6 % and 25.0 %, respectively, the auxin and brassinosteroid response elements in 28.0 % and 11.8 %, respectively, and the *DREB* and *EREF* elements in 48.5 % and 26.5 % of the putative promoters, respectively.

The analysis of the putative promoters showed that the activation or repression of the transcription of a gene in soybean is not likely to be regulated by isolated transcription factors but rather by the interaction of different proteins in a set of DNA-regulatory sequences. In accordance with this hypothesis, this study supported the results of other studies that had proposed crosstalk between the regulation of ethylene metabolism with other development mechanisms, homeostasis and response to various stresses. This affirmation was confirmed by the detection in the possible promoters of different *cis*-acting elements important for responses to other phytohormones, in addition to elements involved in different biotic and abiotic stress responses (heat shock, pathogen resistance, mechanic injuries, etc.). The presence of *cis*-acting elements in the 176 global soybean genes analyzed showed that the *JARE* elements were the most abundant, followed by *EINL*, *DREB* and *ABRE*. The putative promoter analysis indicated that each *cis*-acting element family could contribute in distinct ways to the regulation of the considered soybean genes: *ABRE*, *EINL*, *AREF* and *BRRE* are the most represented in the putative promoters of ethylene biosynthesis genes, and *JARE*, *DREB*, *EREF*, *CE1F* and *GCCF* are the most represented in the putative promoters of ethylene-mediated signal transduction (Figure 4).

Few (11.4 %) of the putative promoters presented *GCCF cis*-acting elements (responsive to ethylene), whereas almost half of them had the very similar *DREB* element, which responds first to drought stress. These proportions were the same in the genes that were differentially expressed in drought stress. Recent ChIP (chromatin immunoprecipitation) experiments showed that the transcription factor ERF1 from *A. thaliana* could interact directly with both *cis*-acting element families. More interestingly, this transcription factor interacted with *GCCF* elements under biotic stress conditions and with *DREB* elements under abiotic stress conditions but never with both at the same time (Cheng et al., 2013).

The data showed that 95.5 % of the putative promoters have the *LREM cis*-acting element (light-responsive elements, not mediated by different types of light) and that 87.5 % have *HEAT* elements (heat shock elements) (Figure 3). In *A. thaliana,* the response to low light intensity could be regulated by ethylene and auxins (induction of AUX22, ACS6, ACS8, ACS9). Similarly, ethylene biosynthesis and ethylene signal transduction, regulated by phytochrome B, are affected by antiphase light and temperature cycles (Bours et al., 2013; Vandenbussche et al., 2003). Complementary studies with etiolated pea stems showed that in addition to light intensity, red light also regulates ethylene biosynthesis and gravitropism (Steed et al., 2004). Additionally, mutants in receptors or orthologs of EIN2 sensitive to ethylene produce high levels of the gas, whereas *ctr1-1* mutants produce lower levels of ethylene than wild plants (Thain et al., 2004). However, although the double mutants *ein3/einl1* have similar phenotypes to *ein2* mutants, they produce low levels of ethylene when grown under long day periods but high levels when grown under dark conditions and even lower levels of ethylene than in *etr1* and *ein2* mutants (An et al., 2010). Thus, it is suggested that there is a parallel route to EIN3/EIL that is responsible for the negative control of ethylene biosynthesis, a mechanism that is light

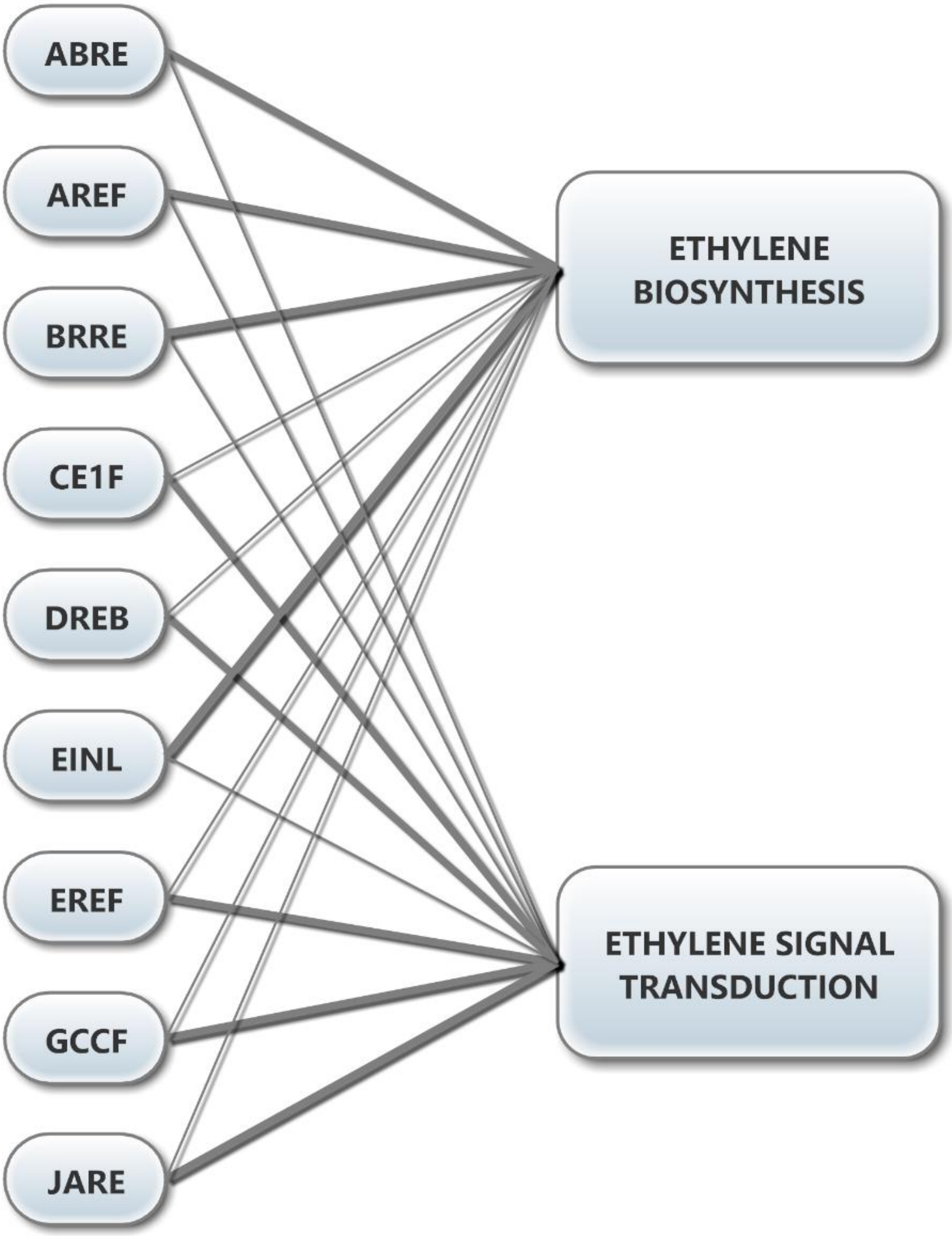

Figure 4. Distribution of *cis*-acting element families important in ethylene biosynthesis and signaling in putative soybean promoters. The diagram corresponds to the number of possible soybean promoters and the number of *cis*-acting elements present in each group analyzed: ethylene biosynthesis and signal transduction. The line thickness is directly related to the contribution of each family of *cis*-acting elements in each group: the thinnest lines correspond to the fewest number of elements and putative promoters that have them, and the thickest line corresponds to the highest number of elements and putative promoters that have them. *ABRE* - ABA response elements; *AREF* - auxin response elements; *BRRE* - brassinosteroid response elements; *CE1F* - coupling elements 1 binding factors; *DREB* - dehydration responsive element binding factors; *EINL* - ethylene insensitive 3 like factors; *EREF* - ethylene response element factors; *GCCF* - *GCC-box* family; *JARE* - jasmonate response elements.

dependent. Transcriptional regulation could be associated with the light-responsive transcription factors that interact with *LREM* elements, which can modulate the response depending on the variation of the *G-box* sequences that commonly flank the *LREM* elements (Jiao et al., 2007). Because more than 77.8 % of the putative promoters have *GBOX* elements and are associated with a high rate of *LREM*, we believe that the mechanisms involving EIN3/EINL, its partners or regulated factors, and other light-responsive factors play important roles in the regulation of soybean ethylene biosynthesis.

Many differentially expressed transcripts identified in soybean transcriptomes have been described in the literature as being important in the response to drought. The functions of these transcripts could be associated with not only ethylene biosynthesis and signaling but also with other metabolic pathways. For example, the enzymes responsible for AdoMet production in ethylene biosynthesis also contribute to other metabolic pathways that are ethylene-independent. Plant polyamines in *A. thaliana* are involved in the response to different environmental stresses, and recent studies have indicated that polyamine signaling is involved in direct interactions with different metabolic pathways and intricate hormonal crosstalks, such as ABA regulation in response to abiotic stresses (Alcázar et al., 2010). Because MAT (SAMS) enzymes provide the substrate for polyamine synthesis, it is very probable that these enzymes are induced by ABA in the response to abiotic stresses, as was demonstrated in tomato plants that had high levels of these enzyme transcripts under NaCl stress conditions and after ABA treatment (Espartero et al., 1994). Thus, it could be suggested that high levels of ABA are related to low levels of ethylene because of a possible redirection of AdoMet toward the biosynthesis of polyamines. We observed that among the *MAT* genes in soybean, 54.6 % have *ABRE* in their putative promoters, indicating induction of these genes by ABA in response to abiotic stresses.

The presence of elements responsive to other phytohormones must also be considered in the regulation of ethylene biosynthesis. Zhang and coworkers (2009) demonstrated that ABA could induce the genes that encode the enzymes ACC synthase and ACC oxidase, stimulating ethylene biosynthesis and fruit ripening (Zhang et al., 2009). Additionally, studies have shown that one of the first actions of auxins is the induction of *ACSs*, which increase ethylene production (Abel and Theologis, 1996). Along with auxins, brassinosteroids and methyl-jasmonate could also induce ACO enzymes, increasing ethylene production in maize and olive plants (Lim et al., 2002; Sanz et al., 1993).

These studies with putative soybean promoters are important not only for a better understanding of ethylene signaling in this crop but also for the production of genetically modified plants with genes regulated under different stress conditions separately and/or simultaneously.

Analysis and validation of soybean transcriptomes in water deficit conditions

Transcriptome databank analysis of water deficit contrasted with soybean genotypes

To investigate the expression of soybean genes, we studied the transcriptome of two cultivars with contrasting responses to drought stress (sensitive to drought BR16 and tolerant to drought EMBRAPA48). The plants were grown hydroponically and under different water stress conditions. The transcriptomes, provided by the GENOSOJA project, were constructed using *subtraction library hybridization* (SSH), which detects differential expression of transcripts under water stress. In this database, 40.9 % of the genes identified were expressed differentially in at least one of the listed situations. Among them, 43.1 % were related to ethylene biosynthesis and 56.9 % to its signal transduction (Figure S5 and Figure S6 in Additional File 1). Furthermore, we found that 25.0 % of differentially expressed genes were detected in sensitive BR16, 47.2 % were detected in drought-resistant EMBRAPA48, and 27.8 % were present in both cultivar databases. These contrasting results might be explained by the genetic basis of each cultivar providing the relative variations in the gene expression or by a discrepancy between the obtained unique sequences and the cultivar databases (42.3 million unique sequences generated, of which 27.8 % are from BR16 and 72.2 % from EMBRAPA48) (Rodrigues et al., 2012).

We observed that 37.5 % of the differentially expressed genes were detected uniquely in roots (among which 3.7 % were from BR16 and 96.3 % were from EMBRAPA48), 26.4 % were detected exclusively in leaves (among which 84.2 % were from BR16, 10.5 % were from EMBRAPA48, and 5.3 % were found in both cultivar databases), and 36.1 % were expressed in both roots and leaves. These results, together with the normalized data presented in Figure S7 (Additional File 1), suggested that the expression of genes in the roots was preferentially observed in the drought-tolerant EMBRAPA48, whereas in the leaves, the differential expression was more proportionate depending on the group of genes and the duration of stress.

Furthermore, the expression of both groups of genes was analyzed. We observed that 28.7 % of the biosynthesis genes were expressed in roots and leaves of the sensitive and tolerant cultivars. Among them, 16.1 % were expressed in only sensitive BR16, mainly in the leaves. Conversely, 43.9 % differential expression was detected exclusively in tolerant EMBRAPA48, mainly in the roots (Figure S5 in Additional File 1).

In ethylene-mediated signal transduction, 61.8 % of the genes were differentially expressed under water stress conditions. In the sensitive cultivar, 33.3 % were differentially expressed, mostly in the leaves, and in the tolerant cultivar, 42.9 % had a differential expression, mainly in the roots (Figure S6 in Additional File 1).

Transcriptome functional validation of the candidate genes

To validate the data obtained *in silico*, the levels of some differentially expressed genes were assessed by RT-qPCR in both leaf and root tissues exposed to drought stress. The plants were grown under the same conditions as those used for transcriptome analysis. The C_t (cycle threshold) values obtained are listed in Tables S10 and S11 (Additional File 5).

The expression of the genes *MAT*, *ACS* and *ACO* were found to have the same differential trends as the data obtained *in silico*, although with variations in the expression profiles (Figure 5A-C. Additional File 1 - Figures S8A-C). This result could be due to limitations in the construction of the GENOSOJA subtractive libraries and general experimental variations. As an example, the expression of *ACS* is different in both cultivars and tissues with RT-qPCR, but it was detected in only the transcriptome of the roots of EMBRAPA48.

Induction kinetics analysis of soybean *ACS* and *ACO* genes confirmed the temporality of the metabolic reactions catalyzed by these enzymes in both cultivars because *ACS* gene expression reached its peak earlier than that of the *ACO* gene (Figure 3B-C). Furthermore, when comparing the two soybean varieties, the expression of these two genes was observed earlier in the drought-tolerant cultivar. This fact could be evidence of ethylene participation in soybean responses to water stress.

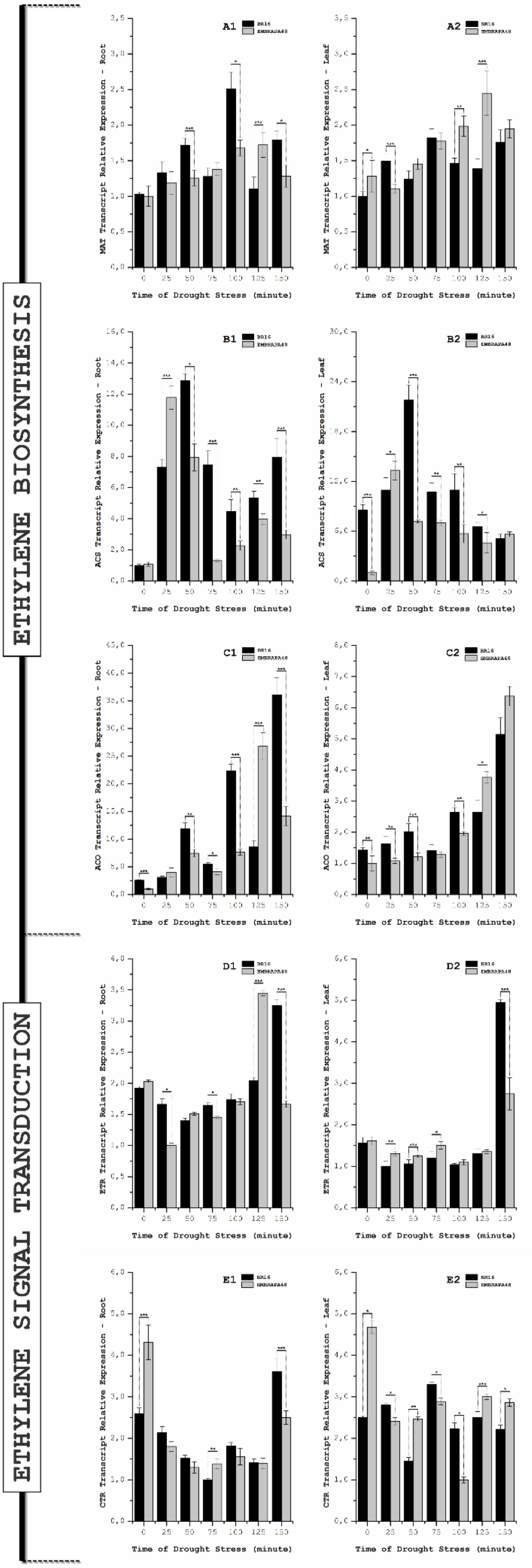

Figure 5. Expression of ethylene-Related genes in soybean under drought stress conditions. The graphs show the expression levels, obtained by RT-qPCR, of five soybean genes related to ethylene biosynthesis [*MAT* (**A**), *ACS* (**B**) and *ACO* (**C**)] and ethylene signal transduction [*ETR* (**D**) and *CTR* (**E**)]. The expression of these genes in the experiment was compared in roots and leaves of soybean cultivars BR16 and EMBRAPA48 after different durations of drought stress. In the column graphs **1** and **2**, the statistical analysis was performed by comparing similar tissues in both cultivars under the same conditions of drought stress (same durations). The asterisks represent the level of statistical significance: (*****) $p \leq 0.05$; (******) $0.01 \leq p < 0.05$; (*******) $0.001 \leq p < 0.01$. Each dot represents the average amount (± standard error) of three experimental replicates (same sample) in three biological samples (different plants), totaling nine replicates. The standard error is not presented with some of the dots because their absolute values are lower than the scale. After normalization based on housekeeping genes, the values given in the graph are relative to the lowest expression, whose value was set at 1 (one). Information about the target genes is presented in Method S1 (Additional File 6).

The same analysis was performed with the genes coding ethylene receptors (*ETR*) and for the protein kinase *CTR* (Figure 5D-E. Additional File 1 - Figure S8D-E). Few differences were observed in the expression patterns of the transcripts of these genes between the cultivars. In the roots of both cultivars, there was a reduction in the level of *ETR* transcripts, comparing stressed and non-stressed plants. The maximal expression was achieved under the longest periods of stress. In the same tissue, the transcripts of *CTR* were reduced, with a significant increase detected only after 150 minutes of water deficit in both cultivars. In the leaves, when comparing the stress and no-stress conditions,

we were able to observe a slight reduction in the levels of transcripts of *ETR* in the first 125 minutes of drought and a peak elevation at the end of the analysis (150 minutes). Relative to the *CTR* transcripts, it was observed that the expression of the drought-tolerant cultivar was higher in the non-stressed state (time zero).

Levels of free ACC and ethylene production

To compare and correlate the data obtained *in silico* with the physiological data, we assessed the levels of free ACC and ethylene in both BR16 and EMBRAPA48 cultivars. The plants were grown under similar conditions as those used for the analysis of transcriptomes. The physiological data showed that both of the cultivars suffered under the water deficit but that the tolerant cultivar responded better, exhibiting increases in the photosynthetic rate, stomatal conductance and evapotranspiration after 75 minutes of stress (Figure S9 in Additional File 1). The water consumption (WUE) showed that before 75 minutes had elapsed, the sensitive cultivar was using its water resources better than the tolerant cultivar, but subsequently, the situation was reversed; thus, the stress caused a greater impact on the susceptible cultivar.

We analyzed the levels of free ACC and ethylene production and found, in general terms, that free ACC was mostly increased in the leaves, whereas ethylene was mostly increased in the roots (Figure 6 and Figure S10 in Additional File 1). We observed that the EMBRAPA48 cultivar had higher levels of free ACC in the leaves and variable levels in the roots. In the roots of non-stressed plants (time zero), the free ACC was higher in BR16 plants, and the ethylene production was higher in the EMBRAPA48 cultivar (Figure 6A-B. Additional File 1 - Figure S10A-B), whereas in the leaves, the level of free ACC was the same in both cultivars, and the quantity of ethylene production was higher

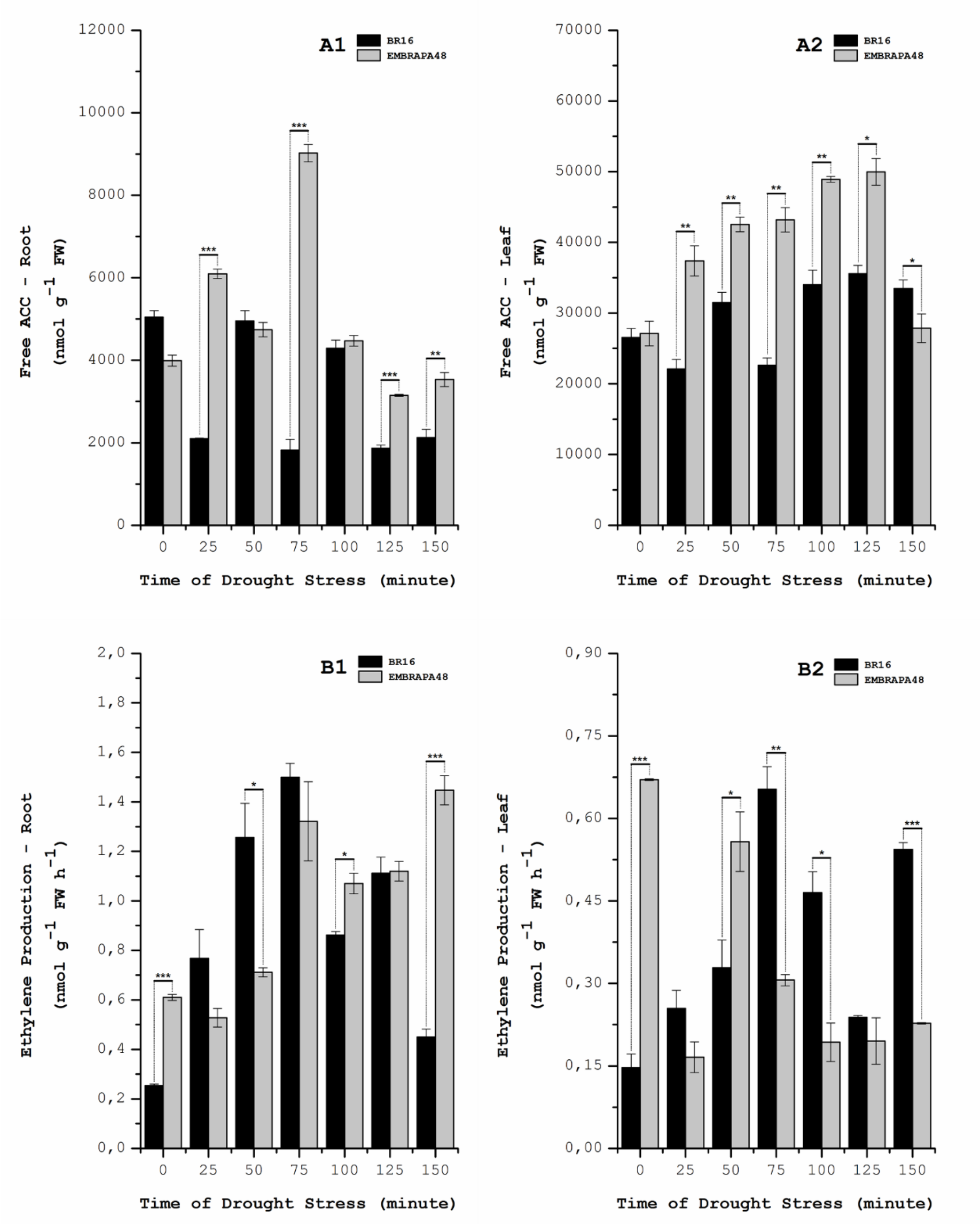

Figure 6. Levels of ethylene and free ACC production in soybean under drought stress conditions. Values were determined for ethylene production and free ACC (1-aminocyclopropane-1-carboxylic acid) in roots and leaves of soybean cultivars BR16 and EMBRAPA48 after the application of different durations of drought stress. The codes **A1** and **A2** represent levels of free ACC; **B1** and **B2** represent levels of ethylene production. The statistical analysis was performed by comparing similar tissues in both cultivars under the same conditions of drought stress (same durations). The asterisks represent the level of statistical significance: (*****) $p \leq 0.05$; (******) $0.05 < p \leq 0.01$; (*******) $0.01 < p \leq 0.001$. Each dot represents the average amount (± standard error) of three replicates in different plants. The standard error is not presented with some dots because their absolute values are lower than the scale.

in EMBRAPA48 (Figure 6A2-B2). Except for the period of 25-50 minutes of stress, it was observed that both free ACC and ethylene production exhibited cyclic behavior in both the leaves and roots of BR16: when free ACC increased, ethylene decreased, and vice versa. Additionally, in BR16, we found that in both tissues, the peak of ethylene production (75 minutes) corresponded to the lowest value of free ACC. In the EMBRAPA48 cultivar, this cyclic pattern was much less evident. We observed that the highest peaks of free ACC were found after 75 minutes in the roots and after 125 minutes in the leaves, whereas the maximal production of ethylene corresponded to 75 and 150 minutes of stress, respectively. In the leaves, the maximal peak of ethylene production occurred at time zero.

When we compared the levels of ethylene production and free ACC, two different situations were observed. First, an increase in free ACC coincided with an increase in ethylene production. Although the hydrolysis of ACC aggregates remains contradictory, the high level of free ACC could be explained by the degradation of these aggregates of malonyl-ACC into free ACC, accompanied by an increase in free ACC production and the conversion of AdoMet into ACC by the ACS enzyme. Thus, the levels of free ACC would exceed the capacity of ACO enzymes to convert it into ethylene, which would be present at its maximal level (Fluhr et al., 1996; Hoffman et al., 1982; Jiao et al., 1986). Conversely, we observed a reduction of the levels of free ACC, together with a reduction in ethylene production. This finding could be explained by the formation of malonyl-ACC and glutamyl-ACC, accompanied by the degradation of the ethylene precursor by ACD enzymes. To support this conclusion, we simultaneously detected the differential expression of GmGGT#002 in the roots and leaves of both cultivars and GmACD#001 in the roots of EMBRAPA48. To understand this trend better, it would be necessary to characterize the molecular pathways involved in ACC conjugation and degradation *in vitro* and *in vivo* to determine the precise mechanisms underlying the regulation of ethylene biosynthesis in response to diverse signals, in addition to the identification of the actual role of the formation of ACC aggregates in this case.

In our work, transcriptome analysis, RT-qPCR and ethylene production revealed that ethylene synthesis depended on the tissue analyzed. After 75 minutes of water deficit, the maximal production of ethylene was observed in leaves of BR16, whereas after the same period of water deficit, in the leaves of EMBRAPA48, ethylene exhibited a significant decrease. In the roots, both cultivars had high levels of ethylene production (Figure 6). Together with only the tolerant cultivar displaying an increase in stomatal conductance, photosynthetic rate and transpiration after the same stress period (Figure S9 in Additional File 1), these findings indicated that in this situation, leaves and roots undertake different responses to ethylene. Additionally, we can suggest that in the leaves, ethylene production could be associated with the response to drought stress because ethylene could regulate stomatal closure (Chen et al., 2013).

Nonetheless, studies have shown that the levels of this phytohormone are low when plants are exposed to water deficit (Sharp, 2002; Spollen et al., 2000). These conflicting observations could be attributed to the system in which the soybean plants were grown. The plants were grown hydroponically, with the roots submerged in a nutrient-containing solution. Some studies have shown that variations in gene expression could occur when hydroponic and soil cultures were compared (Guimarães-Dias et al., 2012). Thus, it is believed that hydroponically grown roots have molecular responses similar to those of roots grown under flooding conditions and that when subjected to water deficit, they exhibit molecular responses different from those shown by roots grown via soil culture.

Additionally, other works have reported that in plants grown under flooding conditions, the levels of ethylene production were higher than those obtained under water deficit conditions (Bailey-Serres and Voesenek, 2008). We also believe that a natural elevation of the temperature caused by rapid water loss could be explained by an increase in ethylene biosynthesis because the activity of enzymes was also rapidly increased, as shown by Antunes (2000) (Antunes and Sfakiotakis, 2000). Because ethylene diffusion is more rapid in the air than in liquid (Hoagland’s solution) and because water deficit and dehydration are more rapid under hydroponic conditions, we believe that the plants do not have sufficient time to begin molecular responses before desiccation occurs. One explanation could be that when short intervals of water deficit (25-50 minutes) are applied, ethylene biosynthesis and signal transduction remain similar to those under normal growing conditions. Therefore, when the stress duration is increased, the signal transduction could be strongly decreased. In fact, the plants were switched from flooding stress to water deficit stress, possibly activating different responses that substituted for the normal water deficit responses because we observed the differential expression of many genes even before the stress was administered. Thus, the analysis of the GENOSOJA database would be best complemented by next-generation sequencing experiments to replace the SSH methodology and cultivation in pot systems, instead of under hydroponic conditions.

CONCLUSIONS

This study was the first to propose accurate models for ethylene biosynthesis and signaling in the soybean. Based on the currently available databases, soybean genes and proteins homologous to almost all of the components of the pathways featured in *A. thaliana* were identified, with the exception of the *ETP* gene. The *cis*-acting elements present in soybean putative promoters were described to infer possible models and the regulation of signaling pathways linked directly to ethylene as well as their communication with other metabolic routes. RT-qPCR experiments were important to the validation of soybean transcriptome data and allowed for the evaluation of the induction kinetics of *ACS* and *ACO* soybean genes. Finally, changes were observed in the levels of production of ethylene and its precursor (in its free form) in soybean cultivars under water stress conditions.

By the integration of all data, many inferences could be made, among which the involvement of ethylene in soybean water stress responses stands out. Furthermore, this work showed that regulation of the ethylene-mediated response could be influenced by diverse exogenous and endogenous factors, indicating that the balance of these various factors determines the quality and intensity of different stimuli responses. Further studies are necessary to continue elucidating *in vivo* molecular mechanisms involved in ethylene coordination in soybean both to confirm our observations and to facilitate biotechnological strategies for the improvement of cultivar tolerance to various stresses.

METHODS

Functional annotation

Based on the Genbank TAIR (*The Arabidopsis Information Resource*; *http://www.arabidopsis.org/*) (Swarbreck et al., 2008), we selected the genes related to ethylene biosynthesis and signaling in *Arabidopsis thaliana*. A BLAST (*Basic Alignment Search Tool)* search was performed with the amino acid (amino acid sequences, equal or over 200 bits score, against protein sequence databases: *Glycine max* [L.] Merrill (*GENOSOJA*: *http://www.lge.ibi.unicamp.br/soybean/*; *SoyBase*: *http://soybase.org/*; *Phytozome version 9.1*: *http://www.phytozome.net/*) and *Oryza sativa* Nipponbare (*Rice Genome Annotation Project*; *http://rice.plantbiology.msu.edu*) (Altschul et al., 1990; do Nascimento et al., 2012; Kawahara et al., 2013).

Subsequently, 176 soybean genes were ranked in three groups according to their ontology with *Blast2GO* software (Gene Ontology) (Conesa et al., 2005): (i) *cell componen*t, with suggestions about their active locations at the cellular and macromolecular complex substructure levels; (ii) *molecular function*, with descriptions of their the catalytic activity or binding at the molecular level; and (iii) *biological processes*, with descriptions of their biological objectives according to one or more ordered sets of molecular features. For this purpose, the soybean nucleotide sequences of each gene were processed with the aid of the *BlastX* tool (used to search the database according to the nucleotide sequences translated into all six possible reading phases) using the *A. thaliana* protein sequences as a database and only selecting those with an *e-value* $\leq e^{-10}$. After the annotation, the functionality of the sequences was analyzed with the aid of the online tool *InterproScan version 5.0* (Jones et al., 2014) and finally determined by the online software *GO Ontology-Slim* (Ashburner et al., 2000).

The protein domain analysis of the amino acid sequences of the genes selected in the three studied organisms was performed using the *PFAM* (*Protein Family*; http://pfam.xfam.org/) bioinformatic tool (Punta et al., 2012). The selection parameter (*e-value* < 1.0) used was the same one defined by this program's website.

The ideogram representing the location of the 176 soybean genes analyzed in the 20 chromosomes was built in proportion to the chromosome size (1.0 cm corresponds to 5.0 megabase), taking into account the location of each gene and the DNA strand in which they are localized (sense and antisense). The positions of the centromeres and the size of each chromosome were obtained from a reference genome (Schmutz et al., 2010).

The sequence alignment analysis and dendrogram construction were performed with the programs *BIOEDIT version 7.0.9.0* and *MEGA version 5*, respectively. The Neighbor-joining analyses were used to calculate the distance matrices for dendrogram construction. Bootstrap analysis with 10^4 replicates was performed to test the robustness of the internal branches. The proposed models for ethylene biosynthesis and signaling in soybean were obtained from the *SoyCyC version 3.0* (*Soybean Metabolic Pathway Database*; *Soybase*; *http://www.soybase.org:8082/*) and *KEGG* (*Kyoto Encyclopedia of Genes and Genomes*; *http://www.genome.jp/kegg/*) (Kanehisa and Goto, 2000; Selkov et al., 1998).

The possible protein phosphorylation sites by MAP kinases (MAPK) and calcium-dependent protein kinase present in the 1-aminocyclopropane-1-carboxylic acid synthase (ACSs) type I and type II, respectively, were determined by the online program *NetPhos version*

2.0 (http://www.cbs.dtu.dk/services/NetPhos/) (Blom et al., 1999). The presence of possible phosphorylation sites was analyzed in the C-terminal region using the amino acids tyrosine, serine and threonine.

Each protein sequence identified in the soybean databases (*A* sequence) was compared individually with those from *A. thaliana* (*TAIR*) and rice (*Rice Genome Annotation Project*) to obtain the homologous *B* and *C* sequences, respectively. The BBH (*Best Bidirectional Hit*) criterion was used, and positive hits were obtained when *B* and/or *C* sequences were compared with the soybean database; the best similarity was with the *A* sequence. Gene duplication was considered to avoid false negatives (Dalquen and Dessimoz, 2013).

The presence of *cis*-acting elements in putative promoter regions was examined [2,000 pairs upstream of the open reading frame bases (ORF)] for each soybean gene selected for this study. This analysis was performed using the bioinformatics tool *MatInspector version 8.0* (Genomatix®) using "*plants*" as matrix group, "*0.85*" as the value for the similarity of the main bases that constitute each *cis*-acting element (core similarity), and "*Optimized +1*" as the value for the similarity matrix (similarity matrix) (Cartharius et al., 2005).

The expression of each gene involved in the biosynthesis and signaling of the ethylene metabolic pathway was accessed in the GENOSOJA database (Rodrigues et al., 2012). The gene expression levels were represented in graphics indicating the *FPKM* (fragments per kilobase of exon per million fragments mapped) normalized read counts for each gene that was differentially expressed in the twelve cDNA libraries (25-50, 75-100 and 125-150 minutes of drought stress).

Plant growth and physiological parameters

Soybean seeds from BR16 and EMBRAPA48, which are sensitive and tolerant to water deficit (Casagrande et al., 2001; Texeira et al., 2008), respectively, were germinated on filter paper (*Germitest*) for seven days in a growth chamber at 25.0 ± 1.0 °C and 100.0 % relative humidity (RH). The seedlings were transferred to 36L boxes containing 50.0 % Hoagland’s solution (Hoagland and Arnon, 1950), which was continuously aerated and replaced weekly. The plantlets were grown until V5 stage (Fehr et al., 1971) in a greenhouse under a natural 12h photoperiod at 30.0 ± 5.0 °C and 60.0 ± 10.0 % RH. The experimental design was a randomized complete block in a 2x7 factorial arrangement involving two cultivars (BR16 and EMBRAPA48) and seven water deficit periods (0, 25, 50, 75, 100, 125 and 150 minutes), respectively, with three replicates. The stress was imposed by removing the plants from the hydroponic solution and leaving them without nutrient solution for up to 150 minutes under air exposure conditions. For each water deficit period, root and leaf samples were collected from three plants, pooled and frozen in liquid nitrogen before storage at -80°C.

The photosynthetic rate (*A*), photosynthetically active radiation (*PAR*), internal CO_2 concentration (C_i), stomatal conductance (g_s) and transpiration rate (*E*) were evaluated using a *LI-6400* Portable Photosynthesis System (Li-Cor, Inc.). The parameters were measured in triplicate on the youngest trifoliate leaf that was totally expanded under a photon flux density of 1,300 µmol m^{-2} s^{-1}. The temperature variation (ΔT) was measured by the difference between the air (T_{ar}) and leaf (T_{leaf}) temperatures. The water use efficiency (*WUE*) was determined by the ratio between *A* and *E*. The data were statistically analyzed by *ANOVA* using the *SAS* and *SANEST* (*Statistical Analysis System version 8.0*) softwares, and the treatments were compared by *Tukey's* test ($p \leq 0.05$).

Total RNA extraction and quantitative Real Time PCR (RT-qPCR)

Total root and leaf RNA from BR-16 and EMBRAPA48 of each treatment was extracted in triplicate using the Trizol (Invitrogen, Inc.) protocol and treated with DNAse I (Invitrogen, Inc.). Total mRNAs were utilized as templates for cDNA synthesis using the enzyme *Moloney Murine Leukemia Virus Reverse Transcriptase* (M-MLV RT) (Invitrogen, Inc.).

Quantitative real-time PCR was performed to validate the genes related to ethylene biosynthesis (*MAT, ACS* and *ACO*) and signaling (*ETR* and *CTR*) pathways in soybean. Primers were designed by *Primer 3 Plus* (Untergasser et al., 2007) software and checked for the presence of putative amplicons from 120 to 200 pb and melting temperature (T_M) of 60.0 ± 2.0 °C (see Additional File 6 - Method S1). To establish the normalization factor, two reference genes were used for root samples (*ACT11* and *UBC2*) and two for leaf samples (*CYP2* and *ELF1A*) (Kulcheski et al., 2010; Miranda et al., 2013). All experiments were carried out in experimental and biological triplicate. The quantitative real-time PCR amplifications were performed using the *ABI Real Time PCR System 7500 Fast* (Applied Biosystem, Inc.) thermal cycler with a comparative cycle threshold ($\Delta\Delta C_t$*). Rox Plus SYBR Green Master Mix 2X* (LGC Inc.) was combined with 4.0 or 10.0 µM of each primer (sense and antisense) and 2.0 µL of cDNA (40 or 80-fold dilution) for each experimental condition (Method S1). The PCR cycling conditions were 95°C for 15 min to activate the hot-start Taq DNA polymerase, 40 cycles at 95

°C for 30 s, 60 °C for 30 s and 72 °C for 3 min (final extension). The raw fluorescence data for all runs were imported into the *Real-Time PCR Miner* software (Zhao and Fernald, 2005) to determine the C_t value and the PCR efficiency. The C_t values were converted by *qBASE v.1.3.5* software (Hellemans et al., 2007). The statistical analysis was performed using the *REST 2009* (*Relative Expression Software Tool* - Qiagen, Inc.) software (Pfaffl et al., 2002) in two ways: first by comparing the relative gene expression values between both cultivars in the same tissue under the same stress conditions and second by comparing the control (without stress) with the stressed samples of the same cultivar.

Determination of ethylene production

For ethylene analysis, 0.5 g root and leaf samples were collected for each stress period in 50 mL glass recipients and sealed with a silicone lid. After 24 hours, the ethylene analysis was performed. First, a 1.0 mL sample of each treatment was obtained using a gastight syringe, and its concentration was determined by a gas chromatograph (GC) equipped with a flame ionization detector (FID), as described by Mainardi and coworkers (2006) (Mainardi et al., 2006). The GC column used was HP-Plot Q (30 m, D.I. 0.53 mm), and the injection conditions were a pressure of 20.0 psi for 2 minutes, ventilation flux of 20.0 mL.min^{-1} after 30 seconds and injector temperature of 200 °C. An isothermal program was run at 30 °C, employing constant fluxes of helium gas of 1.0 mL min^{-1}, a detector temperature of 250 °C and detector air and hydrogen fluxes of 400.0 mL min^{-1} and 40.0 mL min^{-1}, respectively. The ethylene production was estimated in relation to the injection of 0.1 µL L^{-1} of ethylene in synthetic air (Air Liquid Ltd.), and it was represented in nmoles for grams of fresh weight for hour (nmol g^{-1} FW h^{-1}).

Determination of free 1-aminocyclopropane-1-carboxylic acid (ACC)

Liu and coworkers (2012) proposed the method for the determination of free ACC (Liu et al., 2012). The samples were composed of roots and leaves from both cultivars, collected in triplicate, from different plants, stored in 50.0 mL Falcon tubes and frozen in liquid nitrogen (N_2). Approximately 0.5 g of each sample was crushed with N_2, and the powder was transferred to 5.0 mL of a 60.0 % methanol solution (v/v). The samples were stirred for one hour under ambient temperature and centrifuged at 1,4000 x g for 10 minutes at 25 °C. The supernatant was transferred to another tube. The residue was resuspended in 200.0 µL of ultrapure water and transferred to a 1.5 mL microcentrifuge tube, to which was added 300.0 µL of 200.0 mM borate buffer at pH 8.0 and 360.0 µL of 1.0 mM fluorescamine dissolved in acetone. The mixture was vigorously stirred, maintained at 25°C for 10 minutes and then filtered through a 0.45 micron porous membrane into a 2.0 mL glass vial. A 20.0 µL aliquot of the filtered mixture was injected into a liquid chromatograph coupled with a fluorescence detector (Agilent 1100). The sample was eluted through a *C18 Luna* column (5.0 microns, 300 x 4 mm, *Supelco, Sigma-Aldrich, USA*), and the effluent was monitored at an excitation wavelength of 378 nm and an emission wavelength of 475 nm. The results were calculated according to an external standard curve of standard ACC (Sigma-Aldrich, USA) in the range from 0.1 to 10.0 µg. The determination of free ACC is given in nmoles for gram of fresh weight (nmol g^{-1} FW).

REFERENCES

Abel, S., Theologis, A., 1996. Early genes and auxin action. Plant Physiol. 111, 9–17. doi:10.1104/pp.111.1.9

Alcázar, R., Altabella, T., Marco, F., Bortolotti, C., Reymond, M., Koncz, C., Carrasco, P., Tiburcio, A.F., 2010. Polyamines: molecules with regulatory functions in plant abiotic stress tolerance. Planta 231, 1237–1249. doi:10.1007/s00425-010-1130-0

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. doi:10.1016/S0022-2836(05)80360-2

Amrhein, N., Dorzok, U., Kionka, C., Kondziolka, U., Skorupka, H., Tophof, S., 1984. The biochemistry and physiology of 1-aminocyclopropane-1-carboxylic acid conjugation, in: Fuchs, Y., Chalutz, E. (Eds.), Ethylene. Springer Netherlands, Dordrecht, pp. 11–20. doi:10.1007/978-94-009-6178-4_2

Amthor, J.S., 2003. Efficiency of lignin biosynthesis: a quantitative analysis. Ann. Bot. 91, 673–695. doi:10.1093/aob/mcg073

An, F., Zhao, Q., Ji, Y., Li, W., Jiang, Z., Yu, X., Zhang, C., Han, Y., He, W., Liu, Y., Zhang, S., Ecker, J.R., Guo, H., 2010. Ethylene-induced stabilization of ethylene insensitive 3 and EIN3-like 1 is mediated by proteasomal degradation of EIN3 binding F-box 1 and 2 that requires EIN2 in Arabidopsis. Plant Cell 22, 2384–2401. doi:10.1105/tpc.110.076588

Antunes, M.D.C., Sfakiotakis, E.M., 2000. Effect of high temperature stress on ethylene biosynthesis, respiration and ripening of "Hayward" kiwifruit. Postharvest Biol Technol 20, 251–259. doi:10.1016/S0925-5214(00)00136-8

Argueso, C.T., Hansen, M., Kieber, J.J., 2007. Regulation of ethylene biosynthesis. J. Plant Growth Regul. 26, 92–105. doi:10.1007/s00344-007-0013-5

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. doi:10.1038/75556

Bailey-Serres, J., Voesenek, L.A.C.J., 2008. Flooding stress: acclimations and genetic diversity. Annu. Rev. Plant Biol. 59, 313–339. doi:10.1146/annurev.arplant.59.032607.092752

Bari, R., Jones, J.D.G., 2009. Role of plant hormones in plant defence responses. Plant Mol. Biol. 69, 473–488. doi:10.1007/s11103-008-9435-0

Bleecker, A.B., 1999. Ethylene perception and signaling: an evolutionary perspective. Trends Plant Sci. 4, 269–274. doi:10.1016/s1360-1385(99)01427-2

Blom, N., Gammeltoft, S., Brunak, S., 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. 294, 1351–1362. doi:10.1006/jmbi.1999.3310

Bours, R., van Zanten, M., Pierik, R., Bouwmeester, H., van der Krol, A., 2013. Antiphase light and temperature cycles affect phytochrome B-controlled ethylene sensitivity and biosynthesis, limiting leaf movement and growth of Arabidopsis. Plant Physiol. 163, 882–895. doi:10.1104/pp.113.221648

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., Werner, T., 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. Bioinformatics 21, 2933–2942. doi:10.1093/bioinformatics/bti473

Casagrande, E.C., Farias, J.R.B., Neumaier, N., Oya, T., Pedroso, J., Martins, P.K., Breton, M.C., Nepomuceno, A.L., 2001. Expressão gênica diferencial durante déficit hídrico em soja. Rev. Bras. Fisiol. Veg. 13, 168–184. doi:10.1590/S0103-31312001000200006

Chang, C., Kwok, S.F., Bleecker, A.B., Meyerowitz, E.M., 1993. Arabidopsis ethylene-response gene ETR1: similarity of product to two-component regulators. Science 262, 539–544. doi:10.1126/science.8211181

Chen, L., Dodd, I.C., Davies, W.J., Wilkinson, S., 2013. Ethylene limits abscisic acid- or soil drying-induced stomatal closure in aged wheat leaves. Plant Cell Environ. 36, 1850–1859. doi:10.1111/pce.12094

Cheng, M.-C., Liao, P.-M., Kuo, W.-W., Lin, T.-P., 2013. The Arabidopsis ethylene response factor 1 regulates abiotic stress-responsive gene expression by binding to different *cis*-acting elements in response to different stress signals. Plant Physiol. 162, 1566–1582. doi:10.1104/pp.113.221911

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676. doi:10.1093/bioinformatics/bti610

Dalquen, D.A., Dessimoz, C., 2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol. Evol. 5, 1800–1806. doi:10.1093/gbe/evt132

Deikman, J., Xu, R., Kneissl, M.L., Ciardi, J.A., Kim, K.N., Pelah, D., 1998. Separation of *cis*-elements responsive to ethylene, fruit development, and ripening in the 5'-flanking region of the ripening-related E8 gene. Plant Mol. Biol. 37, 1001–1011. doi:10.1023/a:1006091928367

Ding, Y., Virlouvet, L., Liu, N., Riethoven, J.-J., Fromm, M., Avramova, Z., 2014. Dehydration stress memory genes of *Zea mays*; comparison with *Arabidopsis thaliana*. BMC Plant Biol. 14, 141. doi:10.1186/1471-2229-14-141

Do Nascimento, L.C., Costa, G.G.L., Binneck, E., Pereira, G.A.G., Carazzolle, M.F., 2012. A web-based bioinformatics interface applied to the GENOSOJA project: databases and pipelines. Genet. Mol. Biol. 35, 203–211. doi:10.1590/S1415-47572012000200002

Doubt, S.L., 1917. The response of plants to illuminating gas. Botanical Gazette 63, 209–224. doi:10.1086/332006

Espartero, J., Pintor-Toro, J.A., Pardo, J.M., 1994. Differential accumulation of S-adenosylmethionine synthetase transcripts in response to salt stress. Plant Mol. Biol. 25, 217–227. doi:10.1007/BF00023239

Fehr, W.R., Caviness, C.E., Burmood, D.T., Pennington, J.S., 1971. Stage of development descriptions for soybeans, *Glycine max* (L.) Merrill. Crop Sci 11, 929. doi:10.2135/cropsci1971.0011183X001100060051x

Fluhr, R., Mattoo, A.K., Dilley, D.R., 1996. Ethylene — biosynthesis and perception. CRC. Crit. Rev. Plant Sci. 15, 479–523. doi:10.1080/07352689609382368

Forcat, S., Bennett, M.H., Mansfield, J.W., Grant, M.R., 2008. A rapid and robust method for simultaneously measuring changes in the phytohormones ABA, JA and SA in plants following biotic and abiotic stress. Plant Methods 4, 16. doi:10.1186/1746-4811-4-16

Fujimoto, S.Y., Ohta, M., Usui, A., Shinshi, H., Ohme-Takagi, M., 2000. Arabidopsis ethylene-responsive element binding factors act as transcriptional activators or repressors of GCC box-mediated gene expression. Plant Cell 12, 393–404. doi:10.1105/tpc.12.3.393

Gerashchenkov, G.A., Rozhnova, N.A., 2013. The involvement of phytohormones in the plant sex regulation. Russ. J. Plant Physiol. 60, 597–610. doi:10.1134/S1021443713050063

Glick, B.R., 2005. Modulation of plant ethylene levels by the bacterial enzyme ACC deaminase. FEMS Microbiol. Lett. 251, 1–7. doi:10.1016/j.femsle.2005.07.030

Guimarães-Dias, F., Neves-Borges, A.C., Viana, A.A.B., Mesquita, R.O., Romano, E., de Fátima Grossi-de-Sá, M., Nepomuceno, A.L., Loureiro, M.E., Alves-Ferreira, M., 2012. Expression analysis in response to drought stress in soybean: shedding light on the regulation of metabolic pathway genes. Genet. Mol. Biol. 35, 222–232. doi:10.1590/S1415-47572012000200004

Habben, J.E., Bao, X., Bate, N.J., DeBruin, J.L., Dolan, D., Hasegawa, D., Helentjaris, T.G., Lafitte, R.H., Lovan, N., Mo, H., Reimann, K., Schussler, J.R., 2014. Transgenic alteration of ethylene biosynthesis increases grain yield in maize under field drought-stress conditions. Plant Biotechnol. J. 12, 685–693. doi:10.1111/pbi.12172

Hahn, A., Harter, K., 2009. Mitogen-activated protein kinase cascades and ethylene: signaling, biosynthesis, or both? Plant Physiol. 149, 1207–1210. doi:10.1104/pp.108.132241

Hegg, E.L., Que, L., 1997. The 2-His-1-carboxylate facial triad--an emerging structural motif in mononuclear non-heme iron(II) enzymes. Eur. J. Biochem. 250, 625–629. doi:10.1111/j.1432-1033.1997.t01-1-00625.x

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., Vandesompele, J., 2007. qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. Genome Biol. 8, R19. doi:10.1186/gb-2007-8-2-r19

Hoagland, D.R., Arnon, D.I., 1950. The water-culture method for growing plants without soil. Circular. California Agricultural Experiment Station.

Hoffman, N.E., Yang, S.F., McKeon, T., 1982. Identification of 1-(malonylamino) cyclopropane-1-carboxylic acid as a major conjugate of 1-aminocyclopropane-1-carboxylic acid, an ethylene precursor in higher plants. Biochem. Biophys. Res. Commun. 104, 765–770. doi:10.1016/0006-291x(82)90703-3

Hua, J., Chang, C., Sun, Q., Meyerowitz, E.M., 1995. Ethylene insensitivity conferred by Arabidopsis ERS gene. Science 269, 1712–1714. doi:10.1126/science.7569898

Hua, J., Meyerowitz, E.M., 1998. Ethylene responses are negatively regulated by a receptor gene family in *Arabidopsis thaliana*. Cell 94, 261–271. doi:10.1016/s0092-8674(00)81425-7

Jiao, X.Z., Philosoph-Hadas, S., Su, L.Y., Yang, S.F., 1986. The conversion of 1-(malonylamino)-cyclopropane-1-carboxylic acid to 1-aminocyclopropane-1-carboxylic acid in plant tissues. Plant Physiol. 81, 637–641. doi:10.1104/pp.81.2.637

Jiao, Y., Lau, O.S., Deng, X.W., 2007. Light-regulated transcriptional networks in higher plants. Nat. Rev. Genet. 8, 217–230. doi:10.1038/nrg2049

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., Hunter, S., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240. doi:10.1093/bioinformatics/btu031

Ju, C., Yoon, G.M., Shemansky, J.M., Lin, D.Y., Ying, Z.I., Chang, J., Garrett, W.M., Kessenbrock, M., Groth, G., Tucker, M.L., Cooper, B., Kieber, J.J., Chang, C., 2012. CTR1 phosphorylates the central regulator EIN2 to control ethylene hormone signaling from the ER membrane to the nucleus in Arabidopsis. Proc. Natl. Acad. Sci. USA 109, 19486–19491. doi:10.1073/pnas.1214848109

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30. doi:10.1093/nar/28.1.27

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., Childs, K.L., Davidson, R.M., Lin, H., Quesada-Ocampo, L., Vaillancourt, B., Sakai, H., Lee, S.S., Kim, J., Numa, H., Itoh, T., Buell, C.R., Matsumoto, T., 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y) 6, 4. doi:10.1186/1939-8433-6-4

Kaya, C., Tuna, A.L., Yokaş, I., 2009. The role of plant hormones in plants under salinity stress, in: Ashraf, M., Ozturk, M., Athar, H.R. (Eds.), Salinity and water stress, Tasks for vegetation science. Springer Netherlands, Dordrecht, pp. 45–50. doi:10.1007/978-1-4020-9065-3_5

Kende, H., 1993. Ethylene biosynthesis. Annu. Rev. Plant Physiol. Plant Mol. Biol. 44, 283–307. doi:10.1146/annurev.pp.44.060193.001435

Kizis, D., Pagès, M., 2002. Maize DRE-binding proteins DBF1 and DBF2 are involved in *rab17* regulation through the drought-responsive element in an ABA-dependent pathway. Plant J. 30, 679–689. doi:10.1046/j.1365-313x.2002.01325.x

Klee, H.J., Hayford, M.B., Kretzmer, K.A., Barry, G.F., Kishore, G.M., 1991. Control of ethylene synthesis by expression of a bacterial enzyme in transgenic tomato plants. Plant Cell 3, 1187–1193. doi:10.1105/tpc.3.11.1187

Kulcheski, F.R., Marcelino-Guimaraes, F.C., Nepomuceno, A.L., Abdelnoor, R.V., Margis, R., 2010. The use of microRNAs as reference genes for quantitative polymerase chain reaction in soybean. Anal. Biochem. 406, 185–192. doi:10.1016/j.ab.2010.07.020

Leubner-Metzger, G., Petruzzelli, L., Waldvogel, R., Vögeli-Lange, R., Meins, F., 1998. Ethylene-responsive element binding protein (EREBP) expression and the transcriptional regulation of class I beta-1,3-glucanase during tobacco seed germination. Plant Mol. Biol. 38, 785–795. doi:10.1023/a:1006040425383

Liang, X., Abel, S., Keller, J.A., Shen, N.F., Theologis, A., 1992. The 1-aminocyclopropane-1-carboxylate synthase gene family of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA 89, 11046–11050. doi:10.1073/pnas.89.22.11046

Lim, S.H., Chang, S.C., Lee, J.S., Kim, S.-K., Kim, S.Y., 2002. Brassinosteroids affect ethylene production in the primary roots of maize (*Zea mays* L.). J. Plant Biol. 45, 148–153. doi:10.1007/BF03030307

Liu, M., Zhu, S., Zhou, J., 2012. Determination of 1-aminocyclopropane-1-carboxylic acid in apple and peach extracts by high performance liquid chromatography coupled to a fluorescence detector. Anal. Lett. 45, 2324–2333. doi:10.1080/00032719.2012.688083

Liu, Y., Zhang, S., 2004. Phosphorylation of 1-aminocyclopropane-1-carboxylic acid synthase by MPK6, a stress-responsive mitogen-activated protein kinase, induces ethylene biosynthesis in Arabidopsis. Plant Cell 16, 3386–3399. doi:10.1105/tpc.104.026609

Mainardi, J.A., Purgatto, E., Vieira, A., Bastos, W.A., Cordenunsi, B.R., Oliveira do Nascimento, J.R., Lajolo, F.M., 2006. Effects of ethylene and 1-methylcyclopropene (1-MCP) on gene expression and activity profile of alpha-1,4-glucan-phosphorylase during banana ripening. J. Agric. Food Chem. 54, 7294–7299. doi:10.1021/jf061180k

Martin, M.N., Cohen, J.D., Saftner, R.A., 1995. A new 1-aminocyclopropane-1-carboxylic acid-conjugating activity in tomato fruit. Plant Physiol. 109, 917–926. doi:10.1104/pp.109.3.917

McDonnell, L., Plett, J.M., Andersson-Gunnerås, S., Kozela, C., Dugardeyn, J., Van Der Straeten, D., Glick, B.R., Sundberg, B., Regan, S., 2009. Ethylene levels are regulated by a plant encoded 1-aminocyclopropane-1-carboxylic acid deaminase. Physiol. Plant. 136, 94–109. doi:10.1111/j.1399-3054.2009.01208.x

Miranda, V. de J., Coelho, R.R., Viana, A.A.B., de Oliveira Neto, O.B., Carneiro, R.M.D.G., Rocha, T.L., de Sa, M.F.G., Fragoso, R.R., 2013. Validation of reference genes aiming accurate normalization of qPCR data in soybean upon nematode parasitism and insect attack. BMC Res. Notes 6, 196. doi:10.1186/1756-0500-6-196

Miyazaki, J.H., Yang, S.F., 1987. Metabolism of 5-methylthioribose to methionine. Plant Physiol. 84, 277–281. doi:10.1104/pp.84.2.277

Nath, P., Trivedi, P.K., Sane, V.A., Sane, A.P., 2006. Role of ethylene in fruit ripening, in: Khan, N.A. (Ed.), Ethylene action in plants. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 151–184. doi:10.1007/978-3-540-32846-9_8

Oh, S.A., Park, J.H., Lee, G.I., Paek, K.H., Park, S.K., Nam, H.G., 1997. Identification of three genetic loci controlling leaf senescence in *Arabidopsis thaliana*. Plant J. 12, 527–535. doi:10.1046/j.1365-313x.1997.00527.x

Ohme-Takagi, M., Shinshi, H., 1995. Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. Plant Cell 7, 173–182. doi:10.1105/tpc.7.2.173

Olmedo, G., Guo, H., Gregory, B.D., Nourizadeh, S.D., Aguilar-Henonin, L., Li, H., An, F., Guzman, P., Ecker, J.R., 2006. ethylene-insensitive 5 encodes a 5’-3' exoribonuclease required for regulation of the EIN3-targeting F-box proteins EBF1/2. Proc. Natl. Acad. Sci. USA 103, 13286–13293. doi:10.1073/pnas.0605528103

Pfaffl, M.W., Horgan, G.W., Dempfle, L., 2002. Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res. 30, e36. doi:10.1093/nar/30.9.e36

Pommerrenig, B., Feussner, K., Zierer, W., Rabinovych, V., Klebl, F., Feussner, I., Sauer, N., 2011. Phloem-specific expression of Yang cycle genes and identification of novel Yang cycle enzymes in Plantago and Arabidopsis. Plant Cell 23, 1904–1919. doi:10.1105/tpc.110.079657

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., Finn, R.D., 2012. The Pfam protein families database. Nucleic Acids Res. 40, D290–301. doi:10.1093/nar/gkr1065

Qiao, H., Chang, K.N., Yazaki, J., Ecker, J.R., 2009. Interplay between ethylene, ETP1/ETP2 F-box proteins, and degradation of EIN2 triggers ethylene responses in Arabidopsis. Genes Dev. 23, 512–521. doi:10.1101/gad.1765709

Rodrigues, F.A., Marcolino-Gomes, J., de Fátima Corrêa Carvalho, J., do Nascimento, L.C., Neumaier, N., Farias, J.R.B., Carazzolle, M.F., Marcelino, F.C., Nepomuceno, A.L., 2012. Subtractive libraries for prospecting differentially expressed genes in the soybean under water deficit. Genet. Mol. Biol. 35, 304–314. doi:10.1590/S1415-47572012000200011

Roje, S., 2006. S-adenosyl-L-methionine: beyond the universal methyl group donor. Phytochemistry 67, 1686–1698. doi:10.1016/j.phytochem.2006.04.019

Sakai, H., Hua, J., Chen, Q.G., Chang, C., Medrano, L.J., Bleecker, A.B., Meyerowitz, E.M., 1998. *ETR2* is an *ETR1-like* gene involved in ethylene signaling in Arabidopsis. Proc. Natl. Acad. Sci. USA 95, 5812–5817. doi:10.1073/pnas.95.10.5812

Santner, A., Estelle, M., 2010. The ubiquitin-proteasome system regulates plant hormone signaling. Plant J. 61, 1029–1040. doi:10.1111/j.1365-313X.2010.04112.x

Sanz, L.C., Fernández-Maculet, J.C., Gómez, E., Vioque, B., Olías, J.M., 1993. Effect of methyl jasmonate on ethylene biosynthesis and stomatal closure in olive leaves. Phytochemistry 33, 285–289. doi:10.1016/0031-9422(93)85504-K

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., Jackson, S.A., 2010. Genome sequence of the palaeopolyploid soybean. Nature 463, 178–183. doi:10.1038/nature08670

Selkov, E., Grechkin, Y., Mikhailova, N., Selkov, E., 1998. MPW: the metabolic pathways database. Nucleic Acids Res. 26, 43–45. doi:10.1093/nar/26.1.43

Sharp, R.E., 2002. Interaction with ethylene: changing views on the role of abscisic acid in root and shoot growth responses to water stress. Plant Cell Environ. 25, 211–222. doi:10.1046/j.1365-3040.2002.00798.x

Shinozaki, K., Yamaguchi-Shinozaki, K., 2007. Gene networks involved in drought stress response and tolerance. J. Exp. Bot. 58, 221–227. doi:10.1093/jxb/erl164

Solano, R., Stepanova, A., Chao, Q., Ecker, J.R., 1998. Nuclear events in ethylene signaling: a transcriptional cascade mediated by ethylene-insensitive 3 and ethylene-response-factor 1. Genes Dev. 12, 3703–3714. doi:10.1101/gad.12.23.3703

Spollen, W.G., LeNoble, M.E., Samuels, T.D., Bernstein, N., Sharp, R.E., 2000. Abscisic acid accumulation maintains maize primary root elongation at low water potentials by restricting ethylene production. Plant Physiol. 122, 967–976. doi:10.1104/pp.122.3.967

Steed, C.L., Taylor, L.K., Harrison, M.A., 2004. Red light regulation of ethylene biosynthesis and gravitropism in etiolated pea stems. Plant Growth Regul. 43, 117–125. doi:10.1023/b:grow.0000040116.10016.c3

Stepanova, A.N., Alonso, J.M., 2009. Ethylene signaling and response: where different regulatory modules meet. Curr. Opin. Plant Biol. 12, 548–555. doi:10.1016/j.pbi.2009.07.009

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., Huala, E., 2008. The Arabidopsis information resource (TAIR): gene structure and function annotation. Nucleic Acids Res. 36, D1009–14. doi:10.1093/nar/gkm965

Tatsuki, M., Mori, H., 2001. Phosphorylation of tomato 1-aminocyclopropane-1-carboxylic acid synthase, LE-ACS2, at the C-terminal region. J. Biol. Chem. 276, 28051–28057. doi:10.1074/jbc.M101543200

Texeira, L.R., E Braccini, A.D.L., Sperandio, D., Scapim, C.A., Schuster, I., Viganó, J., 2008. Avaliação de cultivares de soja quanto à tolerância ao estresse hídrico em substrato contendo polietileno glicol. Acta Sci. Agron. 30. doi:10.4025/actasciagron.v30i2.1731

Thain, S.C., Vandenbussche, F., Laarhoven, L.J.J., Dowson-Day, M.J., Wang, Z.-Y., Tobin, E.M., Harren, F.J.M., Millar, A.J., Van Der Straeten, D., 2004. Circadian rhythms of ethylene emission in Arabidopsis. Plant Physiol. 136, 3751–3761. doi:10.1104/pp.104.042523

Trusov, Y., Botella, J.R., 2006. Silencing of the ACC synthase gene *ACACS2* causes delayed flowering in pineapple [*Ananas comosus* (L.) Merr.]. J. Exp. Bot. 57, 3953–3960. doi:10.1093/jxb/erl167

Tucker, M.L., Xue, P., Yang, R., 2010. 1-aminocyclopropane-1-carboxylic acid (ACC) concentration and ACC synthase expression in soybean roots, root tips, and soybean cyst nematode (*Heterodera glycines*)-infected roots. J. Exp. Bot. 61, 463–472. doi:10.1093/jxb/erp317

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R., Leunissen, J.A.M., 2007. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. 35, W71–4. doi:10.1093/nar/gkm306

Vandenbussche, F., Vaseva, I., Vissenberg, K., Van Der Straeten, D., 2012. Ethylene in vegetative development: a tale with a riddle. New Phytol. 194, 895–909. doi:10.1111/j.1469-8137.2012.04100.x

Vandenbussche, F., Vriezen, W.H., Smalle, J., Laarhoven, L.J.J., Harren, F.J.M., Van Der Straeten, D., 2003. Ethylene and auxin control the Arabidopsis response to decreased light intensity. Plant Physiol. 133, 517–527. doi:10.1104/pp.103.022665

Voesenek, L.A.C.J., Bailey-Serres, J., 2009. Plant biology: genetics of high-rise rice. Nature 460, 959–960. doi:10.1038/460959a

Wang, K.L.-C., Yoshida, H., Lurin, C., Ecker, J.R., 2004. Regulation of ethylene gas biosynthesis by the Arabidopsis ETO1 protein. Nature 428, 945–950. doi:10.1038/nature02516

Wilkinson, S., Kudoyarova, G.R., Veselov, D.S., Arkhipova, T.N., Davies, W.J., 2012. Plant hormone interactions: innovative targets for crop breeding and management. J. Exp. Bot. 63, 3499–3509. doi:10.1093/jxb/ers148

Wilmowicz, E., Kesy, J., Kopcewicz, J., 2008. Ethylene and ABA interactions in the regulation of flower induction in *Pharbitis nil*. J. Plant Physiol. 165, 1917–1928. doi:10.1016/j.jplph.2008.04.009

Wittkopp, P.J., Kalay, G., 2011. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. 13, 59–69. doi:10.1038/nrg3095

Xu, Z.-S., Xia, L.-Q., Chen, M., Cheng, X.-G., Zhang, R.-Y., Li, L.-C., Zhao, Y.-X., Lu, Y., Ni, Z.-Y., Liu, L., Qiu, Z.-G., Ma, Y.-Z., 2007. Isolation and molecular characterization of the *Triticum aestivum* L. ethylene-responsive factor 1 (TaERF1) that increases multiple stress tolerance. Plant Mol. Biol. 65, 719–732. doi:10.1007/s11103-007-9237-9

Yamagami, T., Tsuchisaka, A., Yamada, K., Haddon, W.F., Harden, L.A., Theologis, A., 2003. Biochemical diversity among the 1-amino-cyclopropane-1-carboxylate synthase isozymes encoded by the Arabidopsis gene family. J. Biol. Chem. 278, 49102–49112. doi:10.1074/jbc.M308297200

Yamaguchi-Shinozaki, K., Shinozaki, K., 1994. A novel *cis*-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. Plant Cell 6, 251–264. doi:10.1105/tpc.6.2.251

Yang, S.F., Hoffman, N.E., 1984. Ethylene biosynthesis and its regulation in higher plants. Annu. Rev. Plant Physiol. 35, 155–189. doi:10.1146/annurev.pp.35.060184.001103

Yoo, S.-D., Cho, Y.-H., Tena, G., Xiong, Y., Sheen, J., 2008. Dual control of nuclear EIN3 by bifurcate MAPK cascades in C_2H_4 signalling. Nature 451, 789–795. doi:10.1038/nature06543

Zhang, G., Chen, M., Li, L., Xu, Z., Chen, X., Guo, J., Ma, Y., 2009. Overexpression of the soybean GmERF3 gene, an AP2/ERF type transcription factor for increased tolerances to salt, drought, and diseases in transgenic tobacco. J. Exp. Bot. 60, 3781–3796. doi:10.1093/jxb/erp214

Zhao, S., Fernald, R.D., 2005. Comprehensive algorithm for quantitative real-time polymerase chain reaction. J. Comput. Biol. 12, 1047–1064. doi:10.1089/cmb.2005.12.1047

SUPPLEMENTARY MATERIAL

Due to their large size, the figures and other supplementary files in this Chapter are permanently deposited at the following link. Therefore, only the subtitles/titles of the files are provided here:

https://bmcplantbiol.biomedcentral.com/articles/10.1186/s12870-015-0597-z#Sec18

Additional File 1: Supplementary Figures (Figures S1-S10). Figure S1. Soybean chromosomal ideogram. In figure, are represented the positions of 176 genes identified in 20 soybean chromosome. Each gene is represented by a generic name (see Tables S5 and S6 in Additional File 2), and the relative position in the chromosome is determined by the color code (names in **black** - plus strand; names in **red** - minus strand). The **yellow** and **red** lines represent respectively genes related to ethylene biosynthesis and signal transduction mediated by this phytohormone. The **dotted black lines** pass through the centromere midpoint in each chromosome. Scale: 1.0 cm equates to 5.0 Mb (megabase); **Figure S2. Gene ontology classification.** The three graphs show the ontological subgroups (level 2) of the 176 soybean selected sequences. *Legends*: **A** - cellular component; **B** - molecular function; **C** - biological process; **Figure S3. Protein orthology by Best Bidirectional Hit (BBH) analysis.** Percentage of soybean BBH positive proteins present in each group analyzed. ^{1} **EBS** and ^{2} **EST** - ethylene biosynthesis and signal transduction proteins, respectively; ^{3} **ALL** - overall BBH positive percentage (considering all soybean proteins analyzed); ^{4} **Double Positive** - soybean proteins BBH positive with *A. thaliana* and *O. sativa* simultaneously. The proteins identified in this experiment are listed in Tables S7 and S8 (Additional File 2); **Figure S4. ACSs Classification.** The figure shows the relationship among amino acid sequence of ACSs (1-aminocyclopropane-

1-carboxilic acid synthase) identified in soybean, *Arabidopsis thaliana* and *Oryza sativa*. This relation allows the classification according to the presence/absence of potential sites of phosphorylation by calcium dependent protein kinase (CPK or CDPK) and/or MPK6 protein (mitogen-activated protein kinase 6 - MAPK6) in C-terminal of these proteins. Thus, these amino acid sequences can be divided into three classes: *type I* (**red circles**; **model A**) - proteins which exhibit extended C-terminal with conserved residues that are targets for phosphorylation by MPK6, as well as a conserved residue that is a phosphorylation site for CPK; *type II* (**yellow circles**; **model B**) - proteins which exhibit only CPK sites; and *type III* (**blue circles**; **model C**) – proteins that lacking both phosphorylation sites (Liu and Zhang, 2004; Tucker et al., 2010). The proteins represented by **green circles** are classified as ACS-like since AtACS10 and AtACS12 possibly do not have ACS activity and are most probably amino acid transferases (AATs) (Yamagami et al., 2003). AtACS1 also does not have ACS activity, by deletions in catalytic core, but AtACS2 does (Yamagami et al., 2003). The **gray rectangle** highlights C-terminal of catalytic core (position at left), the **blue rectangle** the CPK phosphorylation sites and the **orange rectangle** the MPK6 sites. The **underlined amino acid residues** are the most likely to be phosphorylated in each sequence. Each protein is identified by generic name (see Tables S1, S3 and S5 in Additional File 2); **Figure S5. Differential expression of genes related to soybean ethylene biosynthesis in transcriptomes under drought stress conditions.** The graphics represent the expression levels of genes related to ethylene biosynthesis in root and leaf transcriptomes of two soybean cultivars: BR16 and EMBRAPA48, sensitive and tolerant to drought stress, respectively. Each gene is identified by generic name (see Table S5 in Additional File 2). The symbols correspond to: **A1** - root/25 at 50 minutes under drought conditions; **A2** - leaf/25 at 50 minutes under drought conditions; **B1** - root/75 at 100 minutes under drought conditions; **B2** - leaf/75 at 100 minutes under drought conditions; **C1** - root/125 at 150 minutes under drought conditions; **C2** - leaf/125 at 150 minutes under drought conditions; **FPKM** - fragments per kilobase of transcript per million fragments mapped; **Figure S6. Differential expression of genes related to soybean ethylene signal transduction in transcriptomes under drought stress conditions.** The graphics represent the expression levels of genes related to ethylene signal transduction in root and leaf transcriptomes of two soybean cultivars: BR16 and EMBRAPA48, sensitive and tolerant to drought stress, respectively. Each gene is identified by generic name (see Table S6 in Additional File 2). The symbols correspond to: **A1** - root/25 at 50 minutes under drought conditions; **A2** - leaf/25 at 50 minutes under drought conditions; **B1** - root/75 at 100 minutes under drought conditions; **B2** - leaf/75 at 100 minutes under drought conditions; **C1** - root/125 at 150 minutes under drought conditions; **C2** - leaf/125 at 150 minutes under drought conditions; **FPKM** - fragments per kilobase of transcript per million fragments mapped; **Figure S7. <u>Comparison of ethylene biosynthesis and signaling differential gene expression among similar tissues in soybean cultivars under drought stress conditions.</u>** The scatter plots compare the expression levels of ethylene biosynthesis and signal transduction genes among similar tissues of BR16 and EMBRAPA48 soybean cultivars, sensitive and tolerant to drought stress, respectively. Every expressed gene (see transcriptome data in Figures S5 and S6) is represented in each plot by one point whose coordinates correspond to expression levels in similar tissues of both cultivars. The symbols correspond to: **A1** - root/25 at 50 minutes under drought conditions; **A2** - leaf/25 at 50 minutes under drought conditions; **B1** - root/75 at 100 minutes under drought conditions; **B2** - leaf/75 at 100 minutes under drought conditions; **C1** - root/125 at 150 minutes under drought conditions; **C2** - leaf/125 at 150 minutes under drought conditions; **FPKM** - fragments per kilobase of transcript per million fragments mapped; **Figure S8. <u>Expression of ethylene-related genes in soybean under drought stress conditions.</u>** The graphs show the expression levels, obtained by RT-qPCR, of five soybean genes related to ethylene biosynthesis [*MAT* (**A**), *ACS* (**B**) and *ACO* (**C**)] and ethylene signal transduction [*ETR* (**D**) and *CTR* (**E**)]. The expression of these genes in the experiment was compared in roots and leaves of soybean cultivars BR16 and EMBRAPA48 after different durations of drought stress. The statistics were obtained by comparing non-stressed plants (time zero) with stressed plants (at different times of drought stress). The asterisks represent the level of statistical significance: (*****) $p \leq 0.05$; (******) $0.01 \leq p < 0.05$; (*******) $0.001 \leq p < 0.01$. Each dot represents the average amount (± standard error) of three experimental replicates (same sample) in three biological samples (different plants), totaling nine replicates. The standard error is not presented with some of the dots because their absolute values are lower than the scale. After normalization based on housekeeping genes, the values given in the graph are relative to the lowest expression, whose value was set at 1 (one). Information about the target genes is presented in Method S1 (Additional File 6); **Figure S9. <u>Evaluation of physiological parameters in soybean cultivars under drought stress conditions.</u>** During the drought stress experiments in BR16 and EMBRAPA48 soybean cultivars, grown under hydroponic conditions, were determined some relevant physiological parameters: **A** - photosynthetic rate (*A*); **B** - photosynthetically active radiation (internal to the reading chamber - PAR_i); **C** - photosynthetically active radiation (external to the reading chamber - PAR_o); **D** - intercellular CO_2 concentration (C_i); **E** - conductance to H_2O (or estomatic conductance - g_s); **F** - transpiration rate (*E*); **G** - temperature variation (ΔT), where [$\Delta T = T_{air}$ (internal to the reading chamber - T_{air}) – T_{leaf}; average air temperature = 29.7±1.9ºC]; **H** - water use efficiency (*WUE*, ratio among photosynthetic and transpiration rates - *A/E*). Each dot represents the average amount (± standard error) of three replicates in different plants. The absence of representation of standard error occurs in some dots by the fact of their absolute values are lower than the scale; **Figure S10. Levels of ethylene production and free ACC in soybean under drought stress conditions.** Values were determined for ethylene production and free ACC (1-aminocyclopropane-1-carboxylic acid) in roots and leaves of soybean cultivars BR16 and EMBRAPA48 after the application of different durations of drought stress. The codes **A** represents levels of free ACC; **B** represents levels of ethylene production. The statistics were obtained by comparing non-stressed plants (time zero) with stressed plants (at different durations of drought stress). The asterisks represent the level of statistical significance: (*****) $p \leq 0.05$; (******) $0.05 < p \leq 0.01$; (*******) $0.01 < p \leq 0.001$. Each dot represents the average amount (± standard error) of three replicates in different plants. The standard error is not presented with some dots because their absolute values are lower than the scale.

Additional File 2: Supplementary Tables with protein summary and Best Bidirectional Hit (BBH) results (Tables S1-S8). Table S1. *Arabidopsis thaliana* ethylene biosynthesis protein list; **Table S2.** *Arabidopsis thaliana* ethylene signal transduction protein list; **Table S3.** *Oryza sativa* ethylene biosynthesis protein list; **Table S4.** *Oryza sativa* ethylene signal transduction protein list; **Table S5.** Soybean ethylene biosynthesis protein summary; **Table S6.** Soybean ethylene signal transduction protein summary; **Table S7.** BBH Experiment - Soybean ethylene biosynthesis proteins; **Table S8.** BBH Experiment - Soybean proteins related with ethylene signal transduction.

Additional File 3: *In silico* Characterization of ethylene soybean genes. Detailed description of characterization, gene localization and gene onthology (GO) of ethylene soybean genes.

Additional File 4 (excel file): Identification of *cis*-acting elements in soybean putative gene promoters (Table S9). In this table are shown the *cis*-acting elements present in the putative promoters of 176 analyzed genes. The analysis matrix was composed by 100 different elements, distributed in 29 families. Cells highlighted in different colors corresponding to elements in each promoter sequence identified, associated with the number of the identified elements. Thus, sequences with green and red cells respectively represent putative gene promoters in ethylene biosynthesis and signal transduction mediated by this plant hormone. All results are presented in three ways: the total number of matches for each *cis*-acting element in the matrix analyzed (column E); the number of different sequences of putative promoters that represent each *cis*-acting element that compose the matrix (column F); and, at last, the number of different sequences of putative promoters that represent each *cis*-acting element family (column G). These values are totaled, and the total number of matches identified in each sequence analyzed (line 107). (******) N.A. corresponds to not applicable.

Additional File 5: Real Time PCR (RT-qPCR) cycle threshold (C_t) (Tables S10-S11). Table S10. Target and endogenous gene cycle threshold in soybean leaf under drought stress – RT-qPCR; **Table S11.** Target and endogenous gene cycle threshold in soybean root under drought stress – RT-qPCR.

Additional File 6: Real Time PCR (RT-qPCR) primers (Method S1). Gene summary and primers for Real Time PCR.

6. FINAL CONCLUSIONS

In Chapter 01, the study of the core elements of the insect RNAi machinery suggests, for the five orders analyzed, that variability in such elements is an important factor in gene silencing efficiency mediated by this metabolic pathway. By far, compared to the others, Lepidoptera is the most distant order, presenting distinct characteristics of the model species *D. melanogaster*. Among the domains analyzed, dsrm, PAZ, Platform, Ribonuclease III (RIIID) and Helicase, provided the most information about the identified variability, mainly by the presence of mutations in crucial regions for the described activity and the presence of unresolved regions (loops), which can positively and negatively influence the function of each domain. Thus, the stability of the microprocessor complexes responsible for miRNA and siRNA production in insects, whose main components are DCR1, DCR2 and DROSHA, together with their accessory proteins (LOQS, PASHA and R2D2), are the key point in biogenesis efficiency of these small RNAs. This hypothesis corroborates studies that have characterized novel proteins, such as Coleoptera-specific Staufen C, which possibly stabilizes the microprocessor siRNA and can reverse RNAi resistance events. New studies that aim to better characterize the RNAi machinery components of non-model species are essential to identify more particularities in the silencing mechanism and optimize its use as a biotechnological tool.

In Chapter 02 it was possible to highlight points that still need to be better studied related to ethylene, as the characterization of the element of this phytohormone biosynthesis pathway, as regards ethylene participation in drought tolerance. According to the data, in abrupt water deficit conditions, the biosynthesis and signaling mediated by ethylene could be reflection of the cultivation conditions (hydroponics), which may not be ideal for such analysis. The data also suggest that the higher susceptibility to drought of soybean cultivar BR16 may be associated with its lower sensitivity to abscisic acid, due to the detection of high levels of transcripts associated with inhibition of such metabolic pathway. Thus, this study concludes that the regulation of ethylene-mediated response in soybean is influenced by several endogenous and exogenous factors, and the balance between signaling mediated by these factors may determine the quality and intensity of response to these stimuli. Therefore, the *in vitro* and *in vivo* functional elucidation of the molecular mechanisms and metabolic networks coordinated by this phytohormone is essential.

Thus, the analysis of genomic and transcriptomic data from plants and their pathogens can contribute significantly to global agribusiness, mainly by providing knowledge for the generation of tools that optimize sustainably crops breeding.

APPENDIX I

SCIENTIFIC PRODUCTION (2020 – 2015)

OTHER PUBLISHED PAPERS

CHAPTERS IN BOOKS

PLOS ONE

RESEARCH ARTICLE

Comparative gut transcriptome analysis of *Diatraea saccharalis* in response to the dietary source

Daniel D. Noriega[1,2,3]*, Fabricio B. M. Arraes[1,4], José Dijair Antonino[1,5], Leonardo L. P. Macedo[1], Fernando C. A. Fonseca[1,2], Roberto C. Togawa[1], Priscila Grynberg[1], Maria C. M. Silva[1], Aldomario S. Negrisoli Jr[6], Carolina V. Morgante[1,7], Maria F. Grossi-de-Sa[1,3,8]*

1 Embrapa Genetic Resources and Biotechnology, Brasília-DF, Brazil, **2** Department of Cellular Biology, University of Brasília, Brasília-DF, Brazil, **3** Catholic University of Brasília, Brasília-DF, Brazil, **4** Biotechnology Center, UFRGS, Porto Alegre-RS, Brazil, **5** Departamento de Agronomia/Entomologia, UFRPE, Recife-PE, Brazil, **6** Embrapa Tabuleiros Costeiros, Aracaju, Sergipe-SE, Brazil, **7** Embrapa Semi Arid, Petrolina-PE, Brazil, **8** National Institute of Science and Technology–INCT PlantStress Biotech–EMBRAPA, Brasilia-DF, Brazil

* fatima.grossi@embrapa.br (MFGS); daniel.nv07@gmail.com (DDN)

Citation: Noriega DD, Arraes FBM, Antonino JD, Macedo LLP, Fonseca FCA, Togawa RC, et al. (2020) Comparative gut transcriptome analysis of *Diatraea saccharalis* in response to the dietary source. PLoS ONE 15(8): e0235575. https://doi.org/10.1371/journal.pone.0235575

Editor: Pedro L. Oliveira, Universidade Federal do Rio de Janeiro, BRAZIL

Received: April 13, 2020

Accepted: June 17, 2020

Published: August 3, 2020

Data Availability Statement: All fastq files from Illumina raw data are available in the NCBI database. The accession numbers are: SRR10081745, SRR10081744, SRR10081743, SRR10081742, SRR10081741, SRR10081740, SRR10081739 Bioproject: PRJNA564321.

Funding: The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The sugarcane borer (*Diatraea saccharalis*, Fabricius, *1794*) is a devastating pest that causes millions of dollars of losses each year to sugarcane producers by reducing sugar and ethanol yields. The control of this pest is difficult due to its endophytic behavior and rapid development. Pest management through biotechnological approaches has emerged in recent years as an alternative to currently applied methods. Genetic information about the target pests is often required to perform biotechnology-based management. The genomic and transcriptomic data for *D. saccharalis* are very limited. Herein, we report a tissue-specific transcriptome of *D. saccharalis* larvae and a differential expression analysis highlighting the physiological characteristics of this pest in response to two different diets: sugarcane and an artificial diet. Sequencing was performed on the Illumina HiSeq 2000 platform, and a *de novo* assembly was generated. A total of 27,626 protein-coding unigenes were identified, among which 1,934 sequences were differentially expressed between treatments. Processes such as defence, digestion, detoxification, signaling, and transport were highly represented among the differentially expressed genes (DEGs). Furthermore, seven aminopeptidase genes were identified as candidates to encode receptors of Cry proteins, which are toxins of *Bacillus thuringiensis* used to control lepidopteran pests. Since plant-insect interactions have produced a considerable number of adaptive responses in hosts and herbivorous insects, the success of phytophagous insects relies on their ability to overcome challenges such as the response to plant defences and the intake of nutrients. In this study, we identified metabolic pathways and specific genes involved in these processes. Thus, our data strongly contribute to the knowledge advancement of insect transcripts, which can be a source of target genes for pest management.

frontiers in Physiology

ORIGINAL RESEARCH
published: 28 October 2020
doi: 10.3389/fphys.2020.588450

Transcriptome Analysis and Knockdown of the Juvenile Hormone Esterase Gene Reveal Abnormal Feeding Behavior in the Sugarcane Giant Borer

Daniel D. Noriega[1,2,3], Fabricio B. M. Arraes[1,4], José Dijair Antonino[1,5], Leonardo L. P. Macedo[1], Fernando C. A. Fonseca[1,2], Roberto C. Togawa[1], Priscila Grynberg[1], Maria C. M. Silva[1], Aldomario S. Negrisoli[6] and Maria F. Grossi-de-Sa[1,3,7]**

[1] Embrapa Genetic Resources and Biotechnology, Brasília, Brazil, [2] Department of Cellular Biology, University of Brasília, Brasília, Brazil, [3] PPG in Genomic Sciences and Biotechnology, Catholic University of Brasília, Brasília, Brazil, [4] Biotechnology Center, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil, [5] Department of Agronomy/Entomology, Universidade Federal Rural de Pernambuco (UFRPE), Recife, Brazil, [6] Embrapa Coastal Tablelands, Aracaju, Brazil, [7] National Institute of Science and Technology (INCT) PlantStress Biotech, Brazilian Agricultural Research Corporation (EMBRAPA), Brasília, Brazil

OPEN ACCESS

Edited by:
Jin-Jun Wang,
Southwest University, China

Reviewed by:
Shaoli Wang,
Insititute of Vegetables and Flowers (CAAS), China
Mohammad Mehrabadi,
Tarbiat Modares University, Iran

***Correspondence:**
Maria F. Grossi-de-Sa
fatima.grossi@embrapa.br
Daniel D. Noriega
daniel.nv07@gmail.com

Specialty section:
This article was submitted to Invertebrate Physiology, a section of the journal Frontiers in Physiology

Received: 28 July 2020
Accepted: 06 October 2020
Published: 28 October 2020

Citation:
Noriega DD, Arraes FBM, Antonino JD, Macedo LLP, Fonseca FCA, Togawa RC, Grynberg P, Silva MCM, Negrisoli AS and Grossi-de-Sa MF (2020) Transcriptome Analysis and Knockdown of the Juvenile Hormone Esterase Gene Reveal Abnormal Feeding Behavior in the Sugarcane Giant Borer. Front. Physiol. 11:588450. doi: 10.3389/fphys.2020.588450

The sugarcane giant borer (SGB), *Telchin licus licus*, is a pest that has strong economic relevance for sugarcane producers. Due to the endophytic behavior of the larva, current methods of management are inefficient. A promising biotechnological management option has been proposed based on RNA interference (RNAi), a process that uses molecules of double-stranded RNA (dsRNA) to specifically knock down essential genes and reduce insect survival. The selection of suitable target genes is often supported by omic sciences. Studies have shown that genes related to feeding adaptation processes are good candidates to be targeted by RNAi for pest management. Among those genes, esterases are highlighted because of their impact on insect development. In this study, the objective was to evaluate the transcriptome responses of the SGB's gut in order to provide curated data of genes that could be used for pest management by RNAi in future studies. Further, we validated the function of an esterase-coding gene and its potential as a target for RNAi-based control. We sequenced the gut transcriptome of SGB larvae by Illumina HiSeq and evaluated its gene expression profiles in response to different diets (sugarcane stalk and artificial diet). We obtained differentially expressed genes (DEGs) involved in detoxification, digestion, and transport, which suggest a generalist mechanism of adaptation in SGB larvae. Among the DEGs, was identified and characterized a candidate juvenile hormone esterase gene (*Tljhe*). We knocked down the *Tljhe* gene by oral delivery of dsRNA molecules and evaluated gene expression in the gut. The survival and nutritional parameters of the larvae were measured along the developmental cycle of treated insects. We found that the gene *Tljhe* acts as a regulator of feeding behavior. The knockdown of *Tljhe* triggered a forced starvation state in late larval instars that significantly reduced the

frontiers
in Physiology

ORIGINAL RESEARCH
published: 16 November 2020
doi: 10.3389/fphys.2020.591569

RNAi-Mediated Suppression of *Laccase2* Impairs Cuticle Tanning and Molting in the Cotton Boll Weevil (*Anthonomus grandis*)

*Alexandre Augusto Pereira Firmino[1,2†], Daniele Heloísa Pinheiro[1†], Clidia Eduarda Moreira-Pinto[1,3], José Dijair Antonino[1,4], Leonardo Lima Pepino Macedo[1], Diogo Martins-de-Sa[3], Fabrício Barbosa Monteiro Arraes[1,5,6], Roberta Ramos Coelho[1], Fernando Campos de Assis Fonseca[1,3], Maria Cristina Mattar Silva[1,6], Janice de Almeida Engler[6,7], Marília Santos Silva[1], Isabela Tristan Lourenço-Tessutti[1], Walter Ribeiro Terra[8] and Maria Fátima Grossi-de-Sa[1,6,9]**

[1] Embrapa Genetic Resources and Biotechnology, Brasília, Brazil, [2] Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany, [3] Department of Cell Biology, Federal University of Brasília (UnB), Brasília, Brazil, [4] Departamento de Agronomia/Entomologia, Universidade Federal Rural de Pernambuco (UFRPE), Recife, Brazil, [5] Department of Cellular and Molecular Biology, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil, [6] National Institute of Science and Technology – INCT PlantStress Biotech – Embrapa, Brasilia, Brazil, [7] Département Santé des Plantes et Environnement, Institut National de la Recherche Agronomique and Institut Sophia Agrobiotech, Sophia Antipolis, France, [8] Department of Chemistry, University of São Paulo, São Paulo, Brazil, [9] Department of Biological Sciences, Catholic University o Brasília (UCB), Brasília, Brazil

The cotton boll weevil, *Anthonomus grandis*, is the most economically important pest of cotton in Brazil. Pest management programs focused on *A. grandis* are based mostly on the use of chemical insecticides, which may cause serious ecological impacts. Furthermore, *A. grandis* has developed resistance to some insecticides after their long-term use. Therefore, alternative control approaches that are more sustainable and have reduced environmental impacts are highly desirable to protect cotton crops from this destructive pest. RNA interference (RNAi) is a valuable reverse genetics tool for the investigation of gene function and has been explored for the development of strategies to control agricultural insect pests. This study aimed to evaluate the biological role of the *Laccase2* (*AgraLac2*) gene in *A. grandis* and its potential as an RNAi target for the control of this insect pest. We found that *AgraLac2* is expressed throughout the development of *A. grandis* with significantly higher expression in pupal and adult developmental stages. In addition, the immunolocalization of the AgraLac2 protein in third-instar larvae using specific antibodies revealed that AgraLac2 is distributed throughout the epithelial tissue, the cuticle and the tracheal system. We also verified that the knockdown of *AgraLac2* in *A. grandis* resulted in an altered cuticle tanning process, molting defects and arrested development. Remarkably, insects injected with ds*AgraLac2* exhibited defects in cuticle hardening and pigmentation. As a consequence, the development of ds*AgraLac2*-treated insects was compromised, and in cases of severe phenotypic defects, the insects subsequently died. On the contrary,

OPEN ACCESS

Edited by:
Senthil-Nathan, Sengottayan, Manonmaniam Sundaranar University, India

Reviewed by:
Guy Smagghe, Ghent University, Belgium
Subbaratnam Muthukrishnan, Kansas State University, United States

***Correspondence:**
Maria Fátima Grossi-de-Sa
fatima.grossi@embrapa.br

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Invertebrate Physiology, a section of the journal Frontiers in Physiology

Received: *04 August 2020*
Accepted: *20 October 2020*
Published: *16 November 2020*

Citation:
Firmino AAP, Pinheiro DH, Moreira-Pinto CE, Antonino JD, Macedo LLP, Martins-de-Sa D, Arraes FBM, Coelho RR, Fonseca FCdA, Silva MCM, Engler JdA, Silva MS, Lourenço-Tessutti IT, Terra WR and Grossi-de-Sa MF (2020) RNAi-Mediated Suppression of Laccase2 Impairs Cuticle Tanning and Molting in the Cotton Boll Weevil (Anthonomus grandis). Front. Physiol. 11:591569. doi: 10.3389/fphys.2020.591569

Molecular Genetics and Genomics
https://doi.org/10.1007/s00438-020-01677-7

ORIGINAL ARTICLE

Evolutionarily conserved plant genes responsive to root-knot nematodes identified by comparative genomics

Ana Paula Zotta Mota[1,2] · Diana Fernandez[1,3] · Fabricio B. M. Arraes[1,2] · Anne-Sophie Petitot[3] · Bruno Paes de Melo[1,4] · Maria E. Lisei de Sa[1,5] · Priscila Grynberg[1] · Mario A. Passos Saraiva[1] · Patricia Messenberg Guimaraes[1] · Ana Cristina Miranda Brasileiro[1] · Erika Valeria Saliba Albuquerque[1] · Etienne G. J. Danchin[6] · Maria Fatima Grossi-de-Sa[1,7]

Received: 13 November 2019 / Accepted: 4 April 2020

Abstract

Root-knot nematodes (RKNs, genus *Meloidogyne*) affect a large number of crops causing severe yield losses worldwide, more specifically in tropical and sub-tropical regions. Several plant species display high resistance levels to *Meloidogyne,* but a general view of the plant immune molecular responses underlying resistance to RKNs is still lacking. Combining comparative genomics with differential gene expression analysis may allow the identification of widely conserved plant genes involved in RKN resistance. To identify genes that are evolutionary conserved across plant species, we used OrthoFinder to compared the predicted proteome of 22 plant species, including important crops, spanning 214 Myr of plant evolution. Overall, we identified 35,238 protein orthogroups, of which 6,132 were evolutionarily conserved and universal to all the 22 plant species (PLAnts Common Orthogroups—PLACO). To identify host genes responsive to RKN infection, we analyzed the RNA-seq transcriptome data from RKN-resistant genotypes of a peanut wild relative (*Arachis stenosperma*), coffee (*Coffea arabica* L.), soybean (*Glycine max* L.*),* and African rice (*Oryza glaberrima* Steud.) challenged by *Meloidogyne* spp. using EdgeR and DESeq tools, and we found 2,597 (*O. glaberrima*), 743 (*C. arabica*), 665 (*A. stenosperma*), and 653 (*G. max*) differentially expressed genes (DEGs) during the resistance response to the nematode. DEGs' classification into the previously characterized 35,238 protein orthogroups allowed identifying 17 orthogroups containing at least one DEG of each resistant *Arachis*, coffee, soybean, and rice genotype analyzed. Orthogroups contain 364 DEGs related to signaling, secondary metabolite production, cell wall-related functions, peptide transport, transcription regulation, and plant defense, thus revealing evolutionarily conserved RKN-responsive genes. Interestingly, the 17 DEGs-containing orthogroups (belonging to the PLACO) were also universal to the 22 plant species studied, suggesting that these core genes may be involved in ancestrally conserved immune responses triggered by RKN infection. The comparative genomic approach that we used here represents a promising predictive tool for the identification of other core plant defense-related genes of broad interest that are involved in different plant–pathogen interactions.

Keywords Transcriptome · *Meloidogyne* · *Arachis* · Soybean · Coffee · Rice

Introduction

Plants must overcome a wide range of biotic stresses in their natural habitats using a sophisticated perception and immune response system. Under intensive agricultural production, biotic stress conditions can drastically affect plant growth and development, leading to a severe decrease in crop yield. From 2005 to 2015, the effects of climate change substantially affected the agriculture, causing losses of up to $100 billion dollars, in part because of the attack of plants by

Communicated by Stefan Hohmann.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00438-020-01677-7) contains supplementary material, which is available to authorized users.

✉ Maria Fatima Grossi-de-Sa
fatima.grossi@embrapa.br

Extended author information available on the last page of the article

Published online: 24 April 2020

frontiers in Plant Science

ORIGINAL RESEARCH
published: 15 May 2020
doi: 10.3389/fpls.2020.00509

Insights Into Genetic and Molecular Elements for Transgenic Crop Development

*Marcos Fernando Basso[1], Fabrício Barbosa Monteiro Arraes[1,2], Maíra Grossi-de-Sa[1], Valdeir Junio Vaz Moreira[1,2], Marcio Alves-Ferreira[3] and Maria Fatima Grossi-de-Sa[1,4]**

[1] Plant Biotechnology, Embrapa Genetic Resources and Biotechnology, Brasília, Brazil, [2] Department of Molecular Biology and Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, Brazil, [3] Department of Genetic, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil, [4] Department of Genomic Sciences and Biotechnology, Catholic University of Brasilia, Brasília, Brazil

Climate change and the exploration of new areas of cultivation have impacted the yields of several economically important crops worldwide. Both conventional plant breeding based on planned crosses between parents with specific traits and genetic engineering to develop new biotechnological tools (NBTs) have allowed the development of elite cultivars with new features of agronomic interest. The use of these NBTs in the search for agricultural solutions has gained prominence in recent years due to their rapid generation of elite cultivars that meet the needs of crop producers, and the efficiency of these NBTs is closely related to the optimization or best use of their elements. Currently, several genetic engineering techniques are used in synthetic biotechnology to successfully improve desirable traits or remove undesirable traits in crops. However, the features, drawbacks, and advantages of each technique are still not well understood, and thus, these methods have not been fully exploited. Here, we provide a brief overview of the plant genetic engineering platforms that have been used for proof of concept and agronomic trait improvement, review the major elements and processes of synthetic biotechnology, and, finally, present the major NBTs used to improve agronomic traits in socioeconomically important crops.

Keywords: new biotechnological tools, plant genetic transformation, tissue culture, minimal expression cassette, T-DNA delivery

OPEN ACCESS

Edited by:
Junhua Peng,
Huazhi Rice Bio-Tech Co., Ltd., China

Reviewed by:
Ahmad Arzani,
Isfahan University of Technology, Iran
Fengqi Li,
Beijing Academy of Agricultural and Forestry Sciences, China

****Correspondence:***
Maria Fatima Grossi-de-Sa
fatima.grossi@embrapa.br

Specialty section:
This article was submitted to Plant Biotechnology, a section of the journal Frontiers in Plant Science

Received: *17 September 2019*
Accepted: *03 April 2020*
Published: *15 May 2020*

Citation:
Basso MF, Arraes FBM, Grossi-de-Sa M, Moreira VJV, Alves-Ferreira M and Grossi-de-Sa MF (2020) Insights Into Genetic and Molecular Elements for Transgenic Crop Development. Front. Plant Sci. 11:509. doi: 10.3389/fpls.2020.00509

BACKGROUND

Climate change, an increasing world population, and genetic erosion are the main factors indicating a need to improve crop adaptation, tolerance, and productivity. There is a continuing requirement for novel cultivars better adapted to different biomes with improved tolerance to biotic and abiotic stresses and superior yield and quality (Arzani and Ashraf, 2017). Conventional plant breeding, despite being a slow and usually difficult process, has made great contributions over the years. This method has been used mainly to add one simple trait to an otherwise ideal variety/cultivar. In contrast, genetic engineering has provided a complementary

CAPÍTULO 4

Modulação da expressão gênica em plantas via tecnologia CRISPR/dCas9

Carolina Vianna Morgante
Fabricio Barbosa Monteiro Arraes
Clidia Eduarda Moreira-Pinto
Bruno Paes de Melo
Maria Fatima Grossi-de-Sa

Introdução

A regulação da expressão gênica inclui uma diversidade de processos celulares que ocorrem de forma coordenada e em múltiplas etapas para desencadear o aumento ou a redução de um produto gênico específico. A expressão gênica pode ser induzida por estímulos endógenos e ambientais e modulada em diferentes níveis celulares, como na iniciação da transcrição, no processamento do RNA e na modificação pós-traducional da proteína.

A manipulação de genes-alvo é de primordial importância para o entendimento da função gênica e reprogramação das atividades celulares, tanto para o aprofundamento dos conhecimentos básicos sobre processos bioquímicos e moleculares, como para a intensificação de características de interesse agronômico. Neste ponto, a precisão é essencial para que se obtenha o êxito necessário em aplicações da engenharia genética e da biologia sintética.

Nas últimas décadas, tecnologias envolvendo nucleases sítio específicas para a manipulação precisa do DNA sofreram um profundo avanço, surgindo como alternativas promissoras para a indução de mutações sítio-dirigidas e controle fino da expressão gênica. Entre essas tecnologias, destacam-se as de edição de genomas, como a da nuclease dedo de zinco (ZFN, do inglês *Zinc Finger Nuclease*), a de nucleases com efetores do tipo ativador transcricional (TALENs, do inglês *Transcription Activator-Like Effector Nucleases*) e, mais recentemente, a tecnologia CRISPR/Cas (do inglês *Clustered Regularly Interspaced Short Palindromic Repeats*) associada à nuclease Cas. Esta última tem seu caráter revolucionário, sobretudo pela sua especificidade, universalidade e relativa simplicidade (Pickar-Oliver; Gersbach, 2019). Além disso, CRISPR/Cas é uma ferramenta flexível, passível de modificações, o que contribui para seu contínuo aprimoramento e diversifica suas aplicações nos estudos das funções celulares e na biotecnologia.

Freitas *et al. BMC Biotechnology* (2019) 19:79
https://doi.org/10.1186/s12896-019-0561-3

BMC Biotechnology

RESEARCH ARTICLE Open Access

Identification and characterization of the *GmRD26* soybean promoter in response to abiotic stresses: potential tool for biotechnological application

Elinea O. Freitas[1,2†], Bruno P. Melo[1,3†], Isabela T. Lourenço-Tessutti[1], Fabrício B. M. Arraes[1,4], Regina M. Amorim[1], Maria E. Lisei-de-Sá[1,5], Julia A. Costa[1,6], Ana G. B. Leite[1,2], Muhammad Faheem[1,7], Márcio A. Ferreira[8], Carolina V. Morgante[1,9], Elizabeth P. B. Fontes[3] and Maria F. Grossi-de-Sa[1,6*]

Abstract

Background: Drought is one of the most harmful abiotic stresses for plants, leading to reduced productivity of several economically important crops and, consequently, considerable losses in the agricultural sector. When plants are exposed to stressful conditions, such as drought and high salinity, they modulate the expression of genes that lead to developmental, biochemical, and physiological changes, which help to overcome the deleterious effects of adverse circumstances. Thus, the search for new specific gene promoter sequences has proved to be a powerful biotechnological strategy to control the expression of key genes involved in water deprivation or multiple stress responses.

Results: This study aimed to identify and characterize the *GmRD26* promoter (p*GmRD26*), which is involved in the regulation of plant responses to drought stress. The expression profile of the *GmRD26* gene was investigated by qRT-PCR under normal and stress conditions in Williams 82, BR16 and Embrapa48 soybean-cultivars. Our data confirm that *GmRD26* is induced under water deficit with different induction folds between analyzed cultivars, which display different genetic background and physiological behaviour under drought. The characterization of the *GmRD26* promoter was performed under simulated stress conditions with abscisic acid (ABA), polyethylene glycol (PEG) and drought (air dry) on *A. thaliana* plants containing the complete construct of p*GmRD26::GUS* (2.054 bp) and two promoter modules, p*GmRD26A::GUS* (909 pb) and p*GmRD26B::GUS* (435 bp), controlling the expression of the β-glucuronidase (*uidA*) gene. Analysis of GUS activity has demonstrated that p*GmRD26* and p*GmRD26A* induce strong reporter gene expression, as the p*AtRD29* positive control promoter under ABA and PEG treatment.

Conclusions: The full-length promoter p*GmRD26* and the p*GmRD26A* module provides an improved *uidA* transcription capacity when compared with the other promoter module, especially in response to polyethylene glycol and drought treatments. These data indicate that p*GmRD26A* may become a promising biotechnological asset with potential use in the development of modified drought-tolerant plants or other plants designed for stress responses.

Keywords: Stress-responsive promoter, Drought tolerance, Abscisic acid, Promoter modules analysis, Gene-promoter characterization

* Correspondence: mariafatimagrossidesa@gmail.com; fatima.grossi@embrapa.br
†Elinea O. Freitas and Bruno P. Melo contributed equally to this work.
[1]Embrapa Genetic Resources and Biotechnology, Brasilia, DF, Brazil
[6]Catholic University of Brasilia - Post-Graduation Program in Genomic Sciences and Biotechnology, Brasilia, DF, Brazil
Full list of author information is available at the end of the article

www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch

OPEN

Transcriptome and gene expression analysis of three developmental stages of the coffee berry borer, *Hypothenemus hampei*

Received: 14 September 2018
Accepted: 20 August 2019
Published online: 05 September 2019

Daniel D. Noriega[1,2], Paula L. Arias[3], Helena R. Barbosa[2,4], Fabricio B. M. Arraes[2,4], Gustavo A. Ossa[3], Bernardo Villegas[5], Roberta R. Coelho[2], Erika V. S. Albuquerque[2], Roberto C. Togawa[2], Priscila Grynberg[2], Haichuan Wang[6], Ana M. Vélez[6], Jorge W. Arboleda[7], Maria F. Grossi-de-Sa[2,8], Maria C. M. Silva[2] & Arnubio Valencia-Jiménez[5]

Coffee production is a global industry valued at approximately 173 billion US dollars. One of the main challenges facing coffee production is the management of the coffee berry borer (CBB), *Hypothenemus hampei*, which is considered the primary arthropod pest of coffee worldwide. Current control strategies are inefficient for CBB management. Although biotechnological alternatives, including RNA interference (RNAi), have been proposed in recent years to control insect pests, characterizing the genetics of the target pest is essential for the successful application of these emerging technologies. In this study, we employed RNA-seq to obtain the transcriptome of three developmental stages of the CBB (larva, female and male) to increase our understanding of the CBB life cycle in relation to molecular features. The CBB transcriptome was sequenced using Illumina Hiseq and assembled *de novo*. Differential gene expression analysis was performed across the developmental stages. The final assembly produced 29,434 unigenes, of which 4,664 transcripts were differentially expressed. Genes linked to crucial physiological functions, such as digestion and detoxification, were determined to be tightly regulated between the reproductive and nonreproductive stages of CBB. The data obtained in this study help to elucidate the critical roles that several genes play as regulatory elements in CBB development.

Coffee (*Coffea* spp.) is one of the most traded commodities in the world and is the second-most consumed beverage, with more than 9 million tons being consumed annually[1]. Currently cultivated in over 70 countries, the coffee industry is valued at approximately 173 billion US dollars, representing an important source of employment around the world[2]. Crop yield is reduced by more than 30 species of insect pests[3]. Among these pests, the coffee berry borer (CBB), *Hypothenemus hampei* (Ferrari, 1867) (Coleoptera: Curculionidae), is considered the most damaging pest for the coffee industry, causing annual global losses in excess of 500 million US dollars[4].

The CBB life cycle occurs within the coffee seed, with the adult female laying eggs in galleries formed throughout the endosperm. After hatching, the larvae feed on the coffee seed, reducing grain quality and increasing seed susceptibility to pathogen attack[5]. The endophytic behavior of CBB limits the use of traditional management methods due to poor cost-effectiveness, biosafety concerns and the development of insect resistance[4]. Biological

[1]Department of Cellular Biology, University of Brasília, Brasília-DF, Brazil. [2]Embrapa Genetic Resources and Biotechnology, Brasília-DF, Brazil. [3]Departamento de Ciencias Biológicas, Universidad de Caldas, Manizales, Colombia. [4]Biotechnology Center, UFRGS, Porto Alegre-RS, Brazil. [5]Departamento de Producción Agropecuaria, Universidad de Caldas, Manizales, Colombia. [6]University of Nebraska-Lincoln, Nebraska, United States of America. [7]Centro de Investigaciones en Medio Ambiente y Desarrollo – CIMAD, Universidad de Manizales, Manizales, Caldas, Colombia. [8]Catholic University of Brasília - Postgraduate Program in Genomic Sciences and Biotechnology, Brasília-DF, Brazil. Daniel D. Noriega, Paula L. Arias and Helena R. Barbosa contributed equally. Maria C. M. Silva and Arnubio Valencia-Jiménez jointly supervised this work. Correspondence and requests for materials should be addressed to D.D.N. (email: daniel.nv07@gmail.com) or M.F.G.-d.-S. (email: fatima.grossi@embrapa.br) or A.V.-J. (email: arnubio.valencia@ucaldas.edu.co)

Plant Science 270 (2018) 72–84

Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

Review article

Review: Potential biotechnological assets related to plant immunity modulation applicable in engineering disease-resistant crops

Marilia Santos Silva[a,**,1], Fabrício Barbosa Monteiro Arraes[a,b,***,1], Magnólia de Araújo Campos[c], Maira Grossi-de-Sa[d], Diana Fernandez[d], Elizabete de Souza Cândido[e,f], Marlon Henrique Cardoso[e,f,g], Octávio Luiz Franco[e,f,g], Maria Fátima Grossi-de-Sa[a,b,e,g,*]

[a] Embrapa Recursos Genéticos e Biotecnologia (Embrapa Cenargen), Brasília, DF, Brazil
[b] Universidade Federal do Rio Grande do Sul (UFRGS), Post-Graduation Program in Molecular and Cellular Biology, Porto Alegre, RS, Brazil
[c] Universidade Federal de Campina Grande (UFCG), Center of Education and Health Cuité-PB, Brazil
[d] IRD, CIRAD, Univ. Montpellier, IPME, Montpellier, France
[e] Universidade Católica de Brasília (UCB), Post-Graduation Program in Genomic Science and Biotechnology, Brasília, DF, Brazil
[f] Universidade Católica Dom Bosco (UCDB), Campo Grande, MS, Brazil
[g] Universidade de Brasília (UnB), Brasília, DF, Brazil

ARTICLE INFO

Keywords:
Pathogen-associated molecular pattern-triggered immunity (PTI)
Effector-triggered susceptibility (ETS)
Effector-triggered immunity (ETI)
RNA interference
Genetically modified (GM) crop
Disease resistance

ABSTRACT

This review emphasizes the biotechnological potential of molecules implicated in the different layers of plant immunity, including, pathogen-associated molecular pattern (PAMP)-triggered immunity (PTI), effector-triggered susceptibility (ETS), and effector-triggered immunity (ETI) that can be applied in the development of disease-resistant genetically modified (GM) plants. These biomolecules are produced by pathogens (viruses, bacteria, fungi, oomycetes) or plants during their mutual interactions. Biomolecules involved in the first layers of plant immunity, PTI and ETS, include inhibitors of pathogen cell-wall-degrading enzymes (CWDEs), plant pattern recognition receptors (PRRs) and susceptibility (S) proteins, while the ETI-related biomolecules include plant resistance (R) proteins. The biomolecules involved in plant defense PTI/ETI responses described herein also include antimicrobial peptides (AMPs), pathogenesis-related (PR) proteins and ribosome-inhibiting proteins (RIPs), as well as enzymes involved in plant defensive secondary metabolite biosynthesis (phytoanticipins and phytoalexins). Moreover, the regulation of immunity by RNA interference (RNAi) in GM disease-resistant plants is also considered. Therefore, the present review does not cover all the classes of biomolecules involved in plant innate immunity that may be applied in the development of disease-resistant GM crops but instead highlights the most common strategies in the literature, as well as their advantages and disadvantages.

1. Introduction

Plant pathogens, including viruses, bacteria, fungi, and oomycetes are a primary concern in agribusiness [1–3]. The diseases caused by these organisms in plants represent an important and persistent threat to food supplies worldwide [4]. The development of disease-resistant plants through biotechnological approaches aims to obtain economically important crops through elite genetically modified (GM) lines that not only display durable and broad-spectrum resistance to multiple phytopathogens, but that are also biosafe to the environment and consumers. To achieve this goal, several challenges related to transgene must be overcome, such as fine-tuning the choice, origin (i.e., heterologous species and/or non-host plant) and the number of genes to be employed and stacked, as well as gene expression control (e.g., by signal peptides, gene silencing and gene promoters). The current knowledge of the molecular mechanisms involved in plant-pathogen interactions has now provided a large set of biomolecules that can be applied in the development of GM disease-resistant/less susceptible crops.

Plant-pathogen interactions involve a two-way communication process, whereby plants can recognize and induce defense strategies against pathogens, while pathogens can threaten plant functional

[*] Corresponding authors at: Embrapa Recursos Genéticos e Biotecnologia (Cenargen), Brasília, DF, Brazil.
[**] Corresponding author.
[***] Corresponding author at: Universidade Federal do Rio Grande do Sul (UFRGS), Post-Graduation Program in Molecular and Cellular Biology, Porto Alegre, RS, Brazil.
E-mail addresses: marilia.silva@embrapa.br (M.S. Silva), fabricio.arraes@gmail.com (F.B.M. Arraes), fatima.grossi@embrapa.br (M.F. Grossi-de-Sa).
[1] These authors contributed equally.

https://doi.org/10.1016/j.plantsci.2018.02.013
Received 9 October 2017; Received in revised form 1 February 2018; Accepted 4 February 2018
Available online 14 February 2018

Plant Biotechnology Journal

Plant Biotechnology Journal (2017), pp. 1–13 doi: 10.1111/pbi.12694

Transgenic cotton expressing Cry10Aa toxin confers high resistance to the cotton boll weevil

Thuanne Pires Ribeiro[1,2], Fabricio Barbosa Monteiro Arraes[2,3], Isabela Tristan Lourenço-Tessutti[2], Marilia Santos Silva[2], Maria Eugênia Lisei-de-Sá[2,4], Wagner Alexandre Lucena[2,5], Leonardo Lima Pepino Macedo[2], Janaina Nascimento Lima[2], Regina Maria Santos Amorim[2], Sinara Artico[6], Márcio Alves-Ferreira[6], Maria Cristina Mattar Silva[2] and Maria Fatima Grossi-de-Sa[2,7,]*

[1]*Brasilia Federal University (UnB), Brasília, DF, Brazil*
[2]*Embrapa Genetic Resources and Biotechnology, Brasília, DF, Brazil*
[3]*Rio Grande do Sul Federal University, Porto Alegre, RS, Brazil*
[4]*Agricultural Research Company of Minas Gerais State, Uberaba, MG, Brazil*
[5]*Embrapa Cotton, Campina Grande, PB, Brazil*
[6]*Rio de Janeiro Federal University, Rio de Janeiro, RJ, Brazil*
[7]*Catholic University of Brasília, Brasília, DF, Brazil*

Received 22 October 2016;
revised 5 January 2017;
accepted 8 January 2017.

**Correspondence* (Tel +55 61 3448 4705; fax +55 61 3440 3624; email fatima.grossi@embrapa.br)

Accession numbers (GenBank): *AtRNAH*: At1g58050 (TAIR); *Atpp2a*: At1g13320 (TAIR); Cry1Ac: P05068; Cry2Aa: P0A377; *cry10Aa*: AAA22614.1; *Ghubq14*: DW505546; *Ghpp2a1*: DT545658; *uceA1.7*: JN887312.

Keywords: *Gossypium hirsutum*, genetic transformation, *Anthonomus grandis*, insect pest, plant resistance, *Bt* toxin.

Summary

Genetically modified (GM) cotton plants that effectively control cotton boll weevil (CBW), which is the most destructive cotton insect pest in South America, are reported here for the first time. This work presents the successful development of a new GM cotton with high resistance to CBW conferred by Cry10Aa toxin, a protein encoded by entomopathogenic *Bacillus thuringiensis* (*Bt*) gene. The plant transformation vector harbouring *cry10Aa* gene driven by the cotton ubiquitination-related promoter *uce*A1.7 was introduced into a Brazilian cotton cultivar by biolistic transformation. Quantitative PCR (qPCR) assays revealed high transcription levels of *cry10Aa* in both T_0 GM cotton leaf and flower bud tissues. Southern blot and qPCR-based $2^{-\Delta\Delta Ct}$ analyses revealed that T_0 GM plants had either one or two transgene copies. Quantitative and qualitative analyses of Cry10Aa protein expression showed variable protein expression levels in both flower buds and leaves tissues of T_0 GM cotton plants, ranging from approximately 3.0 to 14.0 µg g^{-1} fresh tissue. CBW susceptibility bioassays, performed by feeding adults and larvae with T_0 GM cotton leaves and flower buds, respectively, demonstrated a significant entomotoxic effect and a high level of CBW mortality (up to 100%). Molecular analysis revealed that transgene stability and entomotoxic effect to CBW were maintained in T_1 generation as the Cry10Aa toxin expression levels remained high in both tissues, ranging from 4.05 to 19.57 µg g^{-1} fresh tissue, and the CBW mortality rate remained around 100%. In conclusion, these Cry10Aa GM cotton plants represent a great advance in the control of the devastating CBW insect pest and can substantially impact cotton agribusiness.

Introduction

Cotton (*Gossypium hirsutum*) production is highly influenced by a large number of insect pests, and a major pest in the Americas is the cotton boll weevil (CBW) *Anthonomus grandis* (Coleoptera: Curculionidae), which causes significant losses to cotton production and impacts fiber quality (Azambuja and Degrande, 2014; Bastos *et al.*, 2005; De Lima *et al.*, 2013; Gallo *et al.*, 2002; Habib and Fernandes, 1983; Instituto Mato-grossense do Algodão, 2015; Ribeiro *et al.*, 2010; Soria *et al.*, 2013). The endophytic habit of CBW larvae into cotton reproductive structures can result in crop losses of up to 100%, especially because chemical control is only applicable during the adult weevil stage, when it feeds on immature cotton bolls (Busoli and Michelotto, 2005; Ribeiro *et al.*, 2015).

Bacillus thuringiensis (*Bt*) has contributed to insect pest control since the 1960s (Lacey *et al.*, 2001). There are currently more than 750 characterized *Bt*-encoded entomotoxic crystal proteins (Cry), which are grouped into at least 74 different classes and are collectively active against insects, nematodes, mites and protozoans (Crickmore *et al.*, 1998). Although substantial knowledge on the direct use of *Bt* for the biological control of insects has accumulated over the last decades, its commercial application is limited due to high production costs and instability of the Cry proteins under field conditions (Navon, 2000, 2013).

The broad adoption of genetically modified (GM) *Bt* cotton technology by the world's largest cotton producers, such as China, India and Brazil, has notably brought great economic benefits to producers (James, 2015). Concerning hemipteran insect control, a recent study showed that a *Bt* toxin variant (Cry51Aa2.834_16) could reduce populations of *Lygus* spp. in whole-GM cotton plants evaluated in caged-field trials (Gowda *et al.*, 2016). Although various characterized Cry toxins are active

Please cite this article as: Ribeiro, T.P., Arraes, F.B.M., Lourenço-Tessutti, I.T., Silva, M.S., Lisei-de-Sá, M.E., Lucena, W.A., Macedo, L.L.P., Lima, J.N., Santos Amorim, R.M., Artico, S., Alves-Ferreira, M., Mattar Silva, M.C. and Grossi-de-Sa, M.F. (2017) Transgenic cotton expressing Cry10Aa toxin confers high resistance to the cotton boll weevil. *Plant Biotechnol. J.*, doi: 10.1111/pbi.12694.

Scientific Research Publishing

Agricultural Sciences, 2017, 8, 1195-1225
http://www.scirp.org/journal/as
ISSN Online: 2156-8561
ISSN Print: 2156-8553

AtDREB2A-CA Influences Root Architecture and Increases Drought Tolerance in Transgenic Cotton

Maria Eugênia Lisei-de-Sá[1,2*], Fabricio B. M. Arraes[1,3*], Giovani G. Brito[4], Magda A. Beneventi[1,3], Isabela T. Lourenço-Tessutti[1], Angelina M. M. Basso[1,5], Regina M. S. Amorim[1], Maria C. M. Silva[1], Muhammad Faheem[1], Nelson G. Oliveira[6], Junya Mizoi[7], Kazuko Yamaguchi-Shinozaki[7], Maria Fatima Grossi-de-Sa[1,8#]

[1]Embrapa Recursos Genéticos e Biotecnologia, Brasília, Brazil
[2]Empresa de Pesquisa Agropecuária de Minas Gerais, Uberaba, Brazil
[3]Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
[4]Embrapa Clima Temperado, Pelotas, Brazil
[5]Universidade de Brasília, Brasília, Brazil
[6]Embrapa Agroenergia, Brasília, Brazil
[7]Japan International Research Center for Agricultural Sciences (JIRCAS), Tsukuba, Japan
[8]Universidade Católica de Brasília, Brasília, Brazil
Email: maria-eugenia.sa@colaborador.embrapa.br, fabricio.arraes@gmail.com, giovani.brito@embrapa.br, mabeneventi@gmail.com, isabela.lourenço@embrapa.br, angelina_granger@yahoo.com.br, rmsamorim@yahoo.com.br, mattar@embrapa.br, faheem08@live.com, dr.nelson74@hotmail.com, ajmizoi@mail.ecc.u-tokyo.ac.jp, akys@mail.ecc.u-tokyo.ac.jp, #fatima.grossi@embrapa.br

How to cite this paper: Lisei-de-Sá, M.E., Arraes, F.B.M., Brito, G.G., Beneventi, M.A., Lourenço-Tessutti, I.T., Basso, A.M.M., Amorim, R.M.S., Silva, M.C.M., Faheem, M., Oliveira, N.G., Mizoi, J., Yamaguchi-Shinozaki, K. and Grossi-de-Sa, M.F. (2017) *AtDREB2A-CA* Influences Root Architecture and Increases Drought Tolerance in Transgenic Cotton. *Agricultural Sciences*, **8**, 1195-1225.
https://doi.org/10.4236/as.2017.810087

Received: September 18, 2017
Accepted: October 28, 2017
Published: October 31, 2017

Abstract

Drought is a major environmental factor limiting cotton (*Gossypium hirsutum* L.) productivity worldwide and projected climate changes could increase their negative effects in the future. Thus, targeting the molecular mechanisms correlated with drought tolerance without reducing productivity is a challenge for plant breeding. In this way, we evaluated the effects of water deficit progress on *AtDREB2A-CA* transgenic cotton plant responses, driven by the stress-inducible *rd*29 promoter. Besides shoot and root morphometric traits, gas exchange and osmotic adjustment analyses were also included. Here, we present how altered root traits shown by transgenic plants impacted on physiological acclimation responses when submitted to severe water stress. The integration of *AtDREB2A-CA* into the cotton genome increased total root volume, surface area and total root length, without negatively affecting shoot morphometric growth parameters and nor phenotypic evaluated traits. Additionally, when compared to wild-type plants, transgenic plants (17-T_0 plants and its progeny) highlighted a gradual pattern of phenotypic plasticity to

*These authors have contributed equally to this work.

frontiers
in Plant Science

ORIGINAL RESEARCH
published: 19 February 2016
doi: 10.3389/fpls.2016.00165

Transgenic Cotton Plants Expressing Cry1Ia12 Toxin Confer Resistance to Fall Armyworm (*Spodoptera frugiperda*) and Cotton Boll Weevil (*Anthonomus grandis*)

Raquel S. de Oliveira[1,2], Osmundo B. Oliveira-Neto[2,3], Hudson F. N. Moura[2,4], Leonardo L. P. de Macedo[2], Fabrício B. M. Arraes[2,5], Wagner A. Lucena[2,6], Isabela T. Lourenço-Tessutti[2], Aulus A. de Deus Barbosa[2], Maria C. M. da Silva[2] and Maria F. Grossi-de-Sa[1,2]*

[1] Catholic University of Brasilia, Brasilia, Brazil, [2] Pest-Plant Molecular Interaction Laboratory, Embrapa Genetic Resources and Biotechnology, Brazilian Research Agricultural Corporation, Brasilia, Brazil, [3] UNIEURO – University Center, Brasilia, Brazil, [4] Biology Institute, Brasilia University, Brasilia, Brazil, [5] Federal University of Rio Grande do Sul, Porto Alegre, Brazil, [6] Embrapa Cotton, Campina Grande, Brazil

Gossypium hirsutum (commercial cooton) is one of the most economically important fibers sources and a commodity crop highly affected by insect pests and pathogens. Several transgenic approaches have been developed to improve cotton resistance to insect pests, through the transgenic expression of different factors, including Cry toxins, proteinase inhibitors, and toxic peptides, among others. In the present study, we developed transgenic cotton plants by fertilized floral buds injection (through the pollen-tube pathway technique) using an DNA expression cassette harboring the *cry1Ia12* gene, driven by CaMV35S promoter. The T0 transgenic cotton plants were initially selected with kanamycin and posteriorly characterized by PCR and Southern blot experiments to confirm the genetic transformation. Western blot and ELISA assays indicated the transgenic cotton plants with higher Cry1Ia12 protein expression levels to be further tested in the control of two major *G. hirsutum* insect pests. Bioassays with T1 plants revealed the Cry1Ia12 protein toxicity on *Spodoptera frugiperda* larvae, as evidenced by mortality up to 40% and a significant delay in the development of the target insects compared to untransformed controls (up to 30-fold). Also, an important reduction of *Anthonomus grandis* emerging adults (up to 60%) was observed when the insect larvae were fed on T1 floral buds. All the larvae and adult insect survivors on the transgenic lines were weaker and significantly smaller compared to the non-transformed plants. Therefore, this study provides GM cotton plant with simultaneous resistance against the Lepidopteran (*S. frugiperda*), and the Coleopteran (*A. grandis*) insect orders, and all data suggested that the Cry1Ia12 toxin could effectively enhance the cotton transgenic plants resistance to both insect pests.

Keywords: *Gossypium hirsutum*, genetic cotton transformation, pollen-tube pathway, Cry1Ia12, *Anthonomus grandis*, *Spodoptera frugiperda*

OPEN ACCESS

Edited by:
Scott Nicholas Johnson,
University of Western Sydney,
Australia

Reviewed by:
Octavio Luiz Franco,
Universidade Catolica de Brasilia,
Brazil
Lewis Joel Wilson,
Commonwealth Scientific
and Industrial Research
Organisation – Agriculture, Australia
Katia Scortecci,
Universidade Federal do Rio Grande
do Norte, Brazil

***Correspondence:**
Maria F. Grossi-de-Sa
fatima.grossi@embrapa.br

Specialty section:
This article was submitted to
Crop Science and Horticulture,
a section of the journal
Frontiers in Plant Science

Received: 17 September 2015
Accepted: 31 January 2016
Published: 19 February 2016

Citation:
de Oliveira RS, Oliveira-Neto OB, Moura HFN, de Macedo LLP, Arraes FBM, Lucena WA, Lourenço-Tessutti IT, de Deus Barbosa AA, da Silva MCM and Grossi-de-Sa MF (2016) Transgenic Cotton Plants Expressing Cry1Ia12 Toxin Confer Resistance to Fall Armyworm (Spodoptera frugiperda) and Cotton Boll Weevil (Anthonomus grandis). Front. Plant Sci. 7:165. doi: 10.3389/fpls.2016.00165

In: Jasmonic Acid
Editor: Lucille Morrison

ISBN: 978-1-63463-854-8
© 2015 Nova Science Publishers, Inc.

Chapter 1

JASMONIC ACID: MEDIATED PLANT DEFENSE

***Maria Eugênia Lisei de Sá*[1,2], *Magda Aparecida Beneventi*[1], *Fabrício Barbosa Monteiro Arraes*[2], *Regina Santos de Amorim*[2] *and Maria Fatima Grossi–de–Sa*[2,3]**

[1]Agricultural Research Company of Minas Gerais State, Uberaba, Brazil
[2]Embrapa Genetic Resources and Biotechnology, Brasilia, Brazil
[3]Catholic University of Brasilia, Brasilia, DF, Brazil

ABSTRACT

Compatible plant-pathogen interaction results in various hormonal level changes, triggering a vast array of genes to defend from the invading pathogen. Along with the phytohormones salicylic acid (SA) and ethylene (ET), jasmonic acid (JA) regulate both basal and resistance (*R*) gene-mediated defense responses. There are many evidences that JA and ET signaling work synergistically whereas SA and JA/ET signaling is by an antagonistic cross-talk. Contrarily, it has been shown that both salicylic acid (SA) derivative methyl salicylate (MeSA) and methyl jasmonate (MeJA) are essential for systemic resistance against Tobacco mosaic virus (TMV), possibly acting as the initiating signals for systemic resistance. Silencing of SA or JA biosynthetic and signaling genes in

APPENDIX II

CURRICULUM VITAE

2020

DADOS PESSOAIS

Nome: Fabrício Barbosa Monteiro Arraes

Endereço Profissional: Embrapa Recursos Genéticos e Biotecnologia, Laboratório de Interação Molecular Planta Praga – Parque Estação Biológica – PqEB – Av. W5 Norte
Asa Norte – Brasília-DF – 70.770-917, DF - Brasil
Telefone: 61 3448-4705

E-mail: fabricio.arraes@gmail.com

FORMAÇÃO ACADÊMICA/TITULAÇÃO

2012 - 2014 Mestrado em Biologia Celular e Molecular.
Universidade Federal do Rio Grande do Sul, UFRGS, Porto Alegre, Brasil.
Título: **Biossíntese e Sinalização de Etileno em Soja: Uso de Abordagens *In Silico* e Implicação na Resposta a Seca**, Ano de obtenção: 2014.
Orientadora: Dra. Maria Fátima Grossi-de-Sá.
Bolsista da Fundação De Apoio à Pesquisa do Distrito Federal (FAP-DF).

2001 - 2005 Graduação em Ciências Biológicas.
Universidade de Brasília, UnB, Brasília, Brasil
Título: **Identificação, Isolamento e Caracterização do Gene do Repressor Transcricional Pbtup1 do Fungo Dimórfico e Patogênico *Paracoccidioides brasiliensis***, Ano de obtenção: 2005.
Orientadora: Dra. Maria Sueli Soares Felipe.

FORMAÇÃO COMPLEMENTAR

2010 - 2010 Curso de curta duração em Enfermedades de Almacenamiento Lisosomal. (Carga horária: 40h).
Universidad Nacional de La Plata, UNLP, Argentina.

2007 - 2007 Citometria de Fluxo. . (Carga horária: 24h).
Universidade de Brasília, UnB, Brasília, Brasil.

2005 - 2005 Extensão universitária em Técnicas para Análise de Polimorfismo Genético. (Carga horária: 45h).
Universidade de São Paulo, USP, São Paulo, Brasil.

2004 - 2004 Extensão universitária em I Curso de Verão em Biologia Celular e Molecular. (Carga horária: 84h).
Universidade de São Paulo, USP, São Paulo, Brasil.

2002 - 2002 Curso de curta duração em Marcadores Moleculares. (Carga horária: 3h).
Sociedade Brasileira de Micologia, SBM, Rio De Janeiro, Brasil.

ATUAÇÃO PROFISSIONAL

2015 – 2020 EMBRAPA RECURSOS GENÉTICOS E BIOTECNOLOGIA

Vínculo: Bolsista.
Enquadramento funcional: Bolsista de Doutorado.
Carga horária: 40 horas.
Regime: Dedicação exclusiva.
Atividades: 02/2015 - 01/2020 Pesquisa e Desenvolvimento, Laboratório de Interação Molecular Planta Praga.

2018 – 2019 INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE (INRA)

Vínculo: Bolsista.
Enquadramento funcional: Estagiário.
Carga horária: 44 horas.
Regime: Integral.
Atividades: 02/2018 - 02/2019 Pesquisa na área de genômica estrutural e funcional de insetos, Institut National de la Recherche Agronomique (INRA).

2012 – 2014 EMBRAPA RECURSOS GENÉTICOS E BIOTECNOLOGIA

Vínculo: Bolsista.
Enquadramento funcional: Bolsista de Mestrado.
Carga horária: 40 horas.
Regime: Dedicação exclusiva.
Atividades: 06/2012 - 12/204 Pesquisa e Desenvolvimento, Laboratório de Interação Molecular Planta Praga.

2011 – 2012 EMBRAPA RECURSOS GENÉTICOS E BIOTECNOLOGIA

Vínculo: Bolsista.
Enquadramento funcional: Estagiário.
Carga horária: 30 horas.
Regime: Parcial.
Atividades: 01/2011 - 05/2012 Pesquisa e Desenvolvimento, Laboratório de Interação Molecular Planta Praga.

2009 – 2010 TECNOGENE DIAGNÓSTICOS MOLECULARES

Vínculo: Celetista.
Enquadramento funcional: Biólogo.
Carga horária: 44 horas.
Regime: Integral.
Atividades: 01/2009 - 12/2010 Serviço Técnico Especializado, Tecnogene Diagnósticos Moleculares LTDA.

2007 – 2008 UNIVERSIDADE DE BRASÍLIA – UnB

Vínculo: Bolsista.
Enquadramento funcional: Bolsista/Estagiário Graduado – UnB.
Carga horária: 30 horas.
Regime: Parcial.
Atividades: 01/2008 - 12/2009 Pesquisa e Desenvolvimento, UnB - Laboratório de Virologia Molecular.

2006 – 2006 TECNOGENE DIAGNÓSTICOS MOLECULARES
Vínculo: Estagiário.
Enquadramento funcional: Estagiário Remunerado.
Carga horária: 40 horas.
Regime: Integral.
Atividades: 01/2006 - 12/2006 Estágio, Tecnogene Diagnósticos Moleculares LTDA.

2003 – 2005 UNIVERSIDADE DE BRASÍLIA – UnB
Vínculo: Bolsista.
Enquadramento funcional: Bolsista de Iniciação Tecnológica I.
Carga horária: 30 horas.
Regime: Parcial.
Atividades: 04/2003 - 02/2005 Pesquisa e Desenvolvimento, UnB - Laboratório de Biologia Molecular

---

PRODUÇÃO BIBLIOGRÁFICA

1. ARTIGOS COMPLETOS PUBLICADOS EM PERIÓDICOS

ANO 2020:

ARRAES, FBM; Martins-de-Sa, D; Noriega-Vasquez, DD; Melo, BP; Faheem, M; Macedo, LLP; Morgante, CV; Barbosa, JA; Togawa, RC; Moreira, VJV; Danchin, EGJ; Grossi-de-Sa, MF (2020). **Dissecting protein domain variability in the core RNA interference machinery of five insect orders.** *RNA Biology*.

BASSO, MF; **Arraes, FBM;** Grossi-de-Sa, M; Moreira, VJV; Alves-Ferreira, M; Grossi-de-Sa, MF (2020). **Insights into genetic and molecular elements for transgenic crop development.** *Frontiers in Plant Science*.

FIRMINO, AAP; Pinheiro, DH; Moreira-Pinto, CE; Souza-Jr., JDA; Macedo, LLP; Martins-de-Sa, D; **Arraes, FBM;** Coelho, RR; Fonseca, FCA; Silva, MCM; Engler, JA; Silva, MS; Lourenço-Tessutti, IT; Terra, WR; Grossi-de-Sa, MF (2020). **RNAi-mediated suppression of *Laccase 2* impairs cuticle tanning and molting in the cotton boll weevil (*Anthonomus grandis*).** *Frontiers in Physiology*.

MOTA, APZ; Fernandez, D; **Arraes, FBM;** Petitot, AS; Melo, BP; Lisei-de-Sa, ME; Grynberg, P; Saraiva, MAP; Guimaraes, PM; Brasileiro, ACM; Albuquerque, EVS; Danchin, EGJ; Grossi-de-Sa, MF (2020). **Evolutionarily conserved plant genes responsive to root-knot nematodes identified by comparative genomics.** *Molecular Genetics and Genomics*.

NORIEGA-VASQUEZ, DD; **Arraes, FBM;** Souza-Jr., JDA; Macedo, LLP; Fonseca, FCA; Togawa, RC; Grynberg, P; Silva, MCM; Negrisoli, AS; Morgante, CV; Grossi-de-Sa, MF (2020). **Comparative gut transcriptome analysis of *Diatraea saccharalis* in response to the dietary source.** *PLoS One*.

NORIEGA-VASQUEZ, DD; **Arraes, FBM;** Souza-Jr., JDA; Macedo, LLP; Fonseca, FCA; Togawa, RC; Grynberg, P; Silva, MCM; Negrisoli, AS; Grossi-de-Sa, MF (2020). **Transcriptome analysis and knockdown of the juvenile hormone esterase gene reveal abnormal feeding behavior in the sugarcane giant borer.** *Frontiers in Physiology*.

ANO 2019:

FREITAS, EO; Melo, BP; Lourenço-Tessutti, IT; **Arraes, FBM;** Amorim, RMS; Lisei-de-Sa, ME; Costa, JA; Leite, AGB; Faheem, M; Alves-Ferreira, M; Morgante, CV; Fontes, EPB; Grossi-de-Sa, MF (2019). **Identification and characterization of the *GmRD26* soybean promoter in response to abiotic stresses: Potential tool for biotechnological application.** *BMC Biotechnology*.

NORIEGA-VASQUEZ, DD; Arias, PL; Barbosa, HR; **Arraes, FBM;** Ossa, GA; Villegas, B; Coelho, RR; Albuquerque, EVS; Togawa, RC; Grynberg, P; Wang, H; Vélez, AM.; Arboleda, JW; Grossi-de-Sa, MF; Silva, MCM; Valencia-Jiménez, A (2019). **Transcriptome and gene expression analysis of three developmental stages of the coffee berry borer, *Hypothenemus hampei*.** *Scientific Reports*.

ANO 2018:

SILVA, MS; **Arraes, FBM;** Campos, MA; Grossi-de-Sa, M; Fernandez, D; Cândido, ES; Cardoso, MH; Franco, OL; Grossi-de-Sa, MF (2018). **Review: Potential biotechnological assets related to plant immunity modulation applicable in engineering disease-resistant crops.** *Plant Science*.

ANO 2017:

LISEI-DE-SA, ME; **Arraes, FBM;** Brito, GG; Beneventi, MA; Lourenço-Tessutti, IT; Basso, AMM; Amorim, RMS; Silva, MCM; Faheem, M; Oliveira, NG; Mizoi, J; Yamaguchi-Shinozaki, K; Grossi-de-Sa, MF (2017). ***AtDREB2A-CA* influences root architecture and increases drought tolerance in transgenic cotton.** *Agricultural Sciences*.

RIBEIRO, TP; **Arraes, FBM;** Lourenço-Tessutti, IT; Silva, MS; Lisei-de-Sa, ME; Lucena, WA; Macedo, LLP; Lima, JN; Amorim, RMS; Artico, S; Alves-Ferreira, M; Silva, MCM; Grossi-de-Sa, MF (2017). **Transgenic cotton expressing Cry10Aa toxin confers high resistance to the cotton boll weevil.** *Plant Biotechnology Journal*.

ANO 2016:

OLIVEIRA, RS; Oliveira-Neto, OB; Moura, HFN; Macedo, LLP; **Arraes, FBM;** Lucena, WA; Lourenço-Tessutti, IT; Barbosa, AAD; Silva, MCM; Grossi-de-Sa, MF (2016). **Transgenic cotton plants expressing Cry1Ia12 toxin confer resistance to fall armyworm (*Spodoptera frugiperda*) and cotton boll weevil (*Anthonomus grandis*).** *Frontiers in Plant Science*.

ANO 2015:

ARRAES, FBM; Beneventi, MA; Lisei-de-Sa, ME; Paixao, JFR; Albuquerque, EVS; Marin, SRR; Purgatto, E; Nepomuceno, AL; Grossi-de-Sa, MF (2015). **Implications of ethylene biosynthesis and signaling in soybean drought stress tolerance.** *BMC Plant Biology*.

ANO 2009:

NEIVA, M; **Arraes, FBM;** Souza, JV; Rádis-Baptista, G; Silva, ARBP; Walter, MEMT; Brígido, MM; Yamane, T; López-Lozano, JL; Astolfi-Filho, S (2009). **Transcriptome**

analysis of the Amazonian viper *Bothrops atrox* venom gland using expressed sequence tags (ESTs). *Toxicon*.

ANO 2008:

NICOLA, AM; Andrade, RV; Dantas, AS; Andrade, PA; **Arraes, FBM;** Fernandes, L; Silva-Pereira, I; Felipe, MSS (2008). **The stress responsive and morphologically regulated *hsp90* gene from *Paracoccidioides brasiliensis* is essential to cell viability.** *BMC Microbiology*.

ANO 2007:

ARRAES, FBM; Carvalho, MJA; Maranhão, AQ; Brígido, MM; Pedrosa, FO; Felipe, MSS (2007). Differential metabolism of Mycoplasma species as revealed by their genomes. *Genetics and Molecular Biology*.

ANO 2005:

ALBUQUERQUE, PA; Baptista, AJ; Derengowsky, LS; Procopio-Silva, L; Nicola, AM; **Arraes, FBM;** Souza, DP; Kyaw, CM; Silva-Pereira, I (2005). ***Paracoccidioides brasiliensis* RNA biogenesis apparatus revealed by functional genome analysis.** *Genetics and Molecular Research*.

ARRAES, FBM; Benoliel, B; Burtet, RT; Costa, PL; Galdino, AS; Lima, LH; Marinho-Silva, C; Oliveira-Pereira, L; Pfrimer, P; Procópio-Silva, L; Reis, VCB; Felipe, MSS (2005). **General metabolism of the dimorphic and pathogenic fungus *Paracoccidioides brasiliensis*.** *Genetics and Molecular Research*.

BENOLIEL, B; **Arraes, FBM;** Reis, VCB; Siqueira, SJ; Parachin, NS; Torres, FAG (2005). **Hydrolytic enzymes in *Paracoccidioides brasiliensi*s - Ecological aspects.** *Genetics and Molecular Research*.

FELIPE, MSS; **Arraes, FBM;** Torres, FAG; Maranhão, AQ; Poças-Fonseca, MJ; Campos, EG; Moraes, LMP; Carvalho, MJA; Andrade, RV; Jesuíno, RSA; Pereira, M; Soares, CMA; Brígido, MM (2005). **Functional genome of the human pathogenic fungus *Paracoccidioides brasiliensis*.** *FEMS Immunology and Medical Microbiology*.

FELIPE, MSS; **Arraes, FBM;** Andrade, RV; Nicola, AM; Maranhão, AQ; Torres, FAG; Silva-Pereira, I; Poças-Fonseca, MJ; Campos, EG; Moraes, LMP; Albuquerque, PA; Tavares, AHFP; Silva, SS; Kyaw, CM; Souza, DP; Pereira, M; Jesuíno, RSA; Andrade, EV; Parente, JA; Oliveira, EGS; Barbosa, MS; Martins, NF; Fachin, AL; Cardoso, RS; Passos, GAS; Almeida, NF; Walter, MEMT; Soares, CMA; Carvalho, MJA; Brígido, MM (2005). **Transcriptional profiles of the human pathogenic fungus *Paracoccidioides brasiliensis* in mycelium and yeast cells.** *Journal of Biological Chemistry*.

ANO 2003:

FELIPE, MSS; Andrade, RV; Petrofeza, SS; Maranhão, AQ; Torres, FAG; Albuquerque, PA; **Arraes, FBM;** Arruda, M; Azevedo, MO; Baptista, AJ; Bataus, LAM; Borges, CL; Campos, EG; Cruz, MR; Daher, BS; Dantas, A; Ferreira, MASV; Ghil, GV; Jesuíno, RSA; Kyaw, CM; Leitão, L; Martins, CRF; Moraes, LMP; Neves, EO; Nicola, AM; Alves, ES;

Parente, JA; Pereira, M; Poças-Fonseca, MJ; Resende, R; Ribeiro, BM; Saldanha, RR; Santos, SC; Silva-Pereira, I; Silva, MAS; Silveira, E; Simões, IC; Soares, RBA; Souza, DP; Souza, MT; Andrade, EV; Xavier, MAS; Veiga, HP; Venâncio, EJ; Carvalho, MJA; Oliveira, AG; Inoue, MK; Almeida, NF; Walter, MEMT; Soares, CMA; Brígido, MM (2003). **Transcriptome characterization of the dimorphic and pathogenic fungus *Paracoccidioides brasiliensis* by EST analysis.** *Yeast*.

2. CAPÍTULOS DE LIVROS PUBLICADOS

MORGANTE, CV; **Arraes, FBM;** Moreira-Pinto, CE; Melo, BP; Grossi-de-Sá, MF (2020). **Modulação da expressão gênica em plantas via tecnologia CRISPR/dCas9.** *In:* Tecnologia CRISPR na edição genômica de plantas: biotecnologia aplicada à agricultura.1ª ed., Embrapa, v. 1, p. 125-177.

GROSSI-DE-SÁ, MF; **Arraes, FBM;** Guimarães, PM; Pelegrini, PB (2016). **Biotechnology and GM Crops in Brazil.** *In:* Innovative farming and forestry across the emerging world: the role of genetically modified crops and trees. 1ª ed., International Industrial Biotechnology Network, v. 1, p. 1-128.

LISEI-DE-SA, MEL; Beneventi, MA; **Arraes, FBM;** Amorim, RMS; Grossi-de-Sá, M F (2015). **Jasmonic acid: Mediated plant defense.** *In:* Jasmonic acid–mediated plant Defense. 1ª ed., New York: Nova Publishers, v. 1, p. 1-14.

3. TRABALHOS PUBLICADOS EM ANAIS DE EVENTOS (RESUMO)

ALVES, GSC; Mota, APZ; Fonseca, FCA; Costa, MMC; Togawa, RC; Silva-Jr., OBS; **Arraes, FBM;** Grossi-de-Sá, MF; Brasileiro, ACM; Guimaraes, PM; Miller, RNG (2020). **SMRT-RenSeq for NLR resistance gene characterization in *Arachis*, *Glycine* and *Musa* species.** *In:* XXVIII International Plant and Animal Genome Conference, San Diego.

ALVES, GSC; **Arraes, FBM;** Fonseca, FCA; Souza-Jr., JDA; Togawa, RC; Costa, MMC; Silva-Jr., OBS; Brasileiro, ACM; Guimaraes, PM; Miller, RNG; Grossi-de-Sá, MF (2019). **SMRT RenSeq-based characterization of Nucleotide-Binding-Leucine-Rich repeat gene family resistance genes in *Arachis*, *Glycine* and *Musa* species.** *In:* XXVII International Plant and Animal Genome Conference, San Diego.

ARRAES, FBM; Rocha, S; Martins-de-Sa, D; Faheem, M; Melo, BP; Souza-Jr., JDA; Noriega Vasquez, DD; Grossi-de-Sá, MF; Danchin, EGJ (2018). ***In silico* analysis reveal high variability in RNAi machinery of five different insect orders.** *In:* 7th Brazilian Biotechnology Congress, Brasilia.

ARRAES, FBM; Rocha, S; Martins-de-Sa, D; Faheem, M; Melo, BP; Morgante, CV; Souza-Jr., JDA; Noriega Vasquez, DD; Danchin, EGJ; Grossi-de-Sá, MF (2018). Insect RNAi Machinery: ***In silico* analysis reveal particularities in five different insect orders.** *In:* 12th International Congress of Plant Molecular Biology, Montpellier.

ARRAES, FBM; Faheem, M; Martins-de-Sa, D; Grossi-de-Sá, MF (2017). ***In silico* analysis reveal particularities of RNAi machinery in five different insect orders.** *In:* VI Simpósio Brasileiro de Genética Molecular de Plantas, Ouro Preto.

ARRAES, FBM; Beneventi, MA; Lisei-de-Sa, ME; Paixao, JFR; Albuquerque, EVS; Marin, SRR; Purgatto, E; Nepomuceno, AL; Grossi-de-Sá, MF (2015). **Implications of ethylene biosynthesis and signaling in soybean drought stress tolerance.** *In:* 11th International Congress of Plant Molecular Biology, Foz do Iguaçu.

GROSSI-DE-SÁ, MF; **Arraes, FBM;** Amorim, RMS; Beneventi, MA; Albuquerque, EVS; Oliveira, RS; Nepomuceno, AL; Abdelnoor, R.; Lisei-de-Sa, ME (2013). **Soybean ethylene biosynthesis genes potentially involved in drought tolerance.** *In:* XXI International Plant and Animal Genome Conference, San Diego.

ARRAES, FBM; Lisei-de-Sa, ME; Amorim, RMS; Beneventi, MA; Albuquerque, EVS; Oliveira, RS; Nepomuceno, AL; Abdelnoor, R.; Grossi-de-Sá, MF (2012). ***In silico* analysis of ACC Synthase (ACS) genes identified in brazil soybean genome consortium database.** *In:* 4º Congresso Brasileiro de Biotecnologia, Guarujá.

ARRAES, FBM; Nicola, AM; Reis, VCB; Neves, EO; Derengowiski, LS; Silva-Pereira, I; Poças-Fonseca, MJ; Felipe, MSS (2004). **Identificação, isolamento e caracterização do gene do repressor transcricional *Pbtup1* do fungo dimórfico e patogênico *Paracoccidioides brasiliensis*.** *In:* XXIV Reunião de Genética de Microrganismos, Gramado.

ARRAES, FBM; Pereira, M; Andrade, EV; Jesuíno, RSA; Simões, IC; Teixeira, MM; Bailão, AM; Silva, G; Parente, J; Camargo, J; Barbosa, MS; Castro, NS; Soares, CMA; Brígido, MM; Felipe, MSS (2004). **Metabolic analysis of the transcriptome from the fungal pathogen *Paracoccidioides brasiliensis*.** *In:* XXIV Reunião de Genética de Microrganismos, Gramado.

DERENGOWSKI, LS; Albuquerque, PA; Baptista, AJ; Machado, LS; Nicola, AM; **Arraes, FBM;** Neves, EO; Poças-Fonseca, MJ; Felipe, MSS; Silva-Pereira, I (2004). **Isolamento e caracterização de seqüências promotoras de genes expressos diferencialmente no fungo dimórfico *Paracoccidioides brasiliensis*.** *In:* XXIV Reunião de Genética de Microrganismos, Gramado.

FELIPE, MSS; Andrade, RV; Petrofeza, SS; Carvalho, MJA; Albuquerque, PA; Alves, ES; **Arraes, FBM;** Arruda, M; Azevedo, MO; Baptista, AJ; Bataus, LA; Borges, C; Campos, EG; Daher, BS; Dantas, A; Ferreira, M; Ghil, GV; Inove, MK; Kyaw, CM; Leitão, L; Maranhão, AQ; Martins, CRF; Moraes, LMP; Nicola, AM; Oliveira, AG; Parente, J; Pereira, M; Poças-Fonseca, MJ; Resende, R; Ribeiro, BM; Saldanha, R; Santos, S; Silva-Pereira, I; Silva, MAS; Silveira, E; Simões, IC; Soares, R; Souza, DP; Souza, MT; Torres, FAG; Veiga, HP; Venâncio, EJ; Walter, MMT; Soares, CMA; Brígido, MM (2003). **Initial transcriptome characterization of the dimorphic and pathogenic fungus *Paracoccidioides brasiliensis*.** *In:* XXXII Reunião Anual da Sociedade Brasileira de Bioquímica e Biologia Molecular, Caxambu.

NICOLA, AM; **Arraes, FBM;** DANTAS, A; PEREIRA, Silva-Pereira, I; Felipe, MSS (2003). **The *HSP90* gene in *Paracoccidioides brasiliensis*.** *In:* XXII Congresso Brasileiro de Microbiologia, Florianópolis.

ANDRADE, RV; Petrofeza, SS; Albuquerque, PA; Alves, ES; **Arraes, FBM;** Arruda, M; Azevedo, MO; Baptista, AJ; Bataus, LA; Borges, C; Campos, EG; Daher, BS; Dantas, A; Ferreira, M; Ghil, GV; Inove, MK; Kyaw, CM; Leitão, L; Maranhão, AQ; Martins, CRF;

Moraes, LMP; Nicola, AM; Oliveira, AG; Parente, J; Pereira, M; Poças-Fonseca, MJ; Resende, R; Ribeiro, BM; Saldanha, R; Santos, S; Silva-Pereira, I; Silva, MAS; Silveira, E; Simões, IC; Soares, R; Souza, DP; Souza, MT; Torres, FAG; Veiga, HP; Venâncio, EJ; Walter, MMT; Soares, CMA; Brígido, MM; Felipe, MSS (2002). **Partial transcriptome characterization of the dimorphic and pathogenic fungus *Paracoccidioides brasiliensis*.** *In:* VIII Encontro Internacional Sobre Paracoccidioidomicose (ARBS. Annual Review of Biomedical Sciences), Pirenópolis - GO.

ARRAES, FBM; Andrade, RV; Petrofeza, SS; Azevedo, MO; Baptista, AJ; Maranhão, AQ; Moraes, LMP; Nicola, AM; Saldanha, R; Silva-Pereira, I; Torres, FAG; Andrade, EV; Xavier, MAS; Walter, MMT; Soares, CMA; Brígido, MM; Felipe, MSS (2002). **Ubiquitin genes in the partial transcriptome characterization of *Paracoccidioides brasiliensis*.** *In:* VIII Encontro Internacional Sobre Paracoccidioidomicose, (ARBS. Annual Review of Biomedical Sciences), Pirenópolis - GO.

4. APRESENTAÇÃO DE TRABALHO E PALESTRA

ARRAES, FBM (2020). **Modulação da expressão gênica em plantas via tecnologia CRISPR-dCas9.** Conferência ou palestra, Apresentação de Trabalho. *Cidade:* Buenos Aires; *Evento:* Mejoramiento de plantas mediante edición génica basada en CRISPR-Cas9; *Instituição promotora/financiadora:* Centro Argentino-Brasileiro de Biotecnologia.

ARRAES, FBM (2020). **Modulação da expressão gênica em plantas via tecnologia CRISPR-dCas9.** Conferência ou palestra, Apresentação de Trabalho. *Cidade:* Brasilia; *Evento:* Edição de genomas em plantas (Embrapa); *Instituição promotora/financiadora:* Embrapa Agroenergia.

ARRAES, FBM; Lisei-de-Sa, ME; Amorim, RMS; Beneventi, MA; Albuquerque, EVS; Oliveira, RS; Nepomuceno, AL; Abdelnoor, R; Grossi-de-Sá, MF (2013). ***In silico* analysis of ACC Synthase (ACS) genes identified in Brazil Soybean Genome Consortium Database.** Conferência ou palestra, Apresentação de Trabalho. *Cidade:* Durban; *Evento:* IX World Soybean Research Conference; *Instituição promotora/financiadora:* Society for Soybean Research and Development.

ORGANIZAÇÃO DE EVENTOS, CONGRESSOS, EXPOSIÇÕES E FEIRAS E OLIMPÍADAS

GROSSI-DE-SÁ, MF; **Arraes, FBM;** Morgante, CV (2019). **Genetic engineering to superior crop development: CRISPR-mediated genome editing – A CBAB/CABBIO initiative.** Outro, Organização de evento.

CITAÇÕES

Google Scholar: *Total de citações:* 700. *Total de trabalhos:* 22.