Universidade de Brasília

Instituto De Biologia

Departamento de Biologia Celular

FERRAMENTA COMPUTACIONAL DE IDENTIFICAÇÃO E ANÁLISE DE REDUNDÂNCIA DE PRE-MIRNAS EM PLANTAS

Deborah Ribeiro Bambil

Doutorado em Biotecnologia e Biodiversidade

Brasília, DF – BRASIL

Julho/2023

FERRAMENTA COMPUTACIONAL DE IDENTIFICAÇÃO E ANÁLISE DE REDUNDÂNCIA DE PRE-MIRNAS EM PLANTAS

Deborah Ribeiro Bambil

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Biotecnologia e Biodiversidade da Universidade de Brasília, para obtenção do título de Doutor em Biotecnologia e Biodiversidade.

Brasília, DF – BRASIL

Julho/2023

Deborah Ribeiro Bambil

FERRAMENTA COMPUTACIONAL DE IDENTIFICAÇÃO E ANÁLISE DE REDUNDÂNCIA DE PRE-MIRNAS EM PLANTAS

APROVADA: 25 de julho de 2023

Banca Examinadora:

Dra. Priscila Grynberg – Embrapa

Dr. Roberto Coiti Togawa – Embrapa

Dr. Robert Neil Gerard Miller – UnB

Dr. Thiago José de Carvalho André – UnB

Dr. Lúcio Flávio de Alencar Figueiredo – UnB

(Orientador)

Agradecimentos

A Capes “Coordenação de Aperfeiçoamento Pessoal de Nível Superior” pelo apoio ao desenvolvimento da pesquisa de Doutorado.

Ao Departamento de Botânica da UnB, seus professores e funcionários.

Ao Departamento de Biologia celular da UnB, seus professores e funcionários.

Ao meu orientador Prof. Lúcio Flávio de Alencar Figueiredo, pela orientação, pelos conselhos, por seu apoio em todos os momentos, pela admirável dedicação com o desenvolvimento do meu doutorado, ensinamentos e amizade.

A Profa. Júlia Sonsin pelos momentos agradáveis, conselhos, ensinamentos e amizade.

Ao Prof. Thiago Amorim pelo suporte no Instituto de Brasília e pela amizade.

A minha amiga Mirele Caroline pelo suporte no desenvolvimento dos programas computacionais e pela amizade.

Aos membros da banca Dra. Priscila Grynberg (Embrapa), Dr. Roberto Coiti Togawa (Embrapa), Dr. Robert Neil Gerard Miller (UnB), Dr. Thiago José de Carvalho André (UnB), pelo tempo dedicado à leitura, análise e contribuições nesta tese.

Aos meus pais Carlos e Tânia que sempre me apoiaram.

Ao meu irmão Diego, meus sobrinhos Nicolas e Samuel, que sempre me apoiaram.

A Aline pela parceria e companheirismo em todos os momentos.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

microRNAs (miRNAs) são sequências curtas de RNAs não codificantes que atuam na expressão gênica. O objetivo deste trabalho foi realizar a mineração de pre-miRNAs e miRNAs de plantas analisando a redundância nos pre-miRNAs com a identidade entre 95 a 70% e construir uma ferramenta computacional baseada em modelos de covariância e ocultos, para identificar novos pre-miRNAs. Os pre-miRNAs selecionados tinham de 70 a 300 nt. Assim, 8045 pre-miRNAs de 2623 famílias foram minerados no miRBase, a partir de 8677 e 2942, respectivamente. Uma redução de 11 e 7% de sequências e famílias de pre-miRNAs. As angiospermas possuíam o maior número de famílias de pre-miRNAs (*n*=2202), seguido das gimnospermas (*n*=272), briófitas (*n*=121) e algas (*n*=78). A análise de redundância foi feita pela similaridade em cores classificadas com o algoritmo Deep Learning com a ferramenta Weka. A métrica medida-F, resultante do DL, apresentou o resultado da classificação por cores, que foi usada para fazer a ANOVA, onde o limite de 80% foi significativo em comparação com os outros limites. A ferramenta computacional PmiR-Select foi registrada como propriedade intelectual (nº BR512022001292). Essa ferramenta foi baseada em modelos de covariância que identificou 8470 novos pre-miRNAs no genoma do arroz, que são homólogos a 36 famílias. Dessas, 17 famílias existentes no miRBase para o arroz e 19 seriam de novas famílias, que representam um aumento de 5% de famílias de pre-miRNAs depositados para o arroz (341 famílias). Esses novos pre-miRNAs e suas famílias auxiliam o delineamento e análise de resultados de experimentos de bancada ou computacional. No genoma do ipê rosa (503 Mb), foi utilizada a PmiR-Select, que identificou 305 novos pre-miRNAs homólogos a 22 famílias de pre-miRNAs, enquanto com os modelos ocultos de Markov (HMM) foram identificados 1293 pre-miRNAs de 73 famílias. Dessas 95 familias, somente uma ocorreu em comum entre os dois modelos, fortalecendo a complementaridade deles. A PmiR-Select e o HMM estão analisando o RNA-Seq do ipê rosa e outras três árvores da família Bignoniaceae quanto a plasticidade fenotípica para seca em genes expressos diferencialmente em dois ecossistemas: i) savana e ii) floresta tropical sazonalmente seca. O ipê rosa é nativo e simbólico do bioma cerrado. O uso da PmiR-Select e o do HMM abrem oportunidades para a exploração inicial de novos pre-miRNAs de espécies nativas para diferentes clados, assim como para estratos específicos dos diversos biomas.

Palavras-chave: ipê rosa, *Handroanthus impetiginosus*, hidden Markov models, homologia, mineração de dados, modelos de covariância, pre-miRNAs, redundância.

Abstract

microRNAs (miRNAs) are small sequences of non-coding RNAs that play a role in gene expression. The study aimed to conduct the mining of plant pre-miRNAs and miRNAs by analyzing redundancy in pre-miRNAs with identities ranging from 95 to 70% and to develop a computational tool based on covariance and hidden model approaches to identify novel pre-miRNAs. The selected pre-miRNAs ranged from 70 to 300 nt. Thus, 8045 pre-miRNAs from 2623 families were mined in the miRBase, originating from 8677 and 2942, respectively; this represented an 11% and 7% reduction in pre-miRNA sequences and families. Angiosperms exhibited the highest number of pre-miRNA families (*n*=2202), followed by gymnosperms (*n*=272), bryophytes (*n*=121), and algae (*n*=78). The redundancy analysis assessed color similarity using the Deep Learning algorithm through the Weka tool. The resulting metric, measured-F from the Deep Learning, provided the outcome of color-based classification employed for ANOVA, where the 80% threshold exhibited significance compared to other thresholds. The computational tool, PmiR-Select, was registered as an intellectual property (registration no. BR512022001292). This tool successfully identified 8470 new pre-miRNAs in the rice genome, which are homologous to 36 families. Among these, 17 families already existed in the miRBase for rice, while 19 would be new families, representing a 5% increase in deposited pre-miRNA families for rice (341 families). These novel pre-miRNAs and families could aid in shaping and analyzing results from future bench or computational experiments. In the genome of the pink trumpet tree, utilizing the Hidden Markov Model-based pipeline, 1293 pre-miRNAs from 73 families were identified. Of these 95 families, only one was shared between the two models, reinforcing their complementary nature. The PmiR-Select and HMM methods are employed to analyze RNA-Seq data from pink ipê and three other trees within the Bignoniaceae family. This analysis aims to understand phenotypic plasticity in response to drought, focusing on differentially expressed genes in two distinct ecosystems: i) savanna and ii) seasonally dry tropical forest. The pink ipê is native and symbolic of the cerrado biome. The utilization of the PmiR-Select and HMM approach creates opportunities for the preliminary exploration of new pre-miRNAs from native species across various clades, as well as for specific strata within diverse biomes.

Keywords: covariance models, data mining, pink ipê, *Handroanthus impetiginosus*, hidden Markov models, homology, pre-miRNAs, redundancy.

Introdução geral

Os microRNAs (miRNAs) foram descritos pela primeira vez em 1993, quando foram identificados em um nematóide (*Caenorhabditis elegans* M.). O miRNA identificado foi classificado como um regulador pós-transcricional do gene *lin*-14, envolvido no controle do desenvolvimento larval do nematoide. A descoberta revelou que o gene *lin*-4 não codificava uma proteína, mas expressava um pequeno RNA de aproximadamente 22 nucleotídeos. Esse achado possibilitou a descrição de um novo mecanismo de regulação gênica, no qual o transcrito do gene *lin*-4 regula negativamente a tradução do transcrito de *lin*-14 (Lee et al. 1993). Na época, não foi dado muito destaque pela comunidade científica para essa descoberta; ela foi considerada uma “anomalia de baixa probabilidade da natureza" (Wang et al. 2022). Em plantas, os primeiros miRNAs (*n*=16) foram descritos em 2002 em arabidopsis [*Arabidopsis thaliana* (L.) Heynh]. Esse estudo também demonstrou que alguns desses miRNAs (*n*=8) eram conservados no genoma do arroz (*Oryza sativa* L. – Reinhart et al. 2002). Um exemplo dessa conservação é o miR159 cuja conservação transcende espécies de diferentes clados como licófitas, samambaias, pinheiros e angiospermas; com a mutação de apenas uma base. A mesma conservação ocorre nos sítios de ligação desse miRNA para as espécies dos clados citados (Millar et al. 2022).

Os genomas dos eucariotos contêm milhares de RNAs não codificantes (ncRNAs), que desempenham papéis cruciais na regulação transcricional e pós-transcricional da expressão gênica (Raza et al. 2023). Nesta classe estão agrupados os ncRNAs estruturais: RNA transportador (tRNAs), RNA ribossômico (rRNAs), RNAs nucleares (snRNAs) e RNAs nucleolares (snoRNAs). Já os ncRNAs regulatórios são: ncRNAs longos (lncRNAs), pequenos RNAs interferentes (siRNAs) e os miRNAs. Os miRNAs estão envolvidos na regulação da expressão gênica pela clivagem do mRNA alvo ou inibição da tradução (Bhogireddy et al. 2021; Ponting et al. 2009; Achkar et al. 2016; Waititu et al. 2020). Os miRNAs em plantas são classificados em três categorias: miRNAs primários (pri-miRNAs), precursores (pre-miRNAs) com comprimento máximo de 300 nucleotídeos (nt) e os miRNAs maduros variando entre 20 a 24 nt (Axtell and Meyers, 2018).

O processo de biogênese do miRNA começa no núcleo da célula. Os genes MIR são transcritos no miRNA primário (pri-miRNA) pela enzima RNA polimerase II e HYPONASTIC LEAVES 1 (HYL1), o pri-miRNA é clivado pelas enzimas Dicer-like 1 (DCL1), que se associam às proteínas de ligação Serrate (SE) e HYL1 para clivar e gerar o pre-miRNA. O pre-miRNA é metilado pela enzima metiltransferase nuclear Hua Enhancer 1 (HEN1), prevenindo assim sua degradação. O miRNA Duplex é exportado do núcleo para o

citoplasma com suporte da enzima HASTY (HST). No citoplasma o duplex miRNA: miRNA* é separado e o miRNA maduro é incorporado ao complexo de silenciamento de indução do RNA (RISC), que contém a proteína Argonaute 1 (AGO 1). Para se ligar aos RNAs mensageiros (mRNAs) por alvo de complementariedade, promovendo a clivagem ou inibição de tradução, enquanto o miRNA* é excluído do RISC exposto para degradação (Figura 1.1 – Waititu et al. 2020).

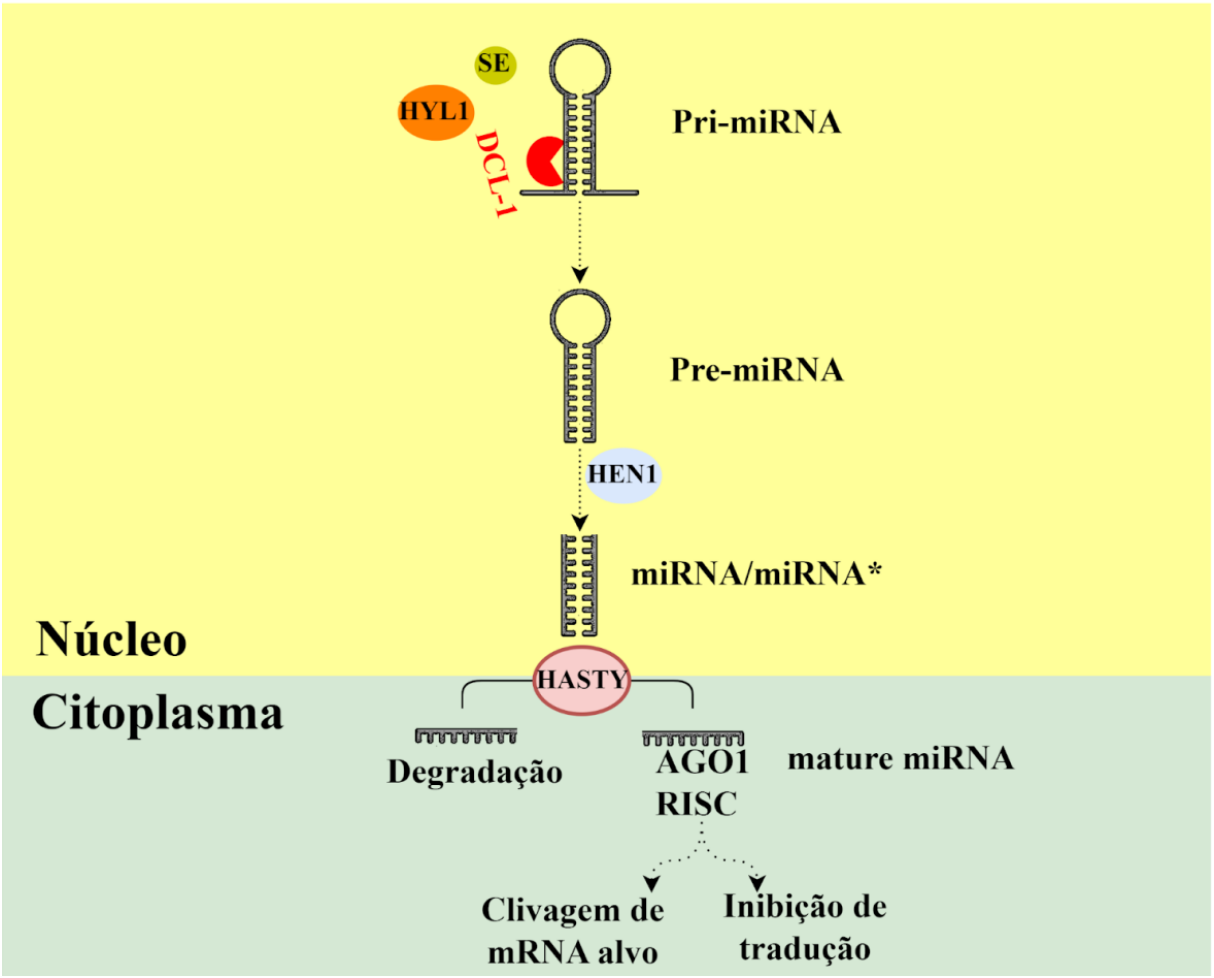


Figura 1. Caracterização do processo de biogênese de miRNAs em plantas. O pri-miRNA é clivado pela enzima Dicer-like 1 (DCL1), juntamente com as proteínas Serrate (SE) e HYL1 (Leaves 1). Em seguida, ocorre a metilação do pre-miRNA pela enzima Hua Enhancer 1 (HEN1). O miRNA duplex é transportado para o citoplasma, o miRNA maduro é incorporado no complexo de silenciamento RISC (RNA-Induced Silencing Complex) em conjunto com a proteína Argonauta 1 (AGO1), para clivagem de miRNA alvo ou inibição de tradução, e miRNA* é degradado (Figura adaptada de Waititu et al. 2020).

Os miRNAs atuam em vários estágios de desenvolvimento da planta, como crescimento, florescimento, desenvolvimento de frutos, grãos e envelhecimento das plantas (Raza et al. 2023; Li and Yu, 2021). Eles também desempenham um papel crucial nas respostas das plantas a estresses bióticos, como a resistência a insetos, bactérias, fungos, nematóides e vírus, atuando como mecanismos de defesa (Khraiwesh and Zhu, 2012; Barah et al. 2013). Como por exemplo, os miRNAs (miR156, miR159, miR162 and miR395) que emergem como importantes reguladores da resposta a herbivoria de pulgões em diferentes estágios do desenvolvimento melão (*Cucumis melo* L. - Sattar et al. 2012).

Além disso, os miRNAs também respondem aos estresses abióticos, como resistência a salinidade, seca, metais pesados, temperaturas extremas, desequilíbrio de nutrientes e fatores sazonais que podem limitar o desenvolvimento das plantas (Fujita et al. 2006; Raza et al. 2023; Wang et al. 2003, Raza et al. 2023). Estudos com o milho (*Zea mays* L.) demonstraram que o miR156 é capaz de regular a resposta ao estresse salino nas raízes (Ding et al. 2009). Em arabidopsis, os miRNAs (miR168, miR171 e miR396) respondem ao estresse causado pela seca e pelo frio (Liu et al. 2008).

Nas últimas décadas, houve um aumento substancial do interesse em estudos relacionados à miRNAs, evidenciado por mais de 31 mil artigos científicos publicados sobre o assunto (https://pubmed.ncbi.nlm.nih.gov/, acessado em 12 de junho de 2023). O miRBase, fundado em 2002, foi o primeiro e é atualmente a maior base de dados de pre-miRNAs e miRNAs disponível. O número de registros nele cresceu cerca de 180 vezes nos últimos 20 anos. Na primeira versão (v.1.1), havia apenas 262 sequências de pre-miRNAs e 266 de miRNAs maduros. Na versão atual (v.22) contém mais de 38 mil pre-miRNAs e quase 49 mil miRNAs maduros.

A identificação de miRNAs maduros envolve a localização no pre-miRNA. Assim, existem algumas limitações na identificação dos miRNAs maduros porque eles não possuem extremidades características. Adiciona-se a isso o tecido e momento específico da expressão em baixos níveis. Já a dificuldade computacional, ocorre devido a busca se basear em pre-miRNAs e miRNAs maduros conhecidos (Sun et al. 2014). Portanto, a abordagem experimental, torna-se necessária para identificar os miRNAs maduros. Na área de tecnologia da informação, visando atender critérios de tratamento dos dados, são descritos protocolos que utilizam processos que dispõem de técnicas para mineração de dados (Brackett and Earley, 2009). A mineração de dados tem uma relevância fundamental para evitar a classificação de falsos positivos. Por essa razão, adotar critérios para classificação de pre-miRNAs de plantas se torna essencial.

Esses processos são difíceis de serem organizados pelo olhar humano e, em muitos casos, são impossíveis de serem realizados manualmente devido ao tamanho das bases de dados. Assim, as ferramentas computacionais podem automatizar esse processo, permitindo que as análises que levariam dias ou até meses sejam realizadas em minutos ou horas (Triguero et al. 2019). A mineração de dados é um processo crucial para a gestão de qualidade dos dados, pois permite a descoberta de padrões e correlações que possibilitam a previsão de resultados em big data e data science. As técnicas utilizadas para a realização desse processo são descritas em manuais especializados, com o objetivo de evitar a duplicação de dados. Essa é uma técnica importante para evitar cópias nas bases de dados, enquanto a normalização de dados é útil para remover a redundância (Brackett and Earley, 2009).

Empresas, como a Microsoft, têm utilizado protocolos semelhantes para criar ferramentas genômicas de alto desempenho, como o projeto Microsoft Genomics (https://www.microsoft.com/en-us/genomics/), que gerencia todos os tipos de dados genômicos na nuvem (Yang-Turner et al. 2018). Essas ferramentas são baseadas em técnicas de mineração de dados, que têm permitido a análise de grandes quantidades de dados em pouco tempo.

1. Ferramentas computacionais e base de dados de pre-miRNAs e miRNAs maduros

Existem diversas ferramentas computacionais disponíveis com diferentes funções no campo da biologia molecular. Algumas delas são voltadas para a classificação de processos biológicos, como a ferramenta UEA sRNA workbench (Stocks et al. 2012). Enquanto outras têm a função de identificação de pre-miRNAs e miRNAs maduros. Aqui, mostramos vinte ferramentas na Figura 2, sendo que oito não estão mais disponíveis em seus repositórios. Todas as ferramentas mencionadas utilizaram o miRBase como base de dados de referência para a construção de seus pipelines para identificação. Além das ferramentas terem sido construídas com a base de dados do miRBase, outras bases de dados foram construídas a partir do miRBase (Tabela 1).

Desde seu lançamento em 2002, o miRBase (https://mirbase.org/ftp/) vem se consolidando como uma referência importante para a anotação e identificação de pre-miRNAs e miRNAs maduros em diversas espécies. A versão inicial do miRBase, a v.1, possuía um depósito de 218 sequências de pre-miRNAs (hairpin) e 266 sequências de miRNAs maduros.

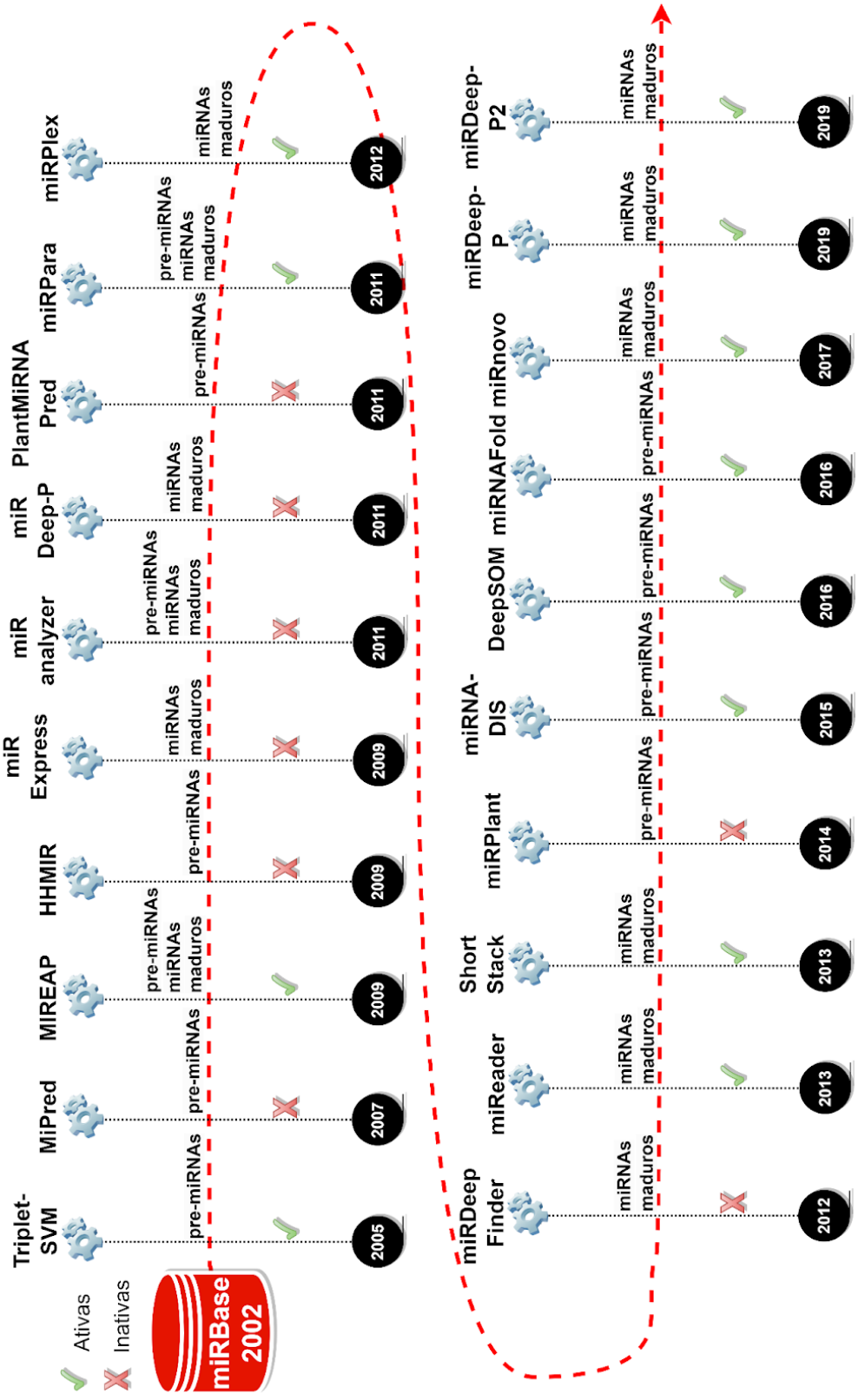


Figura 2: Linha do tempo com as ferramentas computacionais de identificação de pre-miRNAs e miRNAs maduros de plantas (Chen et al. 2019; Morgado and Johannes 2019; Yu et al. 2020).

Tabela 1. Bases de dados de pre-miRNAs e miRNAs maduros construídas a partir do miRBase.

Banco de dados	Classes de miRNAs	Referências
DPMIND	miRNAs maduro	Fei et al. 2018
MepmiRDB	pri-miRNAs, pre-miRNAs e miRNAs maduro	Yu et al. 2019
MirCARTA	pre-miRNAs e miRNAs maduro	Backes et al. 2018
miRVIT	miRNAs maduro	Chitarra et al. 2018
PAMIRDB	miRNAs maduro	Satish and Mukherjee, 2019
PlantCircNet	miRNAs maduro	Zhang et al. 2017
PmiREN	pre-miRNAs e miRNAs maduro	Guo et al. 2020
PMRB	pre-miRNAs e miRNAs maduro	Zhang et al. 2010
TarDB	miRNAs maduro	Liu et al. 2021

Atualmente, na versão v.22.1 do miRBase, o número de sequências depositadas cresceu significativamente, totalizando 38589 sequências de pre-miRNAs e 48885 sequências de miRNAs maduros (http://www.mirbase.org). Desse total, 8677 são sequências de pre-miRNAs de plantas e 10491 são sequências de miRNAs maduros de plantas. Os pre-miRNAs de plantas presentes no miRBase pertencem a espécies dos clados de algas (ALG), briófitas (BRY), gimnospermas (GYM) e angiospermas (ANG - Figura 3). O miRBase continua sendo alimentado periodicamente e serve como o maior repositório de miRNAs.

A família Fabaceae (ex-família Leguminosae) classificada no clado das angiospermas é a família de plantas com a maior quantidade de sequências de pre-miRNAs (*n*=1727) no miRBase. A maioria dos membros dessa família é composta por plantas fixadoras de nitrogênio (Wink, 2013). A família Poaceae (ex-família Gramineae) vem em segundo lugar em número de sequências de pre-miRNAs no miRBase v22.1 (*n*=1313). Os membros desta família são fundamentais como fonte de alimentos primários, ração animal e biocombustíveis (Edwards et al. 2010). Além disso, a família Poaceae é um modelo importante para estudos evolutivos (Hodkinson, 2018).

O miRBase adota uma nomenclatura que utiliza as três primeiras letras para abreviar o nome científico da espécie taxonômica, a que o pre-miRNA pertence (por exemplo, "ath" para *Arabidopsis thaliana*). Em seguida, as três letras representam o prefixo "MIR", indicam que se trata de um pre-miRNA de planta. Os últimos caracteres da nomenclatura dos pre-miRNAs no miRBase são números identificadores da família correspondente, seguido por um caractere alfabético sequencial (por exemplo, 156a). Cabe ao autor (contribuinte) enviar a sequência de pre-miRNAs para publicação na base de dados, sendo sua responsabilidade garantir a

qualidade da anotação e a indicação correta da família em que a sequência será agrupada (Kozomara and Griffiths, 2014).

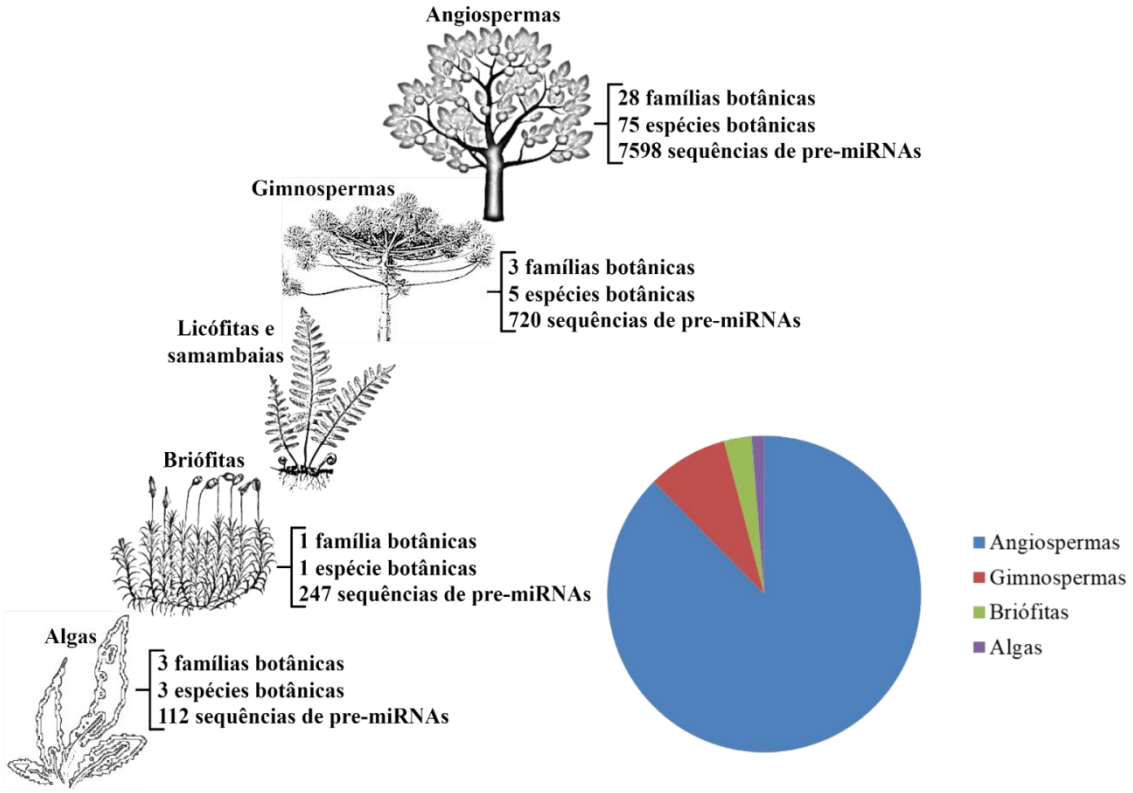


Figura 3: Número de famílias, espécies botânicas e sequências de pre-miRNAs de plantas no miRBase v22.1 por clados de plantas.

Os pre-miRNAs são adicionados sequencialmente no miRBase, o que permite o registro contínuo de novas sequências. A aquisição de dados na base de dados miRBase depende exclusivamente do envio das sequências pelos autores, sem uma revisão rigorosa quanto à qualidade e critérios. Nesse sentido, é cada vez mais importante a mineração de dados antes de incluí-los nos repositórios, a fim de evitar a presença de artefatos computacionais e garantir a integridade dos dados.

Outra base de dados de RNAs é o Rfam que contém famílias de RNA com informações das estruturas secundárias de consenso e modelos de covariância. O Rfam tem atualmente cerca de 1100 famílias de pre-miRNAs de diversos organismos (https://rfam.org/search?q=precursor%20AND%20entry_type:%22Family%22). É importante destacar que existe um projeto de colaboração entre Rfam e o miRBase para integrar as bases de dados (https://rfam.xfam.org/microrna).

1.3 Objetivo geral

O objetivo deste trabalho foi descrever um pipeline com a mineração de dados de pre-miRNAs e miRNAs cuja redundância foi analisada visando construir uma ferramenta computacional baseada em modelos de covariância para identificar novos pre-miRNAs no genomas de plantas.

1.3.1 Objetivos específicos

- Realizar a mineração dos conjuntos de dados de pre-miRNAs e miRNAs maduros de planta disponíveis no miRBase.
- Analisar os limites de identidade para indicar um conjunto de dados sem redundância de pre-miRNA, a fim de obter um conjunto de exemplares.
- Construir modelos de covariância na ferramenta PmiR-Select para a identificação de novos pre-miRNAs em sequências genômicas no arroz e ipê rosa.
- Identificar pre-miRNAs conhecidos com o pipeline baseado nos modelos ocultos de Markov (hidden Markov models), no genoma do ipê rosa.

1. Descrição dos capítulos

No Capítulo 2, é apresentada uma proposta de artigo, a ser submetido. Ele aborda o desenvolvimento e as rotinas da ferramenta computacional PmiR-Select, utilizada para a identificação de novos pre-miRNAs em genomas. O título do artigo é: PmiR-Select - a computational approach to plant pre-miRNAs identification of pre-miRNAs in plant genomes.

No Capítulo 3, é apresentada uma proposta incompleta para ser o segundo artigo, a ser submetido. Ele aborda a identificação de novos pre-miRNAs com a ferramenta PmiR-Select e de pre-miRNAs conhecidos baseados no pipeline em modelos ocultos de Markov. O título do artigo é: Computational identification of pre-miRNAs on pink ipê (*Handroanthus impetiginosus* Mart. ex DC.): a native cerrado species.

Após o capítulo 3 é apresentado uma conclusão geral, considerações finais e perspectivas futuras.

Capítulo 2

Desenvolvimento da ferramenta computacional PmiR-Select para identificação de pre-miRNAs

2.1 Introdução

Nos últimos anos, houve um crescimento de ferramentas computacionais capazes de identificar pre-miRNAs e miRNAs maduros (Morgado and Johannes, 2019). A abordagem computacional tem se mostrado valiosa para a pesquisa em biologia molecular, pois oferece uma opção não onerosa para realizar experimentos (Sablok et al. 2012). Todavia, a validação das identificações feitas por ferramentas computacionais devem ser realizadas por meio de experimentos de bancada (Meng et al. 2012). Assim, a identificação de pre-miRNAs feita por ferramentas computacionais tem se mostrado eficiente nas estruturas de plantas, uma vez que a estrutura de miRNAs maduros é conservada nesses organismos, o que representa uma vantagem esperada para busca por homologia (Zhang et al. 2006; Sun et al. 2014).

Com esse intuito, foi construída uma ferramenta computacional (PmiR-Select) para identificar novos pre-miRNAs baseado em homologia. Inicialmente realizou-se uma mineração de dados, para filtrar pre-miRNAs (70-300 nt) e miRNAs maduros (20-24 nt) de plantas, de acordo com os critérios atualizados (Axtell and Meyers, 2018). Após isso, foi feita uma análise de redundância das sequências de pre-miRNAs por família, considerando identidades entre 95 a 70% (intervalos de 5%), a fim de sugerir um limite sem redundância, para reduzir e facilitar a análise dos dados.

A partir da mineração e eliminação da redundância dos pre-miRNAs, foram construídos modelos de covariância (MC) para a identificação de novos pre-miRNAs em genomas. Os MC foram descritos pela primeira vez em 1994 (Eddy and Durbin, 1994), a partir de sequências conhecidas, baseadas na análise de estrutura secundária conservada que formam ligações entre pares de bases e mantém interações por covariância entre as bases. Essas interações são usadas para identificar novas sequências de ncRNAs, neste caso de pre-miRNAs (Nawrocki and Eddy, 2013).

Os novos pre-miRNAs identificados são comparados com o conjunto de dados de miRNAs maduros a fim de avaliar a conservação, e por fim, os resultados são apresentados em diferentes diretórios que separam os pre-miRNAs identificados em limite de redundância de identidade. A ferramenta PmiR-Select identificou novos pre-miRNAs em sequências genômicas de plantas. A ferramenta PmiR-Select encontra-se disponível na plataforma GitHub (https://github.com/DeborahBambil/PmiRSelect) e foi registrada como propriedade intelectual no Instituto Nacional da Propriedade Industrial (INPI), processo nº BR512022001292, expedida em 31 de maio de 2022.

2.2 Certificado de registro de propriedade intelectual

REPÚBLICA FEDERATIVA DO BRASIL
MINISTÉRIO DA ECONOMIA
INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512022001292-8**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 08/03/2022, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: Análise de miRNA - PmiR Select

Data de publicação: 08/03/2022

Data de criação: 14/06/2021

Titular(es): FUNDAÇÃO UNIVERSIDADE DE BRASILIA

Autor(es): LÚCIO FLÁVIO DE ALENCAR FIGUEIREDO; DEBORAH RIBEIRO BAMBIL; MIRELE CAROLINA SOUZA FERREIRA COSTA

Linguagem: PYTHON

Campo de aplicação: BL-01; BL-02

Tipo de programa: TC-01

Algoritmo hash: SHA-512

Resumo digital hash:
77907CA950DA9E7957BFE1F5E748C10D5F448F2D15A9829AFAAC4E2E1D9EE4F73C0D54B57E2E9D4EA580B9E
26FB775E1C19E428685E078ACD2FDC596A1A2B545

Derivação autorizada: Sim. Existe uso derivado e autorizado dos seguintes softwares e banco de dados: BLAST/ EMBOSS/ GenomeTools/ Infernal/ UnRAR

Expedido em: 31/05/2022

Aprovado por:
Joelson Gomes Pequeno
Chefe Substituto da DIPTO - PORTARIA/INPI/DIRPA Nº 02, DE 10 DE FEVEREIRO DE 2021

2.3 PmiR-Select - a computational approach to plant pre-miRNAs identification in plant genomes

Deborah Bambil[1, 2]*, Mirele Costa[3], Lúcio Flávio de Alencar Figueiredo[4]

[1]Department of Cell Biology, Biology Institute, University of Brasília (UnB), Brasília, 70910-900, DF, Brazil

[2]Federal Institute of Brasília (IFB) Brasília, 70830-450, DF, Brazil

[3]Department of Computation, UnB, Brasília, 70910-900, DF, Brazil

[4]Department of Botany, Biology Institute, UnB, Brasília, 70910-900, DF, Brazil

Corresponding author Deborah Bambil*: Email: deborahbambil@gmail.com

Deborah Bambil - https://orcid.org/0000-0001-8307-0888

Lúcio Flávio de Alencar Figueiredo - https://orcid.org/0000-0001-8868-1717

Mirele Costa - https://orcid.org/0000-0002-1337-4672

Abstract

microRNAs (miRNAs) are small non-coding RNAs involved in regulating gene expression. We described a pipeline to obtain a dataset mined by criteria indicated sequences longer than 300 nt for pre-miRNAs and between 20 and 24 nt for mature miRNAs, also tested from 95 to 70% (5% intervals), for this was evaluate matches between sequences with colored classified alignments with deep learning. The update criteria generated a reduction sequence. However, we estimated an 80% redundancy threshold for a non-redundant dataset and built covariance models (CMs) to identify homology in genome sequences. Showed the tool PmiR-Select, based approach CMs for identified pre-miRNAs, output clustered at homologs families, and separation of identical copies to facilitate analysis. Furthermore, the other datasets with thresholds 70, 75, 85, 90, 95, and 100% in CMs also are available. In these cases, if the user wants to try it in PmiR-Select, even as the dependencies, source code, and user instructions are at https://github.com/DeborahBambil/PmiRSelect.

Keywords: covariance models, homology, plant pre-miRNAs, novel pre-miRNAs, non-redundant

Introduction

The microRNAs (miRNAs) are non-coding RNAs (ncRNAs) that play a crucial role in regulating gene expression (Bhogireddy et al. 2021; Waititu et al. 2020). The discovery of the first miRNA dates back 30 years ago in a nematode (*Caenorhabditis elegans* M. - Lee et al. 1993), which also happens to be the first multicellular organism sequenced. Notably, the first miRNAs (*n*=16) identified in plants were in arabidopsis [*Arabidopsis thaliana* (L.) Heynh] almost 10 years after their discovery in *C. elegans* (The Arabidopsis Genome Initiative, 2000), which was the first sequenced plant. This study also showed that eight miRNAs were conserved in the rice (*Oryza sativa* L.) genome (Reinhart et al. 2002).

The plant miRNA biogenesis begins in the nucleus. MIR genes are transcribed into primary miRNA (pri-miRNA) by polymerase II enzyme and HYPONASTIC LEAVES 1 (HYL1). It is cleaved by Dicer-like enzyme 1 (DCL1), which associates with serrate binding proteins (SE) and HYL1 to cleave and generate pre-miRNA. The pre-miRNA is methylated by the nuclear methyltransferase enzyme, Hua Enhancer 1 (HEN1), thus preventing its degradation. The miRNA duplex is exported from the nucleus to the cytoplasm with the support of the HASTY enzyme (HST). In the cytoplasm, the miRNA:miRNA* duplex is separated, and the mature miRNA is incorporated into the RNA-mediated silencing complex (RISC), which contains the Argonaute 1 (AGO 1) protein. To bind to messenger RNAs (mRNAs) by targeting complementarity, promoting cleavage or inhibition of translation, miRNA* is excluded from the exposed RISC for degradation (Figure 1.1 – Waititu et al. 2020). pre-miRNAs had recently updated criteria with a limit length of 300 nt and mature miRNAs from 20 to 24 nt (Axtell and Meyers 2018).

Plant miRNAs act in the post-transcriptional process and gene expression control (Lee et al. 1993; Reinhart et al. 2002; Bartel 2009). Related to developmental stages such as vegetative growth, flowering, plant senescence, biotic stress such as resistance virus, fungus, nematode, and abiotic stress as the response to salinity, drought, and extreme temperature stress (Raza et al. 2023).

In the last ten years, 32 thousand articles related to miRNAs were published, while more than 24 thousand articles (18%) are about small RNA sequencing. Most of these publications (16 thousand) were related to human miRNAs. Around 2 thousand miRNA papers are specifically for plants (https://pubmed.ncbi.nlm.nih.gov/, accessed on June 12, 2023).

These studies showed strategies used to identify miRNAs being computational and experimental approaches. Bioinformatics has been booming because of its low cost and high efficiency. The structure of mature miRNAs in plants is conserved, an advantage expected by

homologs or orthologs (Zhang et al. 2006). An example of such conservation is observed in miR159, wherein conservation transcends across species from distinct clades, including lycophytes, ferns, pines, and angiosperms, with mutations occurring in solitary nucleotides. Identical conservation is also evident in the binding sites of this miRNA across the aforementioned taxonomic groups (Millar et al., 2022). However, the identification of mature miRNAs involves the location in pre-miRNA. There are some limitations because the mature miRNAs do not have nucleotide-level precision and do not have characteristic ends in miRNA molecules (Sun et al. 2014). Therefore, the experimental approach, such as cloning small plant RNAs, becomes necessary to identify the mature miRNAs. Still, the identification is difficult by cloning due to the sensitivity of miRNAs to respond to specific stresses to cleave or degrade target mRNAs (Sun et al. 2014).

The miRNAs are deposited in public repositories, such as miRBase, the primary public repository of miRNA data, in version 22.1. miRBase (Kozomara et al. 2019) accounted for 48860 mature and 38589 pre-miRNAs (http://www.mirbase.org/ accessed on June 14, 2023). Since 2002, when there were 218 pre-miRNAs, there has been an increase of around 180 times to the current version.

The criteria for miRNA record is based on scientific name. The first letter came from the plant genus, and the two last are from the species name (ath - *A. thaliana*). The three consecutive characters are capital letters corresponding to the plant pre-miRNA. The last three characters are numbers of miRNA family (MIR156) added in miRBase, sequentially (Xu et al. 2018; Griffiths-Jones 2004).

miRNA homolog to one existing in miRBase composes an existing miRNA family, recorded with the same number of the miRNA already deposited. They are distinguished by the organism, like the ath-MIR156a for arabidopsis) and osa-MIR156a for rice. If the miRNA is a homolog to one existing from the same organism, it will be recorded with the same number and new sequential letter (Griffiths-Jones 2004) - for example, the ath-MIR156a and ath-MIR156b. The former has 123 nt placed on chromosome 2 and 183 on chromosome 4; both have 20 nt in mature sequence (https://www.mirbase.org/ - Lin et al. 1999; Mayer et al. 1999).

The pre-miRNA homology sequences are grouped based on similarities. Families of pre-miRNAs in miRBase can contain countless sequences, such as the miR156, which has 371 copies. This family (miR156) has higher cleavage activity and contributes to polymorphism in mature miRNAs or mutations of a single sequence, which happens in pre-miRNA transcriptions to generate mature miRNAs (Cui et al. 2017; Ma et al. 2021).

Thus, part of the pre-miRNA deposited in the miRBase are duplications, and these events are imported for evolutionary studies, which can result in redundancy and become a challenge for analysis (Axtell and Meyers 2018). However, bioinformatics tools can help manage data sets to become non-redundant. Machine learning strategies are increasingly used for biological discovery and make possible computational identification patterns in large volumes of data (Greene et al. 2014). The deep learning technique shows better analytical solutions (Zhang et al. 2018).

In the big data era, the following years' genomic data will change the scenario for exabytes (https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science, accessed on March 02, 2023), with plants' genome projects contributing to this increase, such as 1001 genomes of arabidopsis (https://1001genomes.org/, accessed on March 02, 2023), 3000 rice genomes project (http://rice.uga.edu/, accessed on March 02, 2023), the project for understanding evolution and domestication histories of grape (3525 acc: accessions - Dong et al., 2023) and sorghum association panel project (400 acc - Boatwright et al., 2022), are encouraged.

Here, the pipeline removed long sequences (longer than 300 nt) according to updated criteria for plant pre-miRNAs and mature miRNAs annotations (Axtell and Meyers 2018). Furthermore, studies showed that miRNAs in plants originated from or had expanded families from either tandem duplication events, segmental duplication events, or a combination of the two, and they evolutionarily conserved (Budak and Akpinar 2015; Hajieghrari and Farrokhi, 2022; Sun et al., 2012). Considering this, we added a threshold calculation to the tool for removing redundancy using the classification of alignments by Deeplearning4j (Lang et al. 2019). The PmiR-Select tool provided data mining and non-redundant plant pre-miRNAs dataset in covariance models identified by homology in genomes. A flowchart details the tools developed (Figure 1).

Methodology

Dataset analysis of pre-miRNA and mature miRNA

The pipeline begins applying filters (Figure 1a) in the hairpin (pre-miRNAs) file v.22.1 in the miRBase (Griffiths-Jones 2004). First, plant pre-miRNA were selected from other organisms (available at https://github.com/DeborahBambil/FilterPlant), and long sequences (length of 300 nt - available at https://github.com/DeborahBambil/Filter300NT) were removed, according to updated criteria for plant pre-miRNA annotations (Axtell and Meyers 2018). This plant pre-miRNAs dataset was separated by families (available at

https://github.com/DeborahBambil/SplitFamilies), and the pipeline started the redundancy threshold tests of similarity from 95 to 70% (5% intervals). These six datasets had the redundancy for each pair of sequences recognized in the global alignment analyzed through the skipRedundant at EMBOSS tool. Needleman and Wunsch's algorithm calculated the redundancy, and the shortest redundant sequences were discarded (Sikic and Carugo, 2010).

We classified the appropriate threshold using alignment images generated by the t-Coffee tool (Tommaso et al. 2011), which aligned each miRNA family per the redundancy threshold test. The t-Coffee algorithm calculated the alignments with its library to classify the matches of local and global pair alignments. The alignment qualification made in colors looked to evaluate the matches among sequences. The output generated in ASCII format in colors indicates the alignment with the normalized residue. This scale represents the consensus score for each column. The columns were coded from dark blue (from residue=0 to <9, BAD and AVERAGE alignment - AVG) to dark pink (residue ≥ 9-28, GOOD alignment).

Alignment classified as BAD (blue color tones - hexadecimal colors: #9b92ff, #b4ffb4, #beffbe, #c8ffc8) had a low level of matches in the sequence alignments. The BAD alignment represents a favorable match of sequences retained in this pipeline. AVG (orange color tones - hexadecimal colors: #ffffb7, #ffffad) means an average level of matches in the alignments. Some of these sequences are retained. GOOD (pink color tones - hexadecimal colors: #ffe6e6, #ffdcdc, #ffc8c8) has a high similarity between sequences. They are the majority of sequences eliminated after the redundancy threshold test. The dataset alignments were generated using the Inovtaxon tool, which works with data mining files (Attribute Relation File Format-ARFF) according to the features of colors (Bambil et al. 2020). These ARFF files were classified in the Weka tool with the Deeplearning4J algorithm, configured with ten epochs (dataset processing number). This algorithm runs with different datasets (images, CSV files, plain text, audio, and video). It is supervised to perform pattern classification and recognition tasks based on *n*-dimensional array libraries (Lang et al. 2019).

Looking to uncover the appropriate redundancy threshold, we evaluated the classification from the Deeplearning4J algorithm, with an F-measure metric and harmonic average among true positive, false positive, and false negative variance analysis (One-Way ANOVA, the *p*-value of 0.05 - *R* core team 2020). A *p*-value lower than 5% rejects the null hypothesis when all data samples have the same mean (Powers, 2020). Based on ANOVA results.

Furthermore, we analyzed the mature miRNAs to compare in the PmiR-Select from the pre-miRNAs identified. Thereunto, we filtered mature v.22.1 file only plant mature miRNAs and

a limited length of 20 between 24 nt (Figure 1b), according to updated criteria for plant miRNA annotations (Axtell and Meyers 2018).

Approach and implementation PmiR-Select

In the mined plant pre-miRNAs dataset (*n*=8045), Stockholm alignment with secondary structures was constructed (Yu et al., 2020) to build covariance models (CMs) using known sequences. These CMs are probabilistic models based on the analysis of conserved secondary structures that form bonds between base pairs and maintain covariance interactions between bases; these interactions are used to identify new ncRNA sequences (Nawrocki and Eddy, 2013). Subsequently, they were implemented in PmiR-Select to identify new pre-miRNA within a known pre-miRNA family (Figure 1c - Grüning et al., 2017).

The PmiR-Select started with a genome sequence file for pre-miRNAs identification through the CMs (queries) into the extended sequences and stem-loop structures, which are more evolutionarily conserved (Nawrocki and Eddy 2013, Grüning et al. 2017). For the validation, the new pre-miRNAs will be compared with the dataset of mature miRNAs using the blast tool (E-value 0.01) was utilized for miRNA similarity identification, which is considered significant and deserving of further investigation (Griffiths-Jones 2010). The available pre-miRNA outputs were: i) new pre-miRNAs identified, ii) identical and non-identical copies on different redundancy threshold percentual, and iii) pre-miRNAs identified and data mining related to mature miRNA (Figure 1 - PmiR-Select is available at https://github.com/DeborahBambil/PmiRSelect).

We validated the PmiR-Select with rice (Poaceae family). Rice has a lot of pre-miRNAs identified in different studies (ID4530 - https://www.ncbi.nlm.nih.gov/genome/?term=oryza+sativa accessed on September 28, 2022).

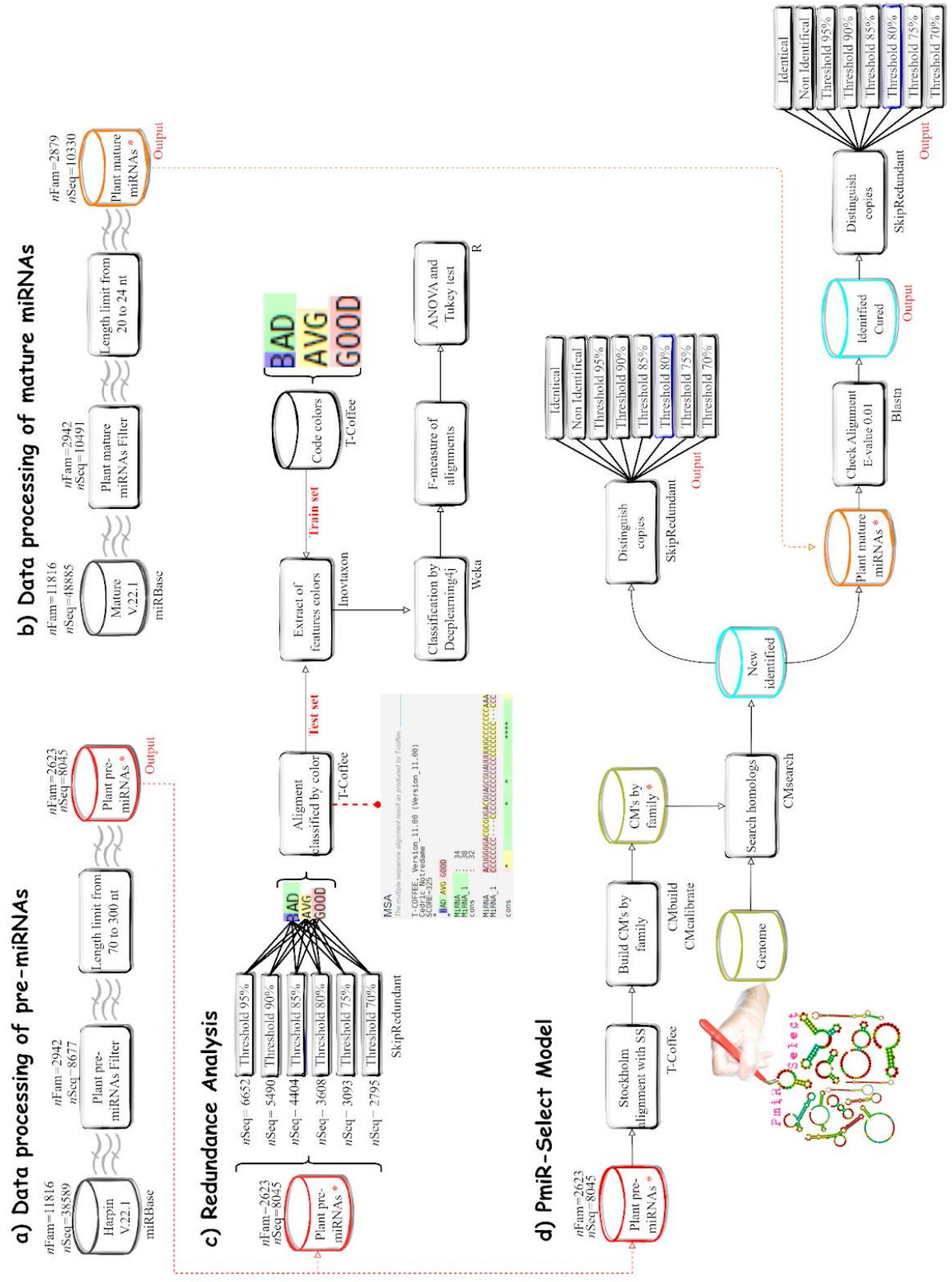


Figure 1. PmiR-Select flow. a) Data mining pre-miRNA: the first step was checking the dataset pre-miRNAs (hairpin v.22.1 MirBASE) and selecting only the plant pre-miRNAs (FilterPlant) and filter pre-miRNAs with a limit length of 70 to 300 nt. a) Data mining mature

miRNA: mature miRNAs (mature V.22.1 MirBASE) and selecting only the plant mature miRNAs (FilterPlant) and filter mature miRNAs with a limit length of 20 between 24 nucleotides. c) An analysis was performed on the data mining plant pre-miRNA dataset using redundancy threshold levels of 95, 90, 85, 80, 75, and 70% (skipRedundant - Sikic and Carugo 2010). For each redundancy threshold dataset, the alignment is classified in color (T-Coffee, Tommaso et al. 2011) by the miRNA family. These colored alignments extracted the color features (Inovtaxon - Bambil et al. 2022) and the color code of the T-Coffee to training input. The classification was performed using Deeplearning4j (Weka), with the F-measure metric being reported. The evaluation was conducted using ANOVA, followed by the Tukey test (R software), which indicated the non-redundant threshold of 80% redundancy. d) Model PmiR-Select: From a data mining of plants, Stockholm alignments containing secondary structures were constructed to build CMs (covariance models) for each pre-miRNA family. The CMs serve as inputs for the tool, along with the user-inserted sequence input 'genome.fa'. The search for new pre-miRNAs begins, and the outputs are distributed into directories that differentiate identity copies. Additionally, a separate directory is used for sequences that have undergone a check against the plant dataset of mature miRNAs, with only the sequences showing significant similarity (e-value 0.01) being included.

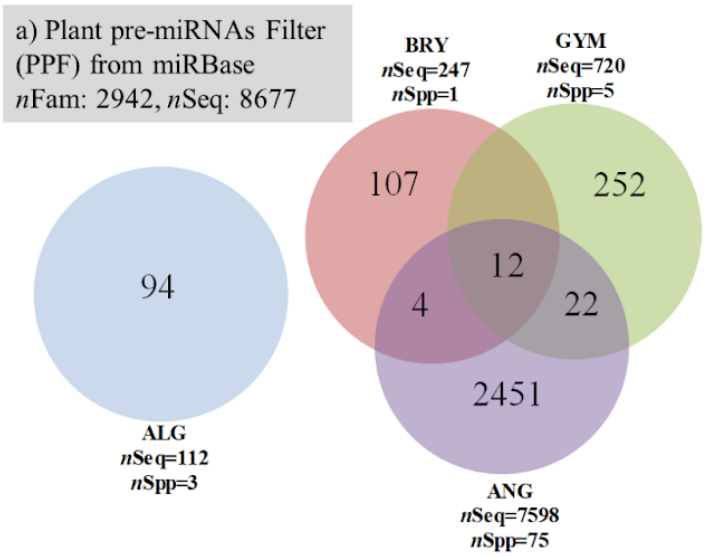
Results and Discussion

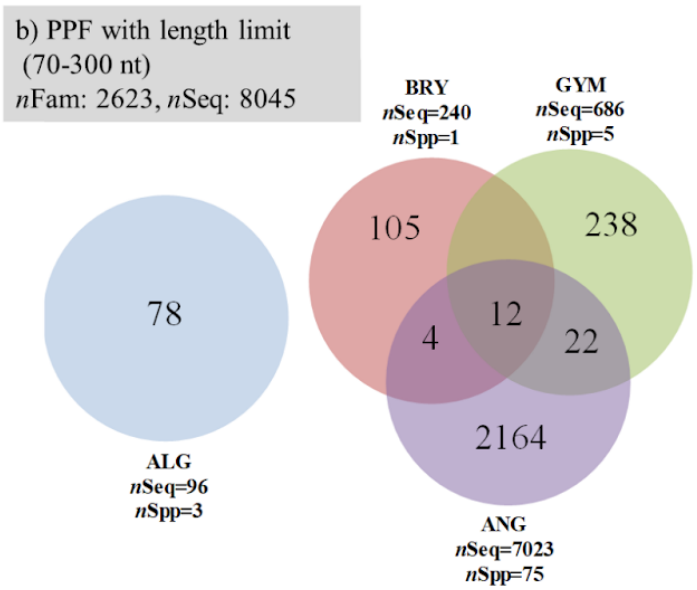
The pipeline began with a hairpin file containing 38589 pre-miRNAs of 14080 families and 265 species from miRBase v.22.1. Most pre-miRNAs belong to the animal kingdom (77%), and very few were vírus (1%). Five pre-miRNAs from the genus *Phytophthora* (plant pathogen fungus) were excluded. They were recorded as plant pre-miRNAs, as the MIR8788 family. This observation was shared with miRBase (Supplementary Table S1). The pipeline mined 8677 (22%) plant pre-miRNAs from 2942 (21%) miRNA families (Figure 1a). This output dataset included pre-miRNAs over 70 to 300 nt from 84 species, 35 plant taxonomic families (PTF), and four clades. These pre-miRNAs were from algae (112 pre-miRNAs, on three species, from three PTF, and 94 pre-miRNA families), bryophytes (247 pre-miRNAs, on one species, from one PTF, and 123 pre-miRNA families), gymnosperms (720 pre-miRNAs, on five species, from three PTF, and 286 pre-miRNA families), and angiosperms (7598 pre-miRNAs, on 75 species, from 28 PTF, and 2489 pre-miRNA families - Figure 2a - Supplementary Table S1).

The algae clade did not share pre-miRNAs with the other three (Figure 2). On the other hand, 38 conserved pre-miRNA families happened in common among the three land clades (Figure 2a).

Twelve pre-miRNAs were common among BRY, GYM, and ANG; four between BRY and ANG; and 22 between GYM and ANG. There was no pre-miRNA recorded between BRY and GYM. Adopting 70 to 300 nt filter for pre-miRNAs (Figure 1a) The exclusivity to their respective clades is evident through identifying 78 unique miRNA families in algae. The three land plant clades shared 12 miRNA families (Figure 2b), confirming some miRNAs' conservation over the last few hundred million years. Angiosperms shared a few exclusive pre-miRNAs families with bryophytes (n=4) and gymnosperms (n=22 - Figure 2b). Conversely, bryophytes and gymnosperms did not have any miRNA in common. This could be explained by the lower number of bryophytes and gymnosperms genomes sequenced and studied so far (16 and 20 species, respectively) related to angiosperm (n=1020 species from 127 families - https://www.plabipd.de/portal/web/guest/sequenced-plant-genomes - accessed on 15.3.2023). Even in angiosperms, only 7% of sequenced genomes have been studied for miRNAs (74 species in 1020 sequenced). It considers the high and exclusive diversity found on angiosperms that are inexistent on the other land plant clades.

Taking in account only pre-miRNA sequences with 70 to 300 nt (Axtel and Meyers, 2018), which resulted in 8045 from 8677 pre-miRNAs, and 2623 from 2942 pre-miRNA families, decreasing by 7% and 10%, respectively (Supplementary Table S3). The small sequences had 70 nt (n=24), and the largest ones had 300 nt (n=8). The pre-miRNA average was 138 nt; the standard deviation was 49, and the 35% coefficient of variation. These pre-miRNAs comprised 84 species from 35 taxonomic plant families: algae (96 pre-miRNAs belong to 78 pre-miRNA families from three species), bryophytes (240 pre-miRNAs belong to 121 pre-miRNA families from one species), gymnosperms (686 pre-miRNAs belong to 272 pre-miRNA families from five species), and angiosperms (7023 pre-miRNAs belong to 2202 pre-miRNA families from 75 species - Figure 2b - Supplementary Table S2).


a) Plant pre-miRNAs Filter (PPF) from miRBase
nFam: 2942, nSeq: 8677
BRY
nSeq=247
nSpp=1
GYM
nSeq=720
nSpp=5
107
252
12
4
22
94
2451
ALG
nSeq=112
nSpp=3
ANG
nSeq=7598
nSpp=75


b) PPF with length limit (70-300 nt)
nFam: 2623, nSeq: 8045
BRY
nSeq=240
nSpp=1
GYM
nSeq=686
nSpp=5
105
238
12
4
22
78
2164
ALG
nSeq=96
nSpp=3
ANG
nSeq=7023
nSpp=75

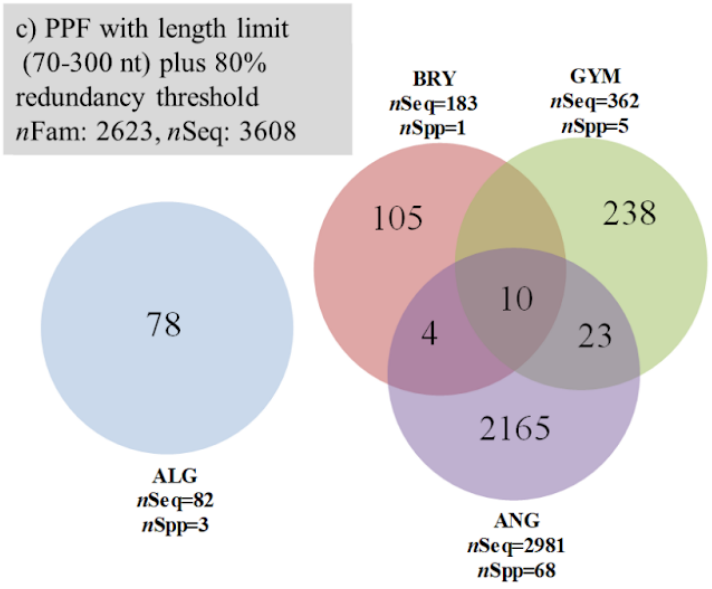

c) PPF with length limit (70-300 nt) plus 80% redundancy threshold
nFam: 2623, nSeq: 3608
BRY
nSeq=183
nSpp=1
GYM
nSeq=362
nSpp=5
105
238
10
4
23
78
2165
ALG
nSeq=82
nSpp=3
ANG
nSeq=2981
nSpp=68

Figure 2. Venn diagrams illustrate the shared pre-miRNA families, sequences, and taxonomic species among plant clades (algae - ALG, bryophytes - BRY, gymnosperms - GYM, and angiosperms - ANG). The dataset used for this analysis included: a) Plant pre-miRNAs filter (PPF) from miRBase, distribution of families. b) PPF with a length limit from 70 to 300 nt. c) PPF with a length limit of 70 - 300 nt, plus 80% redundancy threshold. *n*Fam: Number of pre-miRNA families, *n*Seq: number of pre-miRNA sequences, *n*Spp: number of botanic species.

The most conserved family among the three land clades was the miR169 family, with the highest number of pre-miRNAs recorded (*n*=390). It regulates flowering timing, photosynthesis, protein transport, and transcription factors. miR156 had the second highest number of pre-miRNAs (*n*=362) and is associated with plant development, protein regulation, and transcription factors. miR166 comprised 288 miRNAs implicated in leaf vascular development, transport, and transcription factors. miR171 was the 4th with *n*=263 miRNAs, and it is associated with the growth, metabolism, protein regulation, and transcription factors (Zhou et al. 2010; Djami-Tchatchou et al. 2017; Yousuf et al. 2021 - Supplementary Table S3).

In the next step, the redundancy was removed. The skipRedundant tool suggested a value of percentage identity redundancy of 95% (Lamprecht et al. 2011). However, we performed the redundancy threshold tests from 95 to 70% (interval of 5%). The difference among the percentual redundancy threshold tests using the color alignment was classified by Deeplearning4j. It was statistically significant for colored (BAD). BAD alignment category was the best representative dataset, which was the less redundant. ANOVA, which makes the pairwise comparison among redundancy thresholds, showed no significance among 95 to 85% (p-value=0.07282, $p>0.05$). However, it was significant among 95 to 80% (p-value=0.02884, $p<0.05$ - Figure 3).

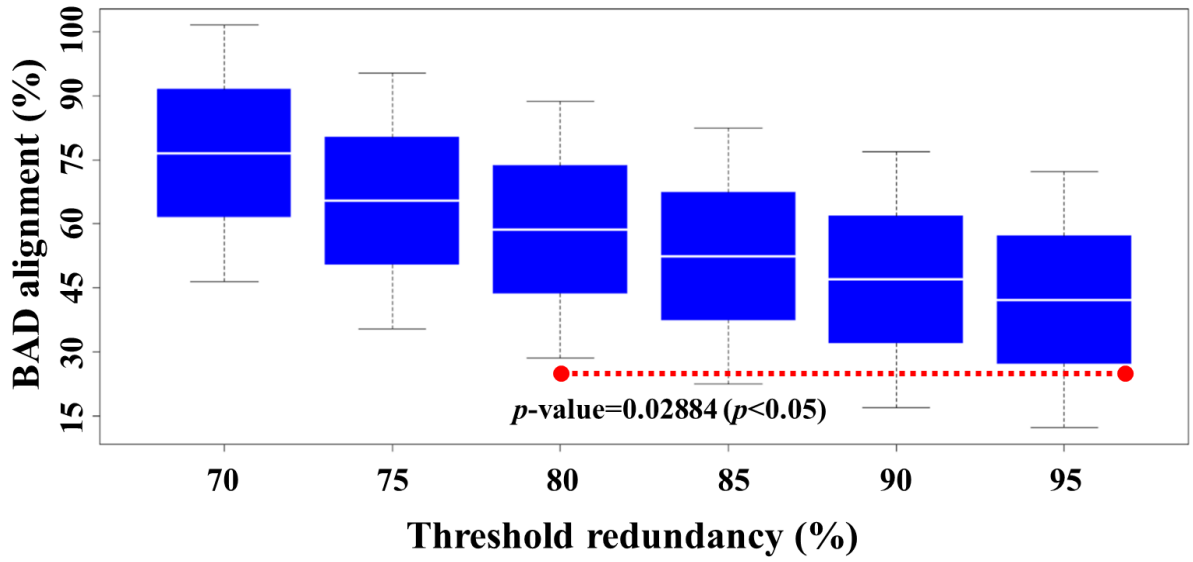


Figure 3. A box plot analysis assessed the statistical significance of the F-measure percentage in color alignment. The results indicated that the colored categories identified as BAD included various blue color tones (hexadecimal colors: #9b92ff, #b4ffb4, #beffbe, #c8ffc8). The red line indicates a noteworthy distinction between the dataset's 80% redundancy (p-value=0.02884) and the range from 95 to 85%.

With more BAD alignment, the 80% redundancy threshold accounted for 3725 pre-miRNAs from 2623 miRNA families and 77 species from 34 taxonomic plant families (Supplementary Table S4). Phylogenetically, these sequences were from algae (82 pre-miRNAs, on three species, from 78 pre-miRNA families), bryophytes (183 pre-miRNAs, on one species, from 119 pre-miRNA families), gymnosperms (362 pre-miRNA, on five species, from 271 pre-miRNA families), and angiosperms (2981 pre-miRNAs, on 68 species, from 2202 pre-miRNA families - Figure 2c). The number of pre-miRNAs decreased substantially by 55% at the dataset's 80% redundancy threshold, compared to the filter that removed sequences' white length from 70 to 300 nt from the 2623 families.

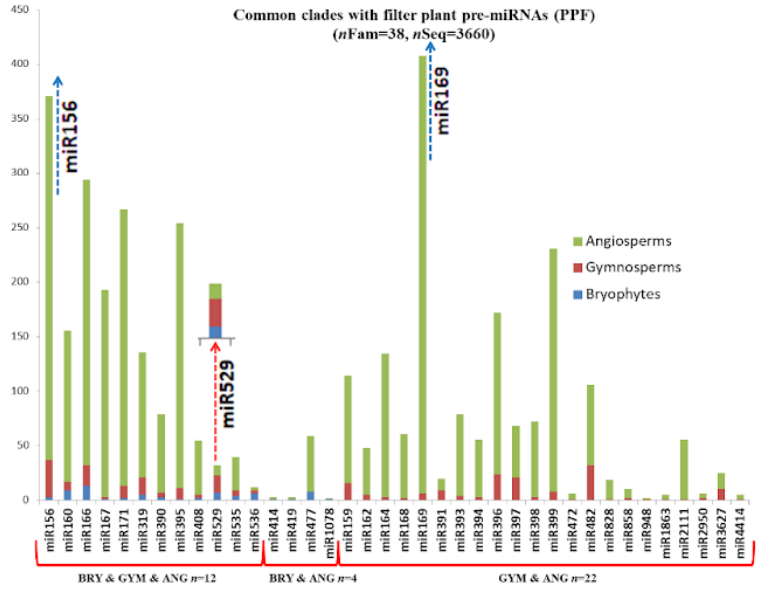
The pre-miRNA families were reduced by 25% after the filter for a length limit from 70 to 300 nt and kept after the 80% redundancy threshold (Figure 3). From 2942 pre-miRNA families from BRY (*n*=247), GYM (*n*=720), and ANG (*n*=7598), 38 happened in common in the three plant clades. Among the 2623 pre-miRNA families, only 10 also occur with angiosperms, gymnosperms, and bryophytes. In the redistribution among clades with an 80% redundancy threshold, only two families originally represented by all three clades became represented by one gymnosperm and another by the angiosperm (Figure 3b, 3c). This result

demonstrates that although an 80% redundancy threshold was applied and the number of sequences was reduced, there was no significant loss of pre-miRNA representatives among the clades.

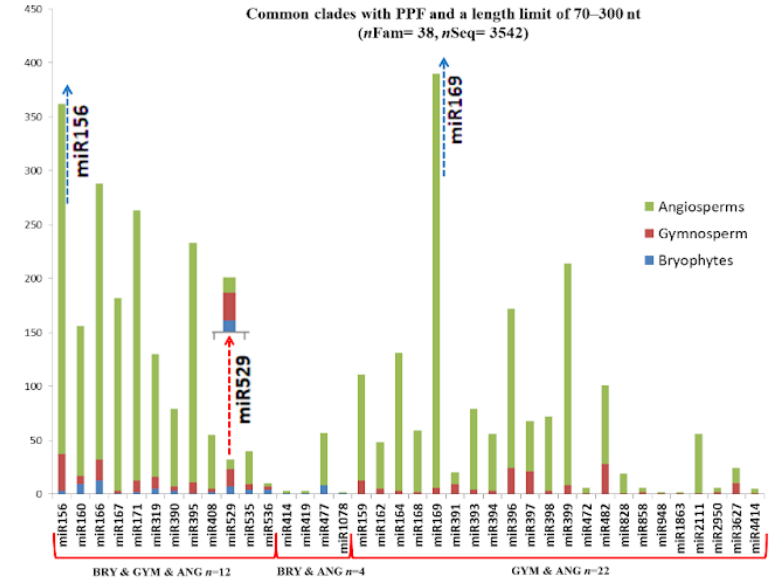
miRBase pre-miRNAs filter from 70 to 300 nt, lightly reduced pre-miRNA families (from 3660 to 3542) among the three plant clades (Figure 4a, 4b). Within angiosperms, the decrease of 104 pre-miRNAs (in families: miR156, miR164, miR166, miR167, miR168, miR169, miR171, miR319, miR395, miR399, miR477, miR182, miR858, and miR1863, miR3627). In bryophytes, two pre-miRNAs were filtered (miR536), and in gymnosperms, 12 pre-miRNAs (in families: miR482, miR159 - Figure 4a and 4b). The family with the highest number of pre-miRNAs is miR169 (*n*=390), followed by miR156 (*n*=362). The predominantly represented clade is that of angiosperms, except in the case of the miR529 family, where pre-miRNAs belonging to the gymnosperms clade predominate (Figure 4b).

There was a strong decrease in pre-miRNA sequences using an 80% redundancy threshold (from 3542 to 585). Angiosperms reduced 2735 pre-miRNAs, bryophytes were reduced by 31 pre-miRNAs, and gymnosperms were reduced by 191 pre-miRNAs (Figure 4b and 4c). The family that had the highest number of reduced pre-miRNAs was miR169 (Angiosperms - *n*=336, Gymnosperms - *n*=4), followed by miR156 (Angiosperms - *n*=292, Gymnosperms - *n*=30, Bryophytes - *n*=2- Figure 4b and 4c). Despite being the family with the highest reduction in pre-miRNAs, miR169 remains the family with the greatest number of pre-miRNAs (n=50), followed by miR166 (*n*=41) and miR156 (*n*=38). miR529 continues to be predominantly represented by gymnosperm pre-miRNAs. It is noteworthy that miR169 belongs to the largest plant family in miRBase. Furthermore, miR156 has 396 identical copies of the mature miRNAs (20 bp) in miRBase. This result indicates that some annotations of pre-miRNAs and mature miRNAs are duplicates, and the 80% redundancy threshold can help distinguish between them.

a)


Common clades with filter plant pre-miRNAs (PPF)
(nFam=38, nSeq=3660)
miR156
miR169
miR529
Angiosperms
Gymnosperms
Bryophytes
miR156 miR160 miR166 miR167 miR171 miR319 miR390 miR395 miR408 miR529 miR535 miR536 miR414 miR419 miR477 miR1078 miR159 miR162 miR164 miR168 miR169 miR391 miR393 miR394 miR396 miR397 miR398 miR399 miR472 miR482 miR828 miR858 miR948 miR1863 miR2111 miR2950 miR3627 miR4414
BRY & GYM & ANG n=12
BRY & ANG n=4
GYM & ANG n=22

b)


Common clades with PPF and a length limit of 70–300 nt
(nFam= 38, nSeq= 3542)
miR156
miR169
miR529
Angiosperms
Gymnosperm
Bryophytes
miR156 miR160 miR166 miR167 miR171 miR319 miR390 miR395 miR408 miR529 miR535 miR536 miR414 miR419 miR477 miR1078 miR159 miR162 miR164 miR168 miR169 miR391 miR393 miR394 miR396 miR397 miR398 miR399 miR472 miR482 miR828 miR858 miR948 miR1863 miR2111 miR2950 miR3627 miR4414
BRY & GYM & ANG n=12
BRY & ANG n=4
GYM & ANG n=22

c)

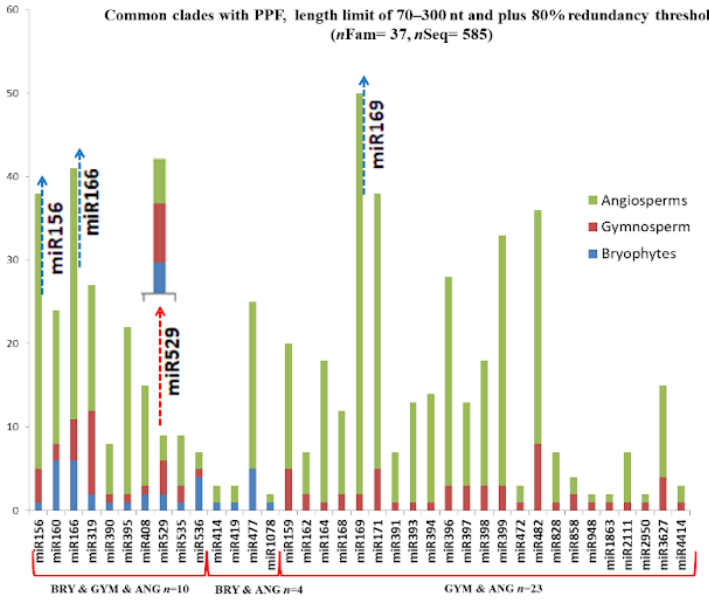

Common clades with PPF, length limit of 70–300 nt and plus 80% redundancy threshold
(nFam= 37, nSeq= 585)
miR156
miR166
miR169
miR529
Angiosperms
Gymnosperm
Bryophytes
miR156 miR160 miR166 miR319 miR390 miR395 miR408 miR529 miR535 miR536 miR414 miR419 miR477 miR1078 miR159 miR162 miR164 miR168 miR169 miR171 miR391 miR393 miR394 miR396 miR397 miR398 miR399 miR472 miR482 miR828 miR858 miR948 miR1863 miR2111 miR2950 miR3627 miR4414
BRY & GYM & ANG n=10
BRY & ANG n=4
GYM & ANG n=23

Figure 4. Plant pre-miRNA sequences distributed among the plant clades (BRY, GYM, and angiosperms ANG) were analyzed using three filters. a) Plant pre-miRNAs filter (PPF) from miRBase, distribution of families b) PPF from 70 to 300 nt. c) PPF from 70 to 300 nt, plus 80% redundancy threshold. *n*Fam: number of families, *n*Seq: number of pre-miRNA sequences. The dotted blue arrow indicates the families with the highest pre-miRNA sequences counted. The dotted red arrow indicates the pre-miRNA family with the highest sequences on a non-angiosperm clade.

Angiosperms accounted for most of the species dataset's pre-miRNAs (78%) on an 80% redundancy threshold. The most abundant species were soybean [*Glycine max* (L.) Merr - *n*=340], followed by rice (*O. sativa*, *n*=315), Norway/European spruce [*Picea abies* (L.) H. Karst - *n*=305], medicago (*Medicago truncatula* G. - *n*=248), algae [*Physcomitrella patens* (H.) Bruch & Schimp. - *n*=183], brachypodium [*Brachypodium distachyon* (L.) P.Beauv - *n*=175], cotton (*Gossypium raimondii* U., *n*=172), and arabidopsis (*A. lyrata* L. and *A. thaliana* - *n*=243).

These species are models for evolutionary studies, such as algae and Norway/European spruce (gymnosperm - Bernhardsson et al. 2019). Medicago and soybean are legume models related to nitrogen fixation (Tang et al. 2014). Soybean is one of the most economically important crops due to its major plant source of proteins and oils. Cereals have some model species crucial for world food security, like rice (C3 photosynthetic metabolism), the first monocot and second plant genome sequenced at the beginning of this century (Goff et al. 2002, Yu et al. 2002). Brachypodium is the genome reference to the cereal, grass, and monocot models for tempered cereals (Garvin *et al*. 2008). Besides being the first multicellular eukaryote and the first plant genome sequenced (The Arabidopsis Genome Initiative, 2000), arabidopsis is the main model plant for developmental genetics and miRNAs studies (Fahlgren et al. 2010).

The most representative taxonomic family in the pre-miRNAs was Fabaceae (*n*=783), followed by Poaceae (*n*=706). Fabaceae (ex Leguminosae family) is the third family with more species in angiosperms (*n*=22342, https://www.catalogueoflife.org/ accessed on January 31, 2023). It is the family with the largest tree species (Beech et al. 2017), and most of its members are nitrogen-fixing (Wink et al. 2013). Although edible pulses have health benefits on the human diet, they are drastically less used than cereals. Most of their edible legumes are scientifically considered orphans or neglected (Foyer et al. 2016). They have a vast under-used role in food security and the environment.

On the other hand, Poaceae (ex Gramineae family) accounts for 12081 species (https://www.catalogueoflife.org/ accessed on January 31, 2023). Grasslands dominate the vegetation worldwide (Edwards et al. 2010), and some species are a model for breeding, evolutive, and taxonomic studies (Hodkinson 2018; Michael and Jackson, 2013). *Homo* evolved in savannas (Strömberg et al. 2022), and cereal domestication marks the birth of civilizations (Morrison, 2016). Green revolution focused on its three most produced cereals (maize, rice, and wheat - Liu et al. 2020). Few Poaceae members are the primary source of food, feed, and biofuel worldwide. Cereals account for half the harvest area planted (FAO Yearbook, 2021).

Rice validation

The rice genome comprises 8470 pre-miRNAs homologous to 36 families and 1469 identical copies. Of these pre-miRNA families, 19 were newly discovered and have not been previously annotated in the miRBase for rice. They are involved in model plants' growth, development, and responses to biotic and abiotic stresses (Table S5). After the alignment with the mature miRNAs, 7279 pre-miRNAs from 13 families and 1460 copies were identical with PmiR-Select (Figure 3). PmiR-Select identified more pre-miRNAs than PmiREN (8470 versus 699 pre-miRNAs - Guo et al. 2020), using the same rice genome (*O. sativa*). PmiREN is a homology-based approach with a secondary structure. It modifies the family assignment to correct it, which may open a path for phylogenetic studies to identify new families (Guo et al. 2020). Most available computational tools use mature miRNAs for identification, half take pre-miRNAs (Morgado and Johannes 2019). Here, we used pre-miRNAs for identification and mature miRNAs to check. We do not pick sequences from mature miRNAs because of the low success of mature miRNAs to match in pre-miRNA sequences due to flawed annotation. No end characteristics indicate the location of mature miRNA in pre-miRNA, as the start and the stop codons at translation or the canonical dinucleotide GT and AG, which flanked introns (Sun et al. 2014; Morgado and Johannes 2019, Frey and Pucker, 2020).

Conclusions and future work

This study described a pipeline to data mining by criteria indicating sequence length limit from 70 to 300 nt for plant pre-miRNAs and from 20 to 24 nt for plant mature miRNAs. We estimated an 80% redundancy threshold for dataset analysis of non-redundant pre-miRNAs, and there was a high reduction in the number of sequences compared to the plant pre-miRNAs dataset (55%). However, few (2) family redistributions among the three clades

(BRY, GYM, and ANG). Nevertheless, annotations without mining can become just a computational artifact in the era of big data.

PmiR-Select a homology-based approach and CMs for identified pre-miRNAs. This technique accommodates the input genomes of diverse organisms, thus enhancing its applicability. Output with datasets of identical copies, non-identical, and threshold redundancies of 95% to 70% (5% intervals). Although our suggestion for analysis was the 80% dataset, this functionality of the output files offers greater flexibility and customization in the analysis, allowing users to evaluate the results according to the specific characteristics of their study. Furthermore, PmiR-Select comes with dependencies, source code, and user instructions at https://github.com/DeborahBambil/PmiRSelect. We are committed to updating PmiR-Select according to miRBase versions. Moreover, the second public repository of miRNAs, the RNA families (Rfam), has a project to collaborate with miRBase to integrate miRNA families (https://rfam.xfam.org/microrna).

Abbreviations

CMs	Covariance models
DCL1	Dicer-like 1
miRNAs	microRNAs
ncRNA	non-coding RNAs
nt	nucleotides
pre-miRNAs	precursors miRNAs
pri-miRNAs	primary miRNAs

Availability and requirements

Project name: PmiR-Select.

Project home page: https://github.com/DeborahBambil/PmiRSelect

Operating system(s): Linux.

Programming language: Python and Shell.

Other requirements: Python 3, infernal, emboss, NCBI-blast+.

Any restrictions to use by non-academics: None.

Supplementary Information

Table S1. Plant pre-miRNAs filtered: 8677 of 2942 families, from 84 plant species, 35 taxonomic families, and four clades.

Table S2. Plant pre-miRNAs filtered by length (from 70 to 300 nt): 8045 of 2623 families, from 84 plant species, 35 taxonomic families, and four clades.

Table S3. Plant pre-miRNAs families list and their functions from pre-miRNAs filtered by length on four clades.

Table S4. Plant pre-miRNAs filtered by length (70 to 300 nt) plus 80% threshold redundancy: 3608 of 2623 miRNA families from 77 plant species, 34 taxonomic families, and four clades.

Table S5. Novel pre-miRNAs identified in rice and their functions using covariance models at PmiR-Select.

Acknowledgments

Not applicable.

Author contributions

All authors of this manuscript have directly participated in the execution of the study. Conceived and designed the experiments: DB and LFAF performed the experiments. DB and LFAF analyzed the data: DB and LFAF wrote the paper. All authors read and approved the final manuscript.

Funding

Brazilian agency CAPES for financial support scholarship to grant for D.B. process number 88887.635040/2021-00.

Availability of data and materials

Declarations

Ethics approval and consent to participate.

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

Axtell MJ, Meyers BC (2018) Revisiting criteria for plant microRNA annotation in the era of big data. The Plant Cell 30(2), 272-284. doi: 10.1105/tpc.17.00851

Bambil D, Pistori H, Bao F, *et al*. (2020) Plant species identification using color learning resources, shape, texture, through machine learning and artificial neural networks. Environment Systems and Decisions 40(4), 480-484. doi: 10.1007/s10669-020-09769-w

Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136:215–233. doi: 10.1016/j.cell.2009.01.002

Beech E, Rivers M, Oldfield S, Smith PP (2017) GlobalTreeSearch: The first complete global database of tree species and country distributions. Journal of Sustainable Forestry 36(5), 454–489. doi: 10.1080/10549811.2017.1310049

Bernhardsson C, Vidalis A, Wang X, et al. (2019) An ultra-dense haploid genetic map for evaluating the highly fragmented genome assembly of Norway spruce (*Picea abies*). G3: Genes, Genomes, Genetics 9(5), 1623-1632. doi: 10.1534/g3.118.200840

Bhogireddy S, Mangrauthia SK, Kumar R, et al. (2021) Regulatory non-coding RNAs: A new frontier in regulation of plant biology. Functional & Integrative Genomics 21, 313-330. doi: 10.1007/s10142-021-00787-8

Boatwright JL, Sapkota S, Jin H, et al. (2022) Sorghum association panel whole-genome sequencing establishes cornerstone resource for dissecting genomic diversity. The Plant Journal 111(3), 888-904. doi: 10.1111/tpj.15853

Budak H and Akpinar BA (2015) Plant miRNAs: biogenesis, organization and origins. Functional and Integrative Genomics v. 15, 523–531. https://doi.org/10.1007/s10142-015-0451-2

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282-5396. doi: 10.1126/science.282.5396.2012

Cui J, You C, Chen X (2017) The evolution of microRNAs in plants. Current Opinion in Plant Biology 35, 61-67. doi: 10.1016/j.pbi.2016.11.006

Djami-Tchatchou AT, Sanan-Mishra N, Ntushelo K, Dubery IA (2017) Functional roles of microRNAs in agronomically important plants potential as targets for crop improvement and protection. Frontiers in Plant Science 8, 378. doi: 10.3389/fpls.2017.00378

Dong Y, Duan S, Xia Q, et al. (2023) Dual domestications and origin of traits in grapevine evolution. Science 379(6635), 892-901. doi: 10.1126/science.add8655

Edwards EJ, Osborn CP, et al. (2010) The origins of C4 grasslands: integrating evolutionary and ecosystem science. Science 328(5978), 587–591. doi: 10.1126/science.1177216

Fahlgren N, Jogdeo S, Kasschau KD, et al. (2010) microRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. The Plant Cell 22(4), 1074-1089. doi: 10.1105/tpc.110.073999

Foyer CH, Lam HM, Nguyen HT, *et al*. (2016) Neglecting legumes has compromised human health and sustainable food production. Nature Plants 2(8), 16112. doi: 10.1038/nplants.2016.112

Garvin DF, Gu YQ, Hasterok R, et al. (2008) Development of genetic and genomic research resources for *Brachypodium distachyon*, a new model system for grass crop research. Crop Science 48, S-69. doi: 10.2135/cropsci2007.06.0332tpg

Greene CS, Tan J, Ung M, et al. (2014) Big data bioinformatics. Journal of Cellular Physiology 229(12), 1896-1900. doi: 10.1002/jcp.24662

Griffiths-Jones S (2004) The microRNA registry. Nucleic Acids Research 32(S1), D109-D111. doi: 10.1093/nar/gkh023

Griffiths-Jones, S (2010) miRBase: microRNA sequences and annotation. Current protocols in bioinformatics 29(1), 12-9. doi: 10.1002/0471250953.bi1209s29

Goff SA, Ricke D, Lan TH, et al., (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296(5565), 92-100. doi: 10.1126/science.1068275

Grüning BA, Fallmann J, Yusuf D, et al. (2017) The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. Nucleic Acids Research 45(W1), W560-W566. doi: 10.1093/nar/gkx409

Guo Z, Kuang Z, Wang Y, et al. (2020) PmiREN: a comprehensive encyclopedia of plant miRNAs. Nucleic Acids Research. 48(D1), D1114-D1121. doi: 10.1093/nar/gkz894

Hajieghrari B, Farrokhi N (2022) Plant RNA-mediated gene regulatory network. Genomics 114(1), 409-442. doi: 10.1016/j.ygeno.2021.12.020

Hodkinson TR (2018) Evolution and taxonomy of the grasses (Poaceae): a model family for the study of species-rich groups. Annual Plant Reviews Online 255-294. doi: 10.1002/9781119312994.apr0622

Kozomara A, Birgaoanu M, Griffiths-Jones S (2019) miRBase: from microRNA sequences to function. Nucleic Acids Research 47(D1), D155-D162. doi: 10.1093/nar/gky1141

Lamprecht AL, Naujokat S, Margaria T, Steffen B (2011) Semantics-based composition of EMBOSS services. Journal of Biomedical Semantics 2(1), 1-21. doi: 10.1186/2041-1480-2-S1-S5

Lang S, Bravo-Marquez F, Beckham C, et al. (2019) Wekadeeplearning4j: A deep learning package for weka based on deeplearnincg4j. Knowledge-Based Systems 178, 48-50. doi: 10.1016/j.knosys.2019.04.013

Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell 75(5), 843-854. doi: 10.1016/0092-8674(93)90529-Y

Lin X, Kaul S, Rounsley S, et al. (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402(6763), 761-768. doi: 10.1038/45471

Ma X, Denyer T, Javelle M, et al. (2021) Genome-wide analysis of plant miRNA action clarifies levels of regulatory dynamics across developmental contexts. Genome Research 31(5), 811-822. doi: 10.1101/gr.270918.120.

Mayer K, Schüller C, Wambutt R, et al. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana.* Nature 402(6763), 769-777. doi: 10.1038/47134

Michael TP, Jackson S (2013) The first 50 plant genomes. The Plant Genome 6(2) doi: 10.3835/plantgenome2013.03.0001in

Millar AA, Lohe A, Wong G (2019) Biology and function of miR159 in plants. Plants 8(8), 255. doi: 10.3390/plants8080255

Morgado L, Johannes F (2019) Computational tools for plant small RNA detection and categorization. Briefings in Bioinformatics 20(4), 1181-1192. doi: 10.1093/bib/bbx136

Nawrocki EP. and Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22), 2933-2935. doi: 10.1093/bioinformatics/btt509

Powers DM (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. ArXiv Preprint 2010-16061. doi: arxiv.org/abs/2010.16061

R core team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.

Raza A, Charagh S, Karikari B, et al. (2023) miRNAs for crop improvement. Plant Physiology and Biochemistry 201, doi: 107857.10.1016/j.plaphy.2023.107857

Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, and Bartel DP (2002) MicroRNAs in plants. Genes and Development 16(13), 1616-1626. doi: 10.1101/gad.1004402

Sikic K, Carugo O (2010) Protein sequence redundancy reduction: comparison of various methods. Bioinformation 5(6), 234. doi: 10.6026/97320630005234

Strömberg CAE, Staver AC (2022) The history and challenge of grassy biomes. Science 377(6606), 592–593. doi: 10.1126/science.add1347

Sun J, Zhou M, Mao Z, Li C (2012) Characterization and evolution of microRNA genes derived from repetitive elements and duplication events in plants. PloS One 7(4), e34092. doi: 10.1371/journal.pone.0034092

Sun X, Zhang Y, Zhu X, et al. (2014) Advances in identification and validation of plant microRNAs and their target genes. Physiologia Plantarum 152(2), 203-218. doi:10.1111/ppl.12191

Tang H, Krishnakumar V, Bidwell S, et al. (2014) An improved genome release (version Mt4. 0) for the model legume *Medicago truncatula*. BMC Genomics 15(1), 1-14. doi: 10.1186/1471-2164-15-312

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814), 796-815. doi: 10.1038/35048692

Tommaso PD, Moretti S, Xenarios I, et al. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Research 39(suppl_2), W13-W17. doi: 10.1093/nar/gkr245

Wink M (2013) Evolution of secondary metabolites in legumes (Fabaceae). South African Journal of Botany 89, 164-175. doi: 10.1016/j.sajb.2013.06.006

World Food and Agriculture (2021) Statistical Yearbook 2021. FAO. doi: https://doi.org/10.4060/cb4477en

Yousuf PY, Shabir PA, Hakeem KR (2021) miRNAomic approach to plant nitrogen starvation. International Journal of Genomics 2314-436X. doi: 10.1155/2021/8560323

Yu J, Hu S, Wang J, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296(5565), 79-92. doi: 10.1126/science.1068037

Yu T, Xu N, Haque N, et al. (2020) Popular computational tools used for miRNA prediction and their future development prospects. Interdisciplinary Sciences: Computational Life Sciences 12, 395-413. doi: 10.1007/s12539-020-00387-

Waititu JK, Zhang C, Liu J, Wang H (2020) Plant non-coding RNAs: Origin, biogenesis, mode of action and their roles in abiotic stress. International Journal of Molecular Sciences 21(21), 8401. doi: 10.3390/ijms21218401

Xu T, Su N, Liu L, et al. (2018) miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence, and family information in different versions of miRBase. BMC Bioinformatics 19(19), 179-188. doi: 10.1186/s12859-018-2531-5

Zhang B, Pan X, Cannon CH, et al. (2006) Conservation and divergence of plant microRNA genes. The Plant Journal 46(2), 243-259. doi: 10.1111/j.1365-313X.2006.02697.x

Zhang Q, Yang LT, Chen Z, et al. (2018) A survey on deep learning for big data. Information Fusion 42, 146-157. doi: 10.1016/j.inffus.2017.10.006

Zhou M, Gu L, Li P, et al. (2010) Degradome sequencing reveals endogenous small RNA targets in rice (*Oryza sativa* L. ssp. indica). Frontiers of Biology in China 5: 67-9. doi: 10.1007/s11515-010-0007-8

2.4 Conclusão

Neste estudo, foi descrito o pipeline para mineração de dados com critérios atualizados para o tamanho dos pre-miRNAs (70-300 nt) e dos miRNAs maduros de plantas (20-24 nt). Houve uma redução de 7 e 11% dos pre-miRNAs (8677 para 8045) e suas famílias (2942 para 2623). A ferramenta PmiR-Select (https:/github.com/DeborahBambil/PmiRSelect) utilizou esse conjunto de dados minerados para construir os modelos de covariância, que foram implantados na identificação de novos pre-miRNAs não anotados em genomas de plantas.

Após a mineração quanto ao tamanho dos pre-miRNAs, foi analisado os limites de redundância de 95 a 70% (intervalos de 5%). Essa análise mostrou que no limite de 80 a 70% de redundância existia uma diferença significativa (p<0.05) comparado aos limites de 95 a 85%. Adotou-se o limite de 80% por se obter mais pre-miRNAs, quando comparado aos limites de 75 a 70%. No conjunto de dados com limite de 80% houve uma redução drástica de 55% dos pre-miRNAs (8045 para 3608).

Entre as 38 famílias de pre-miRNAs em comum nos três clados (BRY, GYM e ANG) ocorreram apenas duas redistribuições das famílias, quando comparado aos conjuntos de pre-miRNAs com sequências 70 a 300 nt. Os pre-miRNAs de ANG foram predominantes em todas as famílias em comum; exceto o miR529, onde prevaleceu as GYM. Os miR169 e miR156 tiveram as maiores ocorrências entre as 38 famílias em comum nos três clados. Antes e após a mineração e após o filtro do limite de redundância de 80% não houve nenhuma família em comum entre as algas com os outros três clados, e entre BRY e GYM.

Embora a sugestão para a análise do conjunto de dados de pre-miRNAs seja com o limite de redundância de 80%, essa funcionalidade dos arquivos de saída oferece maior flexibilidade e personalização na análise, permitindo a avaliação do resultado de acordo com as características do estudo. Por exemplo, uma pré-seleção de exemplares que serão destinados para validação experimental de bancada. Assim na análise do genoma do arroz, foram identificados 8470 pre-miRNAs de 36 familias, sendo 1469 pre-miRNAs idênticos. Com o limite de redundância de 80% ficaram apenas 272 pre-miRNAs das 36 familias.

As vantagens da PmiR-Select são: i) fácil uso, com apenas um comando (bash run.bin) é possível começar a busca por pre-miRNAs, ii) não restringe o tamanho das sequências de entrada, desde que esteja em fasta, iii) os diretórios de saída são exportados com as cópias distinguidas em idênticas e não idênticas, e por limites de 95 a 70% (intervalos de 5%). As desvantagens da PmiR-Select são: i) não está disponível para operação online, ii) não faz a identificação de miRNAs maduros.

Table S1. Plant pre-miRNAs filtered: 8677 of 2942 families, from 84 plant species, 35 taxonomic families, and four clades. Here is the first page.

Table S2. Plant pre-miRNAs filtered by length (from 70 to 300 nt): 8045 of 2623 families, from 84 plant species, 35 taxonomic families, and four clades. Here is the first page.

Table S3. Plant pre-miRNAs families list and their functions from pre-miRNAs filtered by length on four clades. Here is the first page.

Table S4. Plant pre-miRNAs filtered by length (70 to 300 nt) plus 80% threshold redundancy: 3608 of 2623 miRNA families from 77 plant species, 34 taxonomic families, and four clades. Here is the first page.

Table S5. Novel pre-miRNAs identified in rice and their functions using covariance models at PmiR-Select. Here is the first page.

In this appendix, only the first page of the tables has been included. The remaining pages are available through the link or QR code below.

Link to access the tables:

https://1drv.ms/f/s!ApXNAj3cFXgAg_MGyD4VAIz7hGKeNw?e=ghTJer

QrCode to access the tables:

Table S1. Plant pre-miRNAs filtered: 8677 of 2942 families, from 84 plant species, 35 taxonomic families, and four clades. Here is the first page.

Plant miRNAs	Id miRBase	Species	Families	Clades
>ath-MIR156a CAAGAGAAACGCAAAGAAACUGACAGAAGAGAGUGAGCACACAAAGGCAAUUUGCAUAUC AUUGCACUUGCUUCUCUUGCGUGCUCACUGCUCUUUCUGUCAGAUUCCGGUGCUGAUCUC UUU	MI0000178	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156b GCUAGAAGAGGGAGAGAUGGUGAUUGAGGAAUGCAACAGAGAAAACUGACAGAAGAGAGU GAGCACAUGCAGGCACUGUUAUGUGUCUAUAACUUUGCGUGUGCGUGCUCACCUCUCUUU CUGUCAGUUGCCUAUCUCUGCCUGCUUGACCUCUCUCUCUCUCUCUCUCUCAAAUUUG GCU	MI0000179	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156c CGCAUAGAAACUGACAGAAGAGAGUGAGCACACAAAGGCACUUUGCAUGUUCGAUGCAUU UGCUUCUCUUGCGUGCUCACUGCUCUAUCUGUCAGAUUCCGGCU	MI0000180	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156d GAUGGGGGAAAAGAAGUUGACAGAAGAGAGUGAGCACACAAAGGGGAAGUUGUAUAAAAG UUUUGUAUAUGGUUGCUUUUGCGUGCUCACUCUCUUUUUGUCAUAACUUCUCCUUCAU	MI0000181	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156e AGGAGGUGACAGAAGAGAGUGAGCACACAUGGUGGUUUCUUGCAUGCUUUUUUGAUUAGG GUUUCAUGCUUGAAGCUAUGUGUGCUUACUCUCUCUCUGUCACCCCU	MI0000182	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156f GAGUGGUGAGGAAUUGAUGGUGACAGAAGAGAGUGAGCACACAUGGUGGCUUUCUUGCAU AUUUGAAGGUUCCAUGCUUGAAGCUAUGUGUGCUCACUCUCUAUCCGUCACCCCCUUCUC UCCCUCUCCCUC	MI0000183	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR157a GUGUUGACAGAAGAUAGAGAGCACAGAUGAUGAGAUACAAUUCGGAGCAUGUUCUUUGCA UCUUACUCCUUUGUGCUCUCUAGCCUUCUGUCAUCACC	MI0000184	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR157b UGGGAGGCAUUGAUAGUGUUGACAGAAGAUAGAGAGCACAGAUGAUAAGAUACAAUUCCU CGCAGCUUCUUUGCAUCUUACUCCUUUGUGCUCUCUAGCCUUCUGUCAUCACCCGUUAUU UGCCAUCACCCA	MI0000185	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR157c AGGUUUGAGAGUGAUGUUGGUUGUUGACAGAAGAUAGAGAGCACUAAGGAUGACAUGCAA GUACAUACAUAUAUAUCAUCACACCGCAUGUGGAUGAUAAAAUAUGUAUAACAAAUUCAA AGAAAGAGAGGGAGAGAAAGAGAGAGAACCUGCAUCUCUACUCUUUUGUGCUCUCUAUAC UUCUGUCACCACCUUUAUCUCUUCUUCUCUCUAACCU	MI0000186	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR157d GUGGAGGGUGAUAGUGUGGUUGCUGACAGAAGAUAGAGAGCACUAAGGAUGCUAUGCAAA ACAGACACAGAUAUGUGUUUCUAAUUGUAUUUCAUACUUUAACCUCAAAGUUGAUAUAAA AAAAGAAAGAAAGAUAGAAGAGCUAGAAGACUAUCUGCAUCUCUAUUCCUAUGUGCUCUC UAUGCUUCUGUCAUCACCUUUCUUUCUCUAUUUCUCUCUAC	MI0000187	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR158a ACACGUCAUCUCUGUGCUUCUUUGUCUACAAUUUUGGAAAAAGUGAUGACGCCAUUGCUC UUUCCCAAAUGUAGACAAAGCAAUACCGUGAUGAUGUCGU	MI0000188	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR159a GUAGAGCUCCUUAAAGUUCAAACAUGAGUUGAGCAGGGUAAAGAAAAGCUGCUAAGCUAU GGAUCCCAUAAGCCCUAAUCCUUGUAAAGUAAAAAAGGAUUUGGUUAUAUGGAUUGCAUA UCUCAGGAGCUUUAACUUGCCCUUUAAUGGCUUUUACUCUUCUUUGGAUUGAAGGGAGCU CUAC	MI0000189	*Arabidopsis thaliana*	Brassicaceae	Angiosperm

Table S2. Plant pre-miRNAs filtered by length (from 70 to 300 nt): 8045 of 2623 families, from 84 plant species, 35 taxonomic families, and four clades. Here is the first page.

Plant miRNAs	Id miRBase	Species	Families	Clades
>ath-MIR156a CAAGAGAAACGCAAAGAAACUGACAGAAGAGAGUGAGCACACAAAGGCAAUUUGCAUAUC AUUGCACUUGCUUCUCUUGCGUGCUCACUGCUCUUUCUGUCAGAUUCCGGUGCUGAUCUC UUU	MI0000178	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156b GCUAGAAGAGGGAGAGAUGGUGAUUGAGGAAUGCAACAGAGAAAACUGACAGAAGAGAGU GAGCACAUGCAGGCACUGUUAUGUGUCUAUAACUUUGCGUGUGCGUGCUCACCUCUCUUU CUGUCAGUUGCCUAUCUCUGCCUGCUUGACCUCUCUCUCUCUCUCUCUCUCAAAUUUG GCU	MI0000179	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156c CGCAUAGAAACUGACAGAAGAGAGUGAGCACACAAAGGCACUUUGCAUGUUCGAUGCAUU UGCUUCUCUUGCGUGCUCACUGCUCUAUCUGUCAGAUUCCGGCU	MI0000180	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156d GAUGGGGGAAAAGAAGUUGACAGAAGAGAGUGAGCACACAAAGGGGAAGUUGUAUAAAAG UUUUGUAUAUGGUUGCUUUUGCGUGCUCACUCUCUUUUUGUCAUAACUUCUCCUUCAU	MI0000181	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156e AGGAGGUGACAGAAGAGAGUGAGCACACAUGGUGGUUUCUUGCAUGCUUUUUUGAUUAGG GUUUCAUGCUUGAAGCUAUGUGUGCUUACUCUCUCUCUGUCACCCCU	MI0000182	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>ath-MIR156f GAGUGGUGAGGAAUUGAUGGUGACAGAAGAGAGUGAGCACACAUGGUGGCUUUCUUGCAU AUUUGAAGGUUCCAUGCUUGAAGCUAUGUGUGCUCACUCUCUAUCCGUCACCCCCUUCUC UCCCUCUCCCUC	MI0000183	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>osa-MIR156a GGAGGGUGACAGAAGAGAGUGAGCACACGUGGUUGUUUCCUUGCAUAAAUGAUGCCUAUG CUUGGAGCUACGCGUGCUCACUUCUCUCUCUGUCACCUCC	MI0000653	*Oryza sativa*	Poaceae	Angiosperm
>osa-MIR156b UUGUCUUGAGAGGGGAAGAGAUCUCUAUGGGUUUUGGAGGUCUGACAGAAGAGAGUGAGC ACACACGGUGCUUUCUUAGCAUGCAAGAGCCAUGCUGGGAGCUGUGCGUGCUCACUCUCU AUCUGUCAGCCGUUCACCAUGCCCAAUAUGAUUAAUCUCCUUCUCUCAGUUGACAG	MI0000654	*Oryza sativa*	Poaceae	Angiosperm
>osa-MIR156c GGAGGAAGAGAGGGGUGAGAGGUGAGGCUGACAGAAGAGAGUGAGCACACAUGGUGACUU UCUUGCAUGCUGAAUGGACUCAUGCUUGAAGCUAUGUGUGCUCACUUCUCUCUCUGUCAG CCAUUUGAUCUCUCUUUCUCUCUUUCUCC	MI0000655	*Oryza sativa*	Poaceae	Angiosperm
>osa-MIR156d GGAGAAGCUCUCAUGAGAUUGACAGAAGAGAGUGAGCACACGGCGUGAUGGCCGGCAUAA AAUCUAUCCCGUCCUCGCCGCGUGCUCACUCCUCUUUCUGUCACCCUCUUUCUCUCAGGG CUCAACUCC	MI0000656	*Oryza sativa*	Poaceae	Angiosperm
>osa-MIR156e GGCGCGAGGUGACAGAAGAGAGUGAGCACACGGCCGGGCGUGACGGCACCGGCGGGCGUG CCGUCGCGGCCGCGUGCUCACUGCUCUUUCUGUCAUCCGGUGCC	MI0000657	*Oryza sativa*	Poaceae	Angiosperm
>osa-MIR156f AGUUGACAGAAGAGAGUGAGCACACAGCGGCCAGACUGCAUCGAUCUAUCAAUCUUCCCU UCGACAGGAUAGCUAGAUAGAAAGAAAGAGAGGCCGUCGGCGGCCAUGGAAGAGAGAGAG AGAGAGAGAUGAAAUGAUGAUGAUGAUACAGCUGCCGCUGCGUGCUCACUUCUCUUUCUG UCAGCU	MI0000658	*Oryza sativa*	Poaceae	Angiosperm
>osa-MIR156g GGCUGACAGAAGAGAGUGAGCACACAGCGGGCAGACUGCAUCUGAAUUACCUGUUGCGAC GAAGAAGACGACGGACGCAGCUUGCCGUUGCGUGCUCACUUCUCUCUCUGUCAGCU	MI0000659	*Oryza sativa*	Poaceae	Angiosperm
>osa-MIR156h AGUUGACAGAAGAGAGUGAGCACACAGCGGCCAGACUGCAUCGAUCUAUCAAUCUUCCCU UCGACAGGAUAGCUAGAUAGAAAGAAAGAGAGGCCGUCGGCGGCCAUGGAAGAGAGAGAG AGAGAGAGAUGAAAUGAUGAUGAUGAUACAGCUGCCGCUGCGUGCUCACUUCUCUUUCUG UCAGCU	MI0000660	*Oryza sativa*	Poaceae	Angiosperm
>osa-MIR156i GGUGACAGAAGAGAGUGAGCACACGGCCGGGCGGAACGGCACCGGCGGAUGUGCCGUCGC GGCCGCGUGCUCACUGCUCUGUCUGUCAUC	MI0000661	*Oryza sativa*	Poaceae	Angiosperm

Table S3. Plant pre-miRNAs families list and their functions from pre-miRNAs filtered by length on four clades. Here is the first page.

pre-miRNAs	Functions	BRY	GYM	ANG	N	References
156	Resistence to salinity stress, plant development, drought	x	x	x	362	Zhou et al 2010; Djami-Tchatchou et al 2017; Yousuf et al 2021
160	Flower development	x	x	x	156	Zhou et al 2010
166	Implicated in the vascular development of the leaf	x	x	x	288	Zhou et al 2010; Djami-Tchatchou et al 2017; Yousuf et al 2021
167	Flower development	x	x	x	182	Zhou et al 2010
171	Associated in growth, metabolism	x	x	x	263	Zhou et al 2010; Djami-Tchatchou et al 2017; Yousuf et al 2021
319	Nutrient deficiency responses	x	x	x	130	Zhu et al 2021
390	Plant development	x	x	x	79	Zhou et al 2010
395	Response to cadmium ion, sulfate transport	x	x	x	233	Jagadeeswaran et al. 2009; Zhou et al 2010
408	Response to water deprivation, plant reproduction	x	x	x	55	Abdel-Ghany 2008; Axtell et al. 2007; Zhou et al 2010
529	Plant development	x	x	x	32	Axtell et al. 2007; Zhou et al 2010
535	Negatively regulates cold tolerance	x	x	x	40	Sun et al 2020
536	Phytohormone abscisic acid	x	x	x	10	Xia et al 2016
414	Transcriptional regulation	x		x	3	Zhou et al 2010
419	Formation of golgi vesicles	x		x	3	Zhou et al 2010
477	Regulation of plant development, translation	x		x	57	Fattash et al. 2007; Lu et al. 2005; Addo-Quaye et al. 2009; Axtell et al. 2007
1078	Regulatation protein kinases	x		x	2	Arazi et al. 2012

pre-miRNAs	Functions	GYM	ANG	N	References
159	Leaf morphogenesis, embryonic development	x	x	111	Zhou et al 2010; Alves et al. 2009; Schwab et al. 2005; Pantaleo et al. 2010; Axtell et al. 2007
162	Flower development, virus induced gene silencing	x	x	48	Zhou et al 2010
164	Root development	x	x	131	Zhou et al 2010
168	Plant development	x	x	59	Zhou et al 2010
169	Flowering timing, photosynthesis	x	x	390	Zhou et al 2010; Djami-Tchatchou et al 2017; Yousuf et al 2021
391	Plant development	x	x	20	Zhou et al 2010
393	Defense response, flower/root development	x	x	79	Jones-Rhoades, Bartel 2004; Zhou et al 2010
394	Regulation of the cell cycle, transcription	x	x	56	Zhou et al 2010
396	Leaf development	x	x	172	Zhou et al 2010
397	Response cold stress, response to water deprivation	x	x	68	Jagadeeswaran et al. 2009; Pantaleo et al 2010
398	Energy metabolism	x	x	72	Beauclair et al. 2010; Zhou et al 2010
399	Phosphate homeostasis	x	x	214	Jagadeeswaran et al. 2009
472	Response to pathogens infection	x	x	6	Fahlgren et al. 2007
482	Regulation to metabolism	x	x	101	Lu et al. 2005
828	Plant development	x	x	19	Rajagopalan et al. 2006
858	Plant development	x	x	6	Addo-Quaye et al. 2008
948	Response to pathogen infection	x	x	2	Su et al 2017
1863	Expression during short and prolonged heat stress	x	x	2	Mangrauthia et al 2017
2111	Response high-concentration nitrate	x	x	56	Okuma and Kawaguchi 2021
2950	Fiber development	x	x	6	Salih et al 2019
3627	Aluminum tolerance	x	x	24	Zhou et al 2010
4414	Response to ethylene	x	x	5	Zhou et al 2010

References

Abdel-Ghany SE, Pilon M (2008) MicroRNA-mediated systemic down-regulation of copper protein expression in response to low copper availability in Arabidopsis. J Biol Chem 283:15932-45

BRY: Bryophytes, GYM: Gymnosperms, ANG: Angiosperms.

Table S4. Plant pre-miRNAs filtered by length (70 to 300 nt) plus 80% redundancy threshold: 3608 of 2623 miRNA families from 77 plant species, 34 taxonomic families, and four clades. Here is the first page.

Plant miRNAs	Id miRBase	Species	Families	Clades
>ath-MIR156b GCUAGAAGAGGGAGAGAUGGUGAUUGAGGAAUGCAACAGAGAAAACUGACAGAAGAGAGU GAGCACAUGCAGGCACUGUUAUGUGUCUAUAACUUUGCGUGUGCGUGCUCACCUCUCUUU CUGUCAGUUGCCUAUCUCUGCCUGCUUGACCUCUCUCUCUCUCUCUCUCUCAAAUUUG GCU	MI0000179	*Arabidopsis thaliana*	Brassicaceae	Angiosperm
>osa-MIR156l GCUAGGGAGCCGACAGAAGAGAGUGAGCAUAUAUAGUUCUUUCCUUGCAUAUGUGGUCAU AUGUGUGUUGACUGAAGAGAUACAUAUAUAUAGAGAGAGAGAGUUCAUGUGCUUGAAGCU AUAUGUGCUCACUUCUCUUUCUGUCAGCAAAUUAUC	MI0001091	*Oryza sativa*	Poaceae	Angiosperm
>gma-MIR156c ACUUGACCACUAGGCUUAUCUCUUUCCGUUUCUGAGCAUGCAUACUCAUUCACAGCAUCA AAAUGCACAGAUCCUGAUGGAGAUUGCACAGGGCAGGUGAUGCUAGAUUGCACCAUACUC AACUCUGGACUUUGUGAUUGAAGUGUUGACAGAAGAUAGAGAGCACAACCUGAGUCAAAG GAUCC	MI0001772	*Glycine max*	Fabaceae	Angiosperm
>zma-MIR156j CGAGUGGACCUCGGGAGCGAUGACAGAAGAGAGAGAGCACAACCCAGCACCAGCGAGGAA AAGCCUCGCUUCUGCGAGGGCCGUGUGCUCUCUGCUCUCACUGUCAUCGCCCACAGGCCA CCGAA	MI0001808	*Zea mays*	Poaceae	Angiosperm
>ppt-MIR156b GUGAGGCUCGAGUGCAGACACUAUUCAAGUGUGGGCGGGGGGAGCGGGAGUGACAGAAGA GAGUGAGCACGGUUGCGCCUGUACGAGGAUAACGUACUUGCUCAAGUGUGCUCACUCUCU UCAUGUCGCGCCUCUCCCUCGUCGUUGUGACGUCAUUUGUGAUGAGAGGGUGGGGAGGGG GGUGUGUGAGGGAGGAGGAGGGGAGGUGAAGGGUGUGGUG	MI0005696	*Physcomitrella patens*	Funariaceae	Bryophyte
>smo-MIR156d GGAUCUAUUGUUGACAGAAGACAGGGAGCACCAGCAGCCACCAGCUCUAGAGAGGCUUUU GGGUGCUUGUGCUCUCUAUUCUUCUGCCAUCAAUCCGACUCGG	MI0006049	*Selaginella moellendorj*	Selaginellaceae	Gymnosperm
>bna-MIR156b UAGGUUUGAGAGUGAUGCUGGUUGUUGACAGAAGAUAGAGAGCACUAAGGAUGACAUGCA AGUACAUAUGUAUGUAUCAUCACACCGCCUGUGGAUGAUUACAAAAAUAAAACCAAUUCA AAAGAGAGAGAGAGAGAGCCUGCAUGUCUACUCUUUCGUGCUCUCUAUACUUCUGUCACC ACCAUUAUUUCUUCUUCUUUCUUACCUA	MI0006481	*Brassica napus*	Brassicaceae	Angiosperm
>vvi-MIR156h UGCCUCACAAUGACAGAAGAGAGAGAGCAUGCUGGUGGGAAAACAAUUACAACUUUUGAU CAUCUGAUCUGGAAAUGCUUGUAAGCGGCAUUCUCUUGGAUUGUAAUCUGAAUUCUGCCU CUAUCAUCAACCUGCCCACAAACGAUUUCCUUCAACUGAGUGCCUUUCCGGCUUGAGCCU UCUGCAUGAUCAGCUGAGUUCUUUCUGCGCCUUUCAUUGUGUCCUGCC	MI0007939	*Vitis vinifera*	Vitaceae	Angiosperm
>ctr-MIR156 UGAAGAGGAAGAGGAAACAUAACGAGAGAGCUACUGACAGAAGAGAGUGAGCACACGCAG GUAAUUGUAUUAGAAAAUAUCUUUGCAGGUGCGUGCUCGCUCCUCUUCUGUCAGCGUCAU UUUUGCAAGUGCUUCCUCACAACCCCAGCACGUCGUAACCCCCAGCCCUCUUUUGCUCUU CCUUCUAUCA	MI0013293	*Citrus trifoliata*	Rutaceae	Angiosperm
>aly-MIR156a GGGUUGGUUGUGAGUGAAAGAGUUGGGACAAGAGAAACGCAAAGAAACUGACAGAAGAGA GUGAGCACACAAAGGCAAUUUGCAUAUCAUUGCAUUUGCUUCUCUUGCGUGCUCACUGCU CUUUCUGUCAGAUUCCGGUGCUGAUCUCUUUGGCCUGUCCUCGAUCUCUUUGUCUCAAUC UCUC	MI0014502	*Arabidopsis lyrata*	Brassicaceae	Angiosperm
>aly-MIR156f GUAUGUGAUAUUAAGAGGUAUAUAUGAAUCAUAAAUAUGGAUGGUUUGAUUGAUGAGUAA UUGAUGGUGACAGAAGAGAGUGAGCACACAUGGUGGCUUUCUUGCAUAUUUGAUGGUUUC AUGCUUGAAGCUAUGUGUGCUCACUCUCUAUCUGUCACCCCCUUCUCUCUCUCUAUAUAC ACCAUUUAAAUUUAUUUUAUAGAGUUAUAUACAAUUUGAUUUUGAUAC	MI0014507	*Arabidopsis lyrata*	Brassicaceae	Angiosperm

Table S5. Novel pre-miRNAs identified in rice and their functions using covariance models at PmiR-Select. This table has two pages; here is the first page.

Functions	Known pre-miRNA rice	New pre-miRNA rice	N	Plant species	References
Resistance to salinity stress				Sorghum	Sanousi et al. 2016
	miR1861		21	Rice	Teng et al. 2022
		miR2590	8	*Medicago sativa*	Ma et al. 2019
		miR437	313	Rice	Zhu et al. 2012
Resistence against pathogens	miR2118		15	*Medicago truncatula*	Jatan and Lata, 2019
		miR1507	1	Legumes (Cowpea)	Chand et al. 2021
Drought stress response	miR408		4	Pea	
Ascochyta blight resistance	miR482		3	Chickpea	
Fusarium wilt infection	miR530		214	Chickpea	
Suppresses invasion fungal	miR818		1249	*Triticum aestivum*	
		miR1023	5	Wheat	Nair et al. 2020
		miR7696	1		Bej S and Basak J, 2014
Plant development	miR528		2	Rice	Jiao and Peng, 2018
		miR8771	12	Cotton	Zhou et al. 2010
		miR10993	1	Apple	Zhou et al. 2022
		miR1313	1	*Pinus sylvestris*	Wang et al. 2022
	miR418		4	Lettuce	Han et al. 2010
Implications in immunity	miR812		4066	Rice	Krivmane et al. 2020
Metabolic regulation	miR821		1627	*Citrus sinensis*	Campo et al. 2021
Pollen maturation stage	miR1428		13	Rice	Lu et al. 2014
Heat, drought and flood stress response	miR1862		6	Cajanus	Zhang et al. 2017
Water deficit response	miR396		19	Grape	Shanmugavadivel et al. 2016
	miR1882		19	Grape	İnal et al. 2020
Cold resistence		miR1435	717	Rice	Lv et al. 2010
Drough and salt stress response		miR2592	8	*Phaseolus vulgaris*	Kavas et al. 2023
		miR7494	29	*Gossypium spp.*	Bano et al 2021
Methal tolerance		miR2670	3	Solanum torvum	Kang et al. 2017
Regulation in strobilus development		miR3701	2	*Pinus massoniana*	Ye et al. 2020
Regulation biosynthesis		miR7540	6	*Xanthium strumarium*	Fan et al. 2015
Viruses resistance		miR7981	1	Tomato	Prigigallo et al. 2019
Regulation of seed germination		miR10186	1	*Arabidopsis thaliana*	Wu et al. 2021
Phosphate homeostasis	miR399		1		Jagadeeswaran et al. 2009
Negatively regulates rice immunity	miR439		1	Rice	Junhua et al. 2021
Patogen responsive		miR1137	94	Wheat	
Leaf development		miR5185	1	*Cinnamomum burmannii*	Hou et al. 2023
Cellulose synthesis		miR5298	2	Eucalyptus grandis	Zhang et al. 2021
	miR5512		2		

References

Bano N, Fakhrah S, Mohanty CS, Bag SK (2021) Genome-wide identification and evolutionary analysis of gossypium tubby-like protein (TLP) gene family and expression analyses during salt and drought stress. Frontiers in Plant Science 12, 667929. doi: 10.3389/fpls.2021.667929

Bej S and Basak J (2014) MicroRNAs: the potential biomarkers in plant stress response. American Journal of Plant Sciences, 2014.

Campo S, Sánchez-Sanuy F, Camargo-Ramírez R, et al. (2021) A novel Transposable element-derived microRNA participates in plant immunity to rice blast disease. Plant biotechnology journal, 19(9), 1798-1811. doi: 10.1111/pbi.13592

Capítulo 3

Identificação de pre-miRNAs no ipê rosa (*Handroanthus impeiginosus* Mart. ex DC. Mattos) utilizando a ferramenta computacional PmiR-Select e o pipeline baseado nos modelos ocultos de Markov (hidden Markov models)

3.1 Introdução

Neste capítulo, foram identificados pre-miRNAs no ipê rosa (*Handroanthus impeiginosus* Mart. ex DC. Mattos), que é uma espécie nativa do bioma cerrado brasileiro. É encontrada em florestas tropicais, florestas semidecíduas e matas ciliares, distribuídas pela América Central e do Sul. Economicamente ele é utilizado para a produção de madeira para construção civil, naval e na fabricação de móveis. É também utilizado para restauração de áreas degradadas. A beleza das suas flores faz com que esta árvore seja muito apreciada no paisagismo urbano, sendo encontrada em muitas cidades brasileiras (Dal et al. 2015).

O genoma do ipê rosa possui 503 Mb, o que equivale aproximadamente a 90% de seu tamanho total de 557 Mb. Nesse estudo foi predito as estruturas transcritas do RNA (n=35479) e identificados alguns genes para metabólitos especializados. Essa foi à primeira espécie da família Bignoniaceae a ser sequenciada e a primeira árvore da floresta Neotropical (Silva-Junior et al. 2018). Posteriormente, o genoma plastidial do ipê rosa foi sequenciado (159,5 Kb) apresentando 124 genes com sequências altamente conservadas. Análises filogenéticas com 77 desses genes mostram a prevalência de seleção negativa ou evolução neutra na maioria dos genes. Por outro lado, foram detectados indícios de seleção positiva em muitos genes platidiais (Sobreiro et al. 2020). O RNA-Seq do ipê rosa e de outras três árvores pertencentes à família Bignoniaceae revelou que o ipê rosa, que é tipico de floresta tropical sazonalmente seca, foi mais responsivo em nivel do transcriptoma e morfoanatomicamente do que as árvores de savana do estudo (Sobreiro et al. 2021). A identificação de outros transcritos como ncRNAs de pre-miRNAs poderão ajudar a entender o processo regulatório no desenvolvimento de plantas em resposta a estresse biótico e abiótico (Kar and Raichaudhuri, 2021).

O objetivo deste estudo foi identificar pre-miRNAs no ipê rosa; tendo em vista, que essa foi à primeira identificação de pre-miRNAs no ipê rosa. Assim, foi usada a ferramenta computacional PmiR-Select (https://github.com/DeborahBambil/PmiRSelect), baseada em modelos de covariância para fazer a identificação potenciais novos pre-miRNAs (Nawrocki and Eddy 2013) e o pipeline (https://github.com/DeborahBambil/PmiRSelectHMM) baseado em modelos ocultos de Markov (Eddy, 1998). A escolha do ipê rosa para a identificação de pre-miRNAs foi motivada pelo fato de ser uma espécie nativa do cerrado brasileiro, um bioma que possui poucas espécies estudadas, cujo estudo pode servir de modelo para outras espécies a serem estudadas do cerrado e de outros biomas.

3.2 Computational identification of pre-miRNAs on pink ipê (*Handroanthus impetiginous* Mart. ex DC.): a native cerrado species

Deborah Bambil[1,2]* and Lúcio Flávio de Alencar Figueiredo[3]

[1]Department of Cell Biology, University of Brasília (UnB), Brasília, 70910-900, DF, Brazil

[2]Federal Institute of Brasília (IFB) Brasília, 70830-450, DF, Brazil

[3]Department of Botany, UnB, Brasília, 70910-900, DF, Brazil

Corresponding author: Deborah Bambil* - **email:** deborahbambil@gmail.com

Deborah Bambil - https://orcid.org/0000-0001-8307-0888

Lúcio Flávio de Alencar Figueiredo - https://orcid.org/0000-0001-8868-1717

Abstract

microRNAs (miRNAs) represent a class of small non-coding RNAs that play crucial roles in post-transcriptional regulation of gene expression, particularly during plant developmental stages and in response to biotic and abiotic stresses. In the present research, we employed a computational methodology to detect conserved and known precursors (pre-miRNAs), as well as potential novel pre-miRNAs, using the covariance model (CM) through the PmiR-Select tool, and a hidden Markov model (HMM) pipeline in the pink ipê (*Handroanthus impetiginosus*). This native Brazilian species is from the cerrado biome and stands as the first representative of the Bignoniaceae family to undergo genome characterization. Through the application of CM via the PmiR-Select tool, a total of 305 pre-miRNAs belonging to 22 potential novel families were identified within the pink ipê genome. Additionally, utilizing the HMM pipeline, 1293 pre-miRNAs originating from 73 families were discerned. This study provides valuable computational insight into the identification of pre-miRNAs within the pink ipê. The forthcoming experimental validation of these pre-miRNAs holds promise for the advancement of overlooked and resilient plant species across various biomes.

Keywords

Covariance model, hidden Markov models, pink ipê, pre-miRNAs, PmiR-Select

Introduction

The non-coding RNAs (ncRNAs) are important post-transcriptional regulators, regardless of whether they do not encode proteins (Seal et al. 2020). Technology advances have made the identification of ncRNAs more accessible and accurate (Panwar et al. 2014). Including the identification of miRNAs, a small ncRNA, which play a role in the development of plants, as well as in their response to biotic and abiotic stresses (Kar and Raichaudhuri, 2021). The biogenesis of plant miRNAs begins with the primary miRNA (pri-miRNA), which originates the precursor (pre-miRNA) in the cell nucleus. The pre-miRNA is exported to the cytoplasm to undergo cleavage and give rise to mature miRNAs (Dong et al. 2008). Both pre-miRNAs and mature miRNAs have conserved structures in plant families, which is a favourable factor for identification by homology (Sun et al. 2014). According to the updated criteria for pre-miRNAs and mature miRNAs, they have less than 300 nucleotides (nt) in length for the former, ranging from 20 to 24 nt for the latter (Axtel and Meyers, 2018).

The miRBase is the extensive database containing 38589 pre-miRNA hairpins from 265 species, 33% plants (84 plant species) v.22.1. The number of mature miRNAs in miRBase is 20% higher (*n*=48860) with around the same number of families (*n*=270). pre-miRNAs from model species, such as *Arabidopsis thaliana* (L.) Heynh has been used to identify conserved pre-miRNAs in different species (Xuan et al. 2011). However, none of the pre-miRNA and mature miRNA sequences have been annotated from native Brazilian species, despite Brazil having one of the highest diversity and endemism of plants and trees (Beech et al. 2018, BFG 2021). Considering that the identification of pre-miRNAs and mature miRNAs to understand genetic regulatory processes play a crucial role in plant development, and response to biotic and abiotic stresses (Millar, 2020).

Pink ipê is a native species of the Brazilian cerrado biome, found in tropical forests, semi-deciduous forests, and riparian forests distributed across Central and South America (Dal et al. 2015). Sequencing (557 Mb and 37% repetitive DNA) in 2018 made it the first species in the Bignoniaceae family (Silva-Junior et al. 2018). Soon after, its plastid genome was sequenced (159.5 Mb and 124 genes), indicating negative selection or neutral evolution in most genes. On the other hand, signs of positive selection were identified in some genes (Sobreiro et al. 2021). Furthermore, a project involving pink ipê and three more Bignoniaceae tree species revealed different responses to drought from RNA-Seq analysis (Sobreiro et al. 2021). These genomic data will be valuable for understanding its unique adaptations to different environmental conditions.

The pink flowers give the name to the ipê pink tree, while there are other species with white [*Tabebuia roseoalba* (Ridl) Sandwith] and yellow [*Handroanthus serratifolius* (Vahl) S. Grose] colors. The flowers have five petals. The fruits are elongated and contain several seeds in an oblong shape. These flowers supply food for birds. The height of pink ipê ranges from 8 to 30 meters. The bark is thick in brown color, and the diameter ranges from 60 to 100 cm. The tree has an irregular growth (Lorenzi, 2008), as shown in Figure 1a.

Pink ipê is economically valuable for providing wood for civil, naval, and furniture construction (Dal et al. 2015; Souza et al. 2020). In addition, it is commonly used for landscaping and native species for urban afforestation and restoring degraded areas (Souza et al. 2020; Lunardi et al. 2019). The use of plant species has declined in various regions where they naturally occur, such as in the cerrado, which puts these species at risk of extinction (Strassburg 2017). The cerrado is a biodiversity hotspot, covering an area of 200 million hectares and home to over 480 plant and vertebrate species (Strassburg et al. 2017). Agricultural practices (Bueno et al. 2018), invasive species (Pilon et al. 2018), and fires (Musso et al. 2015) represent a threat to the survival of endangered species in the region. In the last year, human activity has caused the loss of 6590 km2 of native cerrado (http://terrabrasilis.dpi.inpe.br/ams/, accessed on June 13, 2023).

The Convention on Biological Diversity established the Aichi strategic plan for Biodiversity 2011-2020, aiming to halt biodiversity loss by restoring, valuing, and conserving it from 2011 to 2020 (O'Connor et al. 2020). Most of the 20 strategic plan targets have not been achieved (Global Biodiversity Outlook 5 - www.cbd.int/GBO5). Genetic studies of endangered native species from the cerrado biome are crucial for both economic use and ecological preservation.

Computational data analysis can provide valuable insights for genetic improvement and species conservation efforts (Ballesteros-Mejia et al. 2020). Approaches such as covariance models (CMs) that use profile stochastic context-free grammars that combine structure with covariance probabilities, enabling the identification of new sequences based on probabilistic models of the conserved sequences and secondary structure of RNA families by statistical correlations between different positions (Nawrocki and Eddy 2013). Other Hidden Markov Model (HMM) profiling approaches are frequently employed in the field of bioinformatics (Potter et al., 2018). This approach was utilized for the identification of pre-miRNAs using the NOVOMIR and HuntMI tools (Gudyś et al., 2013; Teune et al., 2010). The HMM assign their profile statistical models used to model sequences of events with hidden states and look for transitions between those states, being able to identify new and known sequences (Eddy,

1998). This study aims to identify the conserved known pre-miRNAs with a pipeline based on HMM and potential novel ones with the PmiR-Select tool based on CM in pink ipê. This work is the first effort to identify pre-miRNAs in a native species of the cerrado biome.

Materials and methods

Genome search

Pink ipê conserved pre-miRNAs were identified from the genome published by Silva-Júnior et al. (2018). It has a total length of 557 Mb (N50=81316 bp), with 13206 scaffolds, 13204 contigs, and 35479 messenger RNA transcripts (Silva-Junior et al. 2018), and employed a homology-based approach using PmiR-Select (https://github.com/DeborahBambil/PmiRSelect). For potential novel pre-miRNA identification and model-based hidden Markov models (HMM - https://github.com/DeborahBambil/PmiRSelectHMM) to known plant pre-miRNAs (Figure 1).

RNA-Seq search

Pre-miRNAs are also being searched in RNA-Seq sequences annotated in drought conditions to identify the response to water deficit (Sobreiro et al. 2021). The files analyzed were SRR11144483 (Deficit *H. impetiginosus*, Individual 1, 3.1 Gb), SRR11144482 (Deficit *H. impetiginosus*, Individual 2, 3.7 Gb), SRR11144481 (Deficit *H. impetiginosus*, Individual 3, 7.7 Gb), SRR11144486 (Irrigated *H. impetiginosus*, Individual 1, 3.7 Gb), SRR11144485 (Irrigated *H. impetiginosus*, Individual 2, 4.6 Gb), and SRR11144484 (Irrigated *H. impetiginosus*, Individual 3, 6.1 Gb (https://www.ncbi.nlm.nih.gov/sra/?term=handroanthus+impetiginosus).

Identification of pre-miRNAs

The pre-miRNAs were identified with the PmiR-Select (https://github.com/DeborahBambil/PmiRSelect) tool that leverages data mining constructed from miRBase based on covariance models (CM) to identify homologs of pre-miRNAs in genomes, through the use of CM, which are probabilistic models used to assign new RNA sequences to known RNA families. This method enables the identification of highly conserved RNA secondary structures, thereby facilitating the identification of new pre-miRNA sequences (Figure 1b - Nawrocki and Eddy 2013).

The HMM model was employed to identify hidden sequences within families of known plant pre-miRNAs (https://github.com/DeborahBambil/PmiRSelectHMM). The HMM was constructed using the characteristics of pre-miRNAs molecules with the HMMER tool (Eddy, 1998), utilizing the conserved secondary structure in the Stockholm format through the T-coffee tool (Tommaso et al. 2011). Subsequently, the build HMMs for each plant pre-miRNA families were utilized to identify pre-miRNAs within the pink ipe genome with HMM search (Eddy, 1998). The identified sequences underwent further validation of plant pre-miRNAs families using Blast (E-value of 0.01 - Figure 1c - Mahram and Herbordt, 2015). The output corresponds to identified pre-miRNAs, distinguishing between identical and non-identical copies, and by similarity threshold from 95% to 70% (in 5% intervals), and identified curated, in this output corresponds to pre-miRNAs compared with plant mature miRNAs.

Results and discussion

The pink ipê genome analysis using PmiR-Select (https://github.com/DeborahBambil/PmiRSelect) identified 305 pre-miRNA sequences belonging to 22 pre-miRNA families (Figure 1b; Table S1). 3% (*n*=11) of these pre-miRNAs were identical to seven pre-miRNA families. Non-identical copies accounted for 294 pre-miRNAs (Figure 1b; Table S2). The miR530 family, associated with stress responses caused by pathogens (Chand et al. 2021), was identified with the highest number of pre-miRNAs (*n*=44), followed by the miR812 family related to plant immunity (Krivmane et al. 2020 - *n*=44), using the PmiR-Select tool. Upon comparing these results with mature miRNAs, two pre-miRNAs from the same family were selected. It's worth noting that these two pre-miRNAs are not identical copies. This approach was used to identify pre-miRNAs in rice (Bambil and de Alencar, in submission) when pre-miRNAs were identified in families with PmiR-Select (https://github.com/DeborahBambil/PmiRSelect). Such pipelines incorporate CMs into their models (Ding et al. 2011; Joshi et al. 2016; Tang and Sun 2019).

The analysis of the pink ipê genome through the pipeline with HMM (https://github.com/DeborahBambil/PmiRSelectHMM) identified 1293 pre-miRNAs sequences belonging to 73 families (Figure 1c; Table S3). 5% (*n*=73) of these pre-miRNAs were identical to 32 pre-miRNA families. Non-identical copies accounted for 1220 pre-miRNAs (Figure 1c; Table S4), and with the 80% redundancy threshold, a total of 616 pre-miRNAs were selected. The miR1023 family, associated with plant development (Nair et al. 2020), exhibited the highest number of pre-miRNAs (*n*=129) through the HMM approach.

Furthermore, this same family was also identified using the PmiR-Select tool. When comparing these findings with mature miRNAs, no pre-miRNAs were selected.

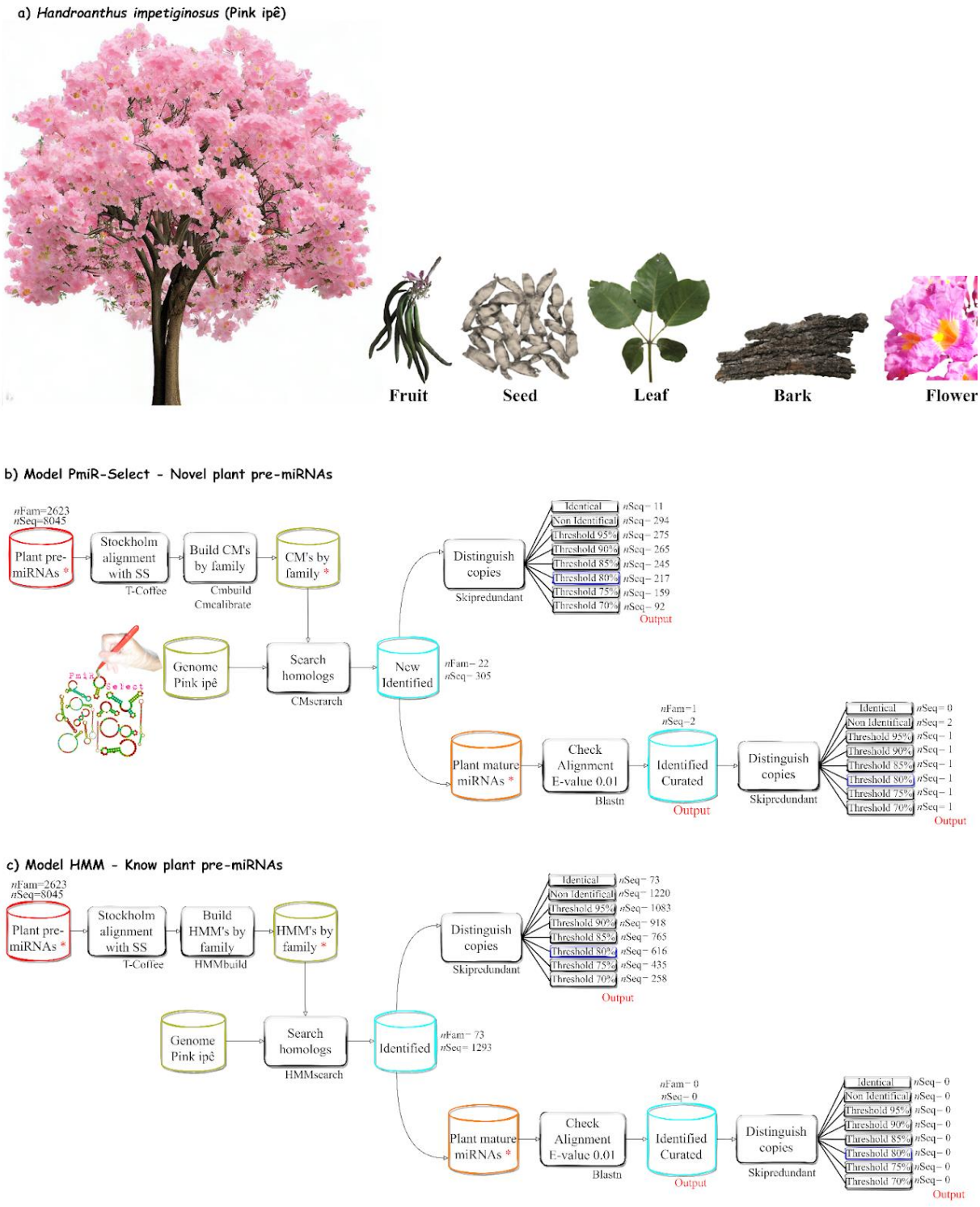


Figure 1. (a) Parts of the plant body of pink ipê: flower, leaf, seed, fruit, and cork. (b) The identification of pre-miRNAs in the pink ipê genome using the PmiR-Select tool (https://github.com/DeborahBambil/PmiRSelect), based on the covariance model, to identify

novel pre-miRNAs. (c) The identification of pre-miRNAs in the pink ipê genome using the pipeline based on hidden Markov models (https://github.com/DeborahBambil/PmiRSelectHMM) to conserved known pre-miRNAs.

However, we used both approaches (CM and HMM) to more accurately identify new pre-miRNAs with CM and others that can be considered known with HMM because it is the first to identify pre-miRNAs of a species native to the cerrado biome. Out of the 95 families (CM *n*=22 and HMM *n*=73), only two pre-miRNA families. With this publication, we aspire to motivate further research on native species of the cerrado, which is one of the world's most biodiverse biomes. We emphasize that computational analyses with RNA-Seq are still in progress.

Abbreviations

CM	Covariance model
HMM	hidden Markov models
miRNAs	microRNAs
ncRNA	Non-coding RNAs
nt	Nucleotides
pre-miRNAs	Precursors miRNAs

Supplementary Information

Table S1. pre-miRNAs identified in the pink ipê genome using the PmiR-Select

Table S2. pre-miRNAs non-identical identified in the pink ipê genome using the PmiR-Select

Table S3. pre-miRNAs identified in the pink ipê genome with hidden Markov models from homolog families.

Table S4. pre-miRNAs non-identical identified in the pink ipê genome through hidden Markov models from homolog families.

Acknowledgments

Not applicable.

Author contributions

All authors of this manuscript have directly participated in the execution of the study. Conceived and designed the experiments: D.B. and L.F.A.F. performed the experiments. D.B. and L.F.A.F. analyzed the data: D.B. and L.F.A.F. wrote the paper. All authors read and approved the final manuscript.

Funding

Brazilian agency CAPES for financial support scholarship to grant for D.B. process number 88887.635040/2021-00.

Availability of data and materials

This published article and its supplementary information files include all data generated or analyzed during this study.

Declarations

Ethics approval and consent to participate.

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

Axtell MJ and Meyers BC (2018) Revisiting criteria for plant microRNA annotation in the era of big data. The Plant Cell 30(2), 272-284. doi: 10.1105/tpc.17.00851

Ballesteros-Mejia L, Lima JS, Collevatti RG (2020) Spatially-explicit analyses reveal the distribution of genetic diversity and plant conservation status in cerrado biome. Biodiversity and Conservation 29(5), 1537-1554. doi: 10.1007/s10531-018-1588-9

Beech, E., Rivers, M., Oldfield, S. & Smith, P.P. (2017). GlobalTreeSearch: The first complete global database of tree species and country distributions. *Journal of Sustainable Forestry*, 36, 454–489.

BFG. (2021). Brazilian flora 2020: leveraging the power of a collaborative scientific network. *TAXON*.

Bond WJ and Parr CL (2010) Beyond the forest edge: ecology, diversity and conservation of the grassy biomes. Biological Conservation 143(10), 2395-2404. doi: 10.1016/j.biocon.2009.12.012

Bueno ML, Dexter KG, Pennington RT, et al. (2018) The environmental triangle of the Cerrado domain: ecological factors driving shifts in tree species composition between forests and savannas. Journal of Ecology 106(5), 2109-2120. doi: 10.1111/1365-2745.12969

Balatti P (2015) First report of *Alternaria alternata* causing black spot on pink lapacho (*Handroanthus impetiginosus*). Australasian Plant Disease Notes 10(1), 1-2. doi: 10.1007/s13314-015-0159-0

Chand H. Nayyar JU, Mantri N, et al. (2021) Non-Coding RNAs in legumes: their emerging roles in regulating biotic/abiotic stress responses and plant growth and development. Cells 10(7), 1674. doi: 10.3390/cells10071674

Dal BG, Franco E, Larrán S, Balatti P (2015) First report of Alternaria alternata causing black spot on pink lapacho (*Handroanthus impetiginosus*). Australasian Plant Disease Notes 10(1), 1-2. doi: 10.1007/s13314-015-0159-0

Ding J, Zhou S, Guan J (2011) miRFam: an effective automatic miRNA classification method based on n-grams and a multiclass SVM. BMC Bioinformatics 12(1), 1-11.doi: 10.1186/1471-2105-12-216

Dong Z, Han MH, Fedoroff N (2008) The RNA-binding proteins HYL1 and SE promote accurate in vitro processing of pri-miRNA by DCL1. Proceedings of the National Academy of Sciences 105(29), 9970-9975. doi: 10.1073pnas.0803356105

Eddy SR (1998) Profile hidden Markov models. Bioinformatics (Oxford, England) 14(9), 755-763. doi:10.1093/bioinformatics/14.9.755

Fernandes GW, Vale MM, Overbeck GE, et al. (2017) Dismantling Brazil's science threatens global biodiversity heritage. Perspectives in Ecology and Conservation 15(3), 239-243. doi: 10.1016/j.pecon.2017.07.004

Gautam V, Singh A, Verma S, et al. (2017) Role of miRNAs in root development of model plant *Arabidopsis thaliana*. Indian Journal of Plant Physiology 22(4), 382-392. doi: 10.1007/s40502-017-0334-8

Gudyś A, Szcześniak MW, Sikora M, Makałowska I (2013) HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. BMC Bioinformatics 14(1), 1-10. doi: 10.1186/1471-2105-14-83

Joshi RK, Megha S, Basu U, Rahman MH, Kav NN (2016) Genome wide identification and functional prediction of long non-coding RNAs responsive to S*clerotinia sclerotiorum* infection in *Brassica napus*. PLoS One 11(7), e0158784. doi: 10.1371/journal.pone.0158784

Kar MM and Raichaudhuri A (2021) Role of microRNAs in mediating biotic and abiotic stress in plants. Plant Gene 26, 100277. doi: 10.1016/j.plgene.2021.100277

Krivmane B, Šņepste I, Šķipars V, et al. (2020) Identification and in silico characterization of novel and conserved microRNAs in methyl jasmonate-stimulated scots pine (Pinus sylvestris L.) needles. Forests 11(4), 384. doi: /10.3390/f11040384

Lorenzi H (2008) Árvores brasileiras: manual de identificação e cultivo de plantas arboreas do Brasil. Instituto Plantarum Nova Odessa, SP, Vol. 1. 5, 2008.

Lunardi VO, Silva EEM, Silva STA, Lunardi DG (2019) *Handroanthus impetiginosus* (Bignoniaceae) As an important floral resource for synanthropic birds in the brazilian semiarid. Oecologia Australis. doi: 10.4257/OECO.2019.2301.12

Mahram A and Herbordt MC (2015) NCBI BLASTP on high-performance reconfigurable computing systems. ACM Transactions on reconfigurable Technology and Systems (TRETS) 7(4), 1-20. doi: 10.1145/2629691

Millar AA (2020) The function of miRNAs in plants. Plants 9(2), 198. doi: 10.3390/plants9020198

Musso C, Miranda HS, Aires SS, et al. (2015) Simulated post-fire temperature affects germination of native and invasive grasses in cerrado (Brazilian savanna). Plant Ecology & Diversity 8(2), 219-227. doi: 10.1080/17550874.2014.910714

Nair MM, Krishna TS, Alagu M (2020) Bioinformatics insights into microRNA mediated gene regulation in Triticum aestivum during multiple fungal diseases. Plant Gene, 21, 100219. doi: 10.1016/j.plgene.2019.100219

Nawrocki EP and Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22), 2933-2935. doi: 10.1093/bioinformatics/btt509

O'Connor B, Secades C, Penner J, et al. (2015) Earth observation as a tool for tracking progress towards the Aichi Biodiversity Targets. Remote Sensing in Ecology and Conservation 1(1), 19-28. doi: 10.1002/rse2.4

Panwar BA, Arora A, Raghava GPS (2014) Prediction and classification of ncRNAs using structural information. BMC Genomics 15(1):1-13. doi: 1471-2164/15/127

Pilon NAL, Buisson E, Durigan G (2018) Restoring Brazilian savanna ground layer vegetation by topsoil and hay transfer. Restoration Ecology 26(1), 73-81. doi: 10.1111/rec.12534

Potter SC, Luciani A, Eddy SR, et al. (2018) HMMER web server: 2018 update. Nucleic Acids Research 46(W1), W200-W204. doi:10.1093/nar/gky448

Seal RL, Chen LL, Griffiths-Jones S, et al. (2020) A guide to naming human non-coding RNA genes. The EMBO Journal 39(6), e103777. doi: 10.15252/embj.2019103777

Silva-Junior OB, Grattapaglia D, Novaes E, Collevatti RG (2018) Genome assembly of the pink ipê (*Handroanthus impetiginosus*, Bignoniaceae), a highly valued, ecologically keystone neotropical timber forest tree. Gigascience 7(1), gix125. doi: 10.1093/gigascience/gix125

Sobreiro MB, Vieira LD, Nunes R, et al. (2020). Chloroplast genome assembly of *Handroanthus impetiginosus*: comparative analysis and molecular evolution in Bignoniaceae. *Planta*, **252**, 91. doi: 10.1007/s00425-020-03498-9

Sobreiro MB, Collevatti RG, Dos Santos YL, et al. (2021) RNA-Seq reveals different responses to drought in Neotropical trees from savannas and seasonally dry forests. BMC Plant Biology 21(1), 1-17. doi: 10.1186/s12870-021-03244-7

Souza JMA, Sampaio PDTB, Degterev IA, et al. (2020) Longitudinal distribution of lapachol in the stalk of ipê species (*Handroanthus* spp.). European Journal of Wood and Wood Products 78(3), 609-611. doi: 10.1007/s00107-020-01530-z

Strassburg BB, Brooks T, Feltran-Barbieri R, et al. (2017) Moment of truth for the Cerrado hotspot. Nature Ecology & Evolution 1(4), 1-3. doi: 10.1038/s41559-017-0099

Sun X, Zhang Y, Zhu X, et al. (2014) Advances in identification and validation of plant microRNAs and their target genes. Physiologia Plantarum 152(2), 203-218. doi:10.1111/ppl.12191

Tang X and Sun Y (2019) Fast and accurate microRNA search using CNN. BMC Bioinformatics 20(23), 1-14. doi: 10.1186/s12859-019-3279-2

Teune JH and Steger G (2010) NOVOMIR: de novo prediction of microRNA-coding regions in a single plant-genome. Journal of Nucleic Acids 2090-0201. doi: 10.4061/2010/495904

Tommaso PD, Moretti S, Xenarios I, et al. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Research 39, W13-W17. doi: 10.1093/nar/gkr245

Xuan P, Guo M, Liu X, Huang, et al. (2011) PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. Bioinformatics 27(10), 1368-1376. doi: 10.1093/bioinformatics/btr153

3.3 Conclusão

A partir do genoma disponível do ipê rosa, foram utilizadas duas ferramentas computacionais para a identificação inédita de pre-miRNAs ipê rosa. A primeira ferramenta foi a PmiR-Select (https://github.com/DeborahBambil/PmiRSelect) que utiliza modelos de covariância com pre-miRNAs de 70 a 300 nt. Com essa abordagem, foram identificadas 305 potenciais novos pre-miRNAs, distribuídos em 22 famílias e foram identificadas 11 cópias idênticas de pre-miRNAs. Com o filtro do limite de redundância de 80%, foram selecionados 217 pre-miRNAs. A família do miR530, relacionada a respostas ao estresse causado por patógeno, foi identificada com o maior número de pre-miRNAs (*n*=44), seguida do miR812 (*n*=44) relacionado a imunidade da planta. Ao compararmos esses resultados com os miRNAs maduros, dois pre-miRNAs, pertencentes a uma mesma família, foram selecionados, esses dois pre-miRNAs não são cópias idênticas.

Com a segunda abordagem, baseada no pipeline dos modelos ocultos de Markov (hidden Markov models - HMM https://github.com/DeborahBambil/PmiRSelectHMM), foram identificados 1293 pre-miRNAs de 73 famílias. Um total de 73 pre-miRNAs (5%) são cópias idênticas. Com o limite de redundância de 80%, foram selecionados 616 pre-miRNAs de 73 famílias. Na comparação desses resultados com os miRNAs maduros, não foram identificados pre-miRNAs. A família do miR1023, relacionada ao desenvolvimento da planta, obteve o maior número de pre-miRNAs (*n*=129), essa família também foi a única encontrada com a ferramenta PmiR-Select. Corroborando a importância do uso de mais de um modelo na identificação de pre-miRNAs. Das 95 famílias (73 HMM + 22 PmiR-Select) somente uma ocorreu em comum entre os dois modelos, fortalecendo a complementaridade deles.

Devido ao grande tamanho dos arquivos (39655 Mb) a análise do RNA-Seq está sendo concluída com os dois modelos (covariância da PmiR-Select e HMM). Sendo o ipê rosa e o amarelo os mais responsivos ao estresse de seca, espera-se uma correlação dos miRNAs com os transcritos responsivos envolvidos com essa condição de estresse. Este estudo com o RNA-Seq poderá proporcionar perspectivas não identificadas na diversidade dessas quatro árvores estudadas que podem representar um passo significativo no entendimento dos mecanismos de regulação genética de algumas espécies de árvores a seca. Isso poderá contribuir para estudos de conservação e manejo de ecossistemas onde o ipê rosa é encontrado.

Table S1. pre-miRNAs identified in the pink ipê genome using the PmiR-Select through covariance models from homolog families. Here is the first page.

Table S2. pre-miRNAs non-identical identified in the pink ipê genome using the PmiR-Select through covariance models from homolog families. Here is the first page.

Table S3. pre-miRNAs identified in the pink ipê genome through hidden Markov models from homolog families. Here is the first page.

Table S4. pre-miRNAs non-identical identified in the pink ipê genome through hidden Markov models from homolog families. Here is the first page.

In this appendix, only the first page of the tables has been included. The remaining pages are available through the link or QR code below.

Link to access the tables:

https://1drv.ms/f/s!ApXNAj3cFXgAg_MHJoUxzaXPKdsfdw?e=K0mBXn

QrCode to access the tables:

Table S1. pre-miRNAs identified in the pink ipê genome using the PmiR-Select through covariance models from homolog families. Here is the first page.

pre-miRNA identified	pre-miRNA family
>NKXS01003504.1/93484-93364 AGGGAAAAUGAGAAGGAUAGAGAGAGGAAGAAGAGUUUAUAUUCCACCCCACUUUUUUU CUAUUGUUUACACAUUUUUCUAAUAUGAUAUGAUUUUUUUUUCUUUUCUUUUCUUUUUUUUU U	miR390
>NKXS01004161.1/58501-58630 ACGUUCUUCUUCUUCUUCUUUUUUUCUUUUUUUUUUUUCUUUUUUUCUUUGGAGGGGGUUUG AAUGUGGCAGGAGAAAAAAGAAAAAGAAAAAGCAAAAGAAAAAUGGAGUUGGGUAGGG GUAGUGGUGAU	miR396
>NKXS01009554.1/10875-11017 UUUUUCUUUUUUUCUCUUCUUCUUCUUUUUUUUUUUUCCCCCUUCUUUGUUUUUUUUUUUCAA UUUUUGGUAAUUCUUGGACAAACAUGAAGAGGAAGAGAGAAAGGGAUAGACACAAAGUG AAUGGAGCGGAAUGAAUAGGCA	
>NKXS01002948.1/10647-10513 UCCUUUUUUCUCUCUUUUUUCUUUCUUCUCUUUUCUCUCUUUUUUUCUUCUCUCUUCUUGCA UGACAUGAGGGCUUGAUUUUCUAAAAAAAAAAUUGUUUAAUCAUAUAGAAAAAUGUUAGG AAUUAUUGGUCAUU	
>NKXS01007841.1/5658-5537 GUAUGGUUGUGGUAGGGGGGUGGAAGGGGGAAGGAGAGAGAAAGGUUAGAGGAAGAAG AUGACUAAAAUUUUCUUCUCUUUCCUUUUUUUGUUUUUUUUUUUUUUGUUUUUUGAUUGAGC CUG	miR408
>NKXS01000284.1/16975-16839 AAAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGGAGAGAAUUUUAAUAA AAAAAAAAUAUCACAUUGCCGUUAUUUCUUUCUUUCCUUUCCUUUUCUUUCCUUUUCUUUU UGCCCUAUUUAUGUGUGU	
>NKXS01005832.1/22851-22763 GUAGACAUGGCGUGGUAUUGCUUUGGCUCAGCGCUUUUCAAGUAAACAGUUUGGCGAUGA GCCGAAUCAAUAUCACUCUUGUAUGCUUU	miR479
>NKXS01005832.1/16084-15994 GUAGACAUGGCGUGGUAUUGCUUUGGCUCAGCGCUUUUCAAGUAAGCAGUUUGGCGUUGA GCCGAAUCAAUAUCACUCUUGUAUGCUUUUU	
>NKXS01005322.1/36690-36564 AGAGAGAGAGAGAGAGAGAGAUAGAGAGAGAGAGAGAGAGAGAGAGAAUUUUAAUCGG AAUUUCACACUGUCAUUAUUCCUUUCUUUCCUUCUUUCUUUCUUUUCCUUUUCCUUCCUUUU CCUUCCU	miR482
>NKXS01000158.1/134582-134481 GGGGGGGGGGGGGGGGGGGGGGGCGGGGUAGAGGAAGAAGAUGUUUUAAUUUUUUUUUU AAUUUUUUUUCCUUUUAUUUAUUUAUUUAUUAUUUUUUUU	
>NKXS01001436.1/68521-68633 GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGUAAGAAAAAGAAAAAAAGGAUUUAAA UGAAAUGCCCCAGCUCCUCUUUUUUCCAUUAAUUUUCAAUUAAUUUUCUUUUCUU	

Table S2. pre-miRNAs non-identical identified in the pink ipê genome using the PmiR-Select through covariance models from homolog families. Here is the first page.

Pre-miRNA identified	Pre-miRNA family
>NKXS01003504.1_93484-93364 AGGGAAAAUGAGAAGGAUAGAGAGAGGAAGAAGAGUUUAUAUUCCACCCCACU UUUUUUC UAUUGUUUACACAUUUUUCUAAUAUGAUAUGAUUUUUUUUUCUUUUCUUUUCUUUU UUUUU U	miR390
>NKXS01004161.1_58501-58630 ACGUUCUUCUUCUUCUUCUUUUUUUCUUUUUUUUUUUUCUUUUUUUCUUUGGAGGG GGUUU GAAUGUGGCAGGAGAAAAAAGAAAAAGAAAAAGCAAAAGAAAAAUGGAGUUG GGUAGGGG UAGUGGUGAU	miR396
>NKXS01009554.1_10875-11017 UUUUUCUUUUUUUCUCUUCUUCUUCUUUUUUUUUUUUCCCCCUUCUUUGUUUUUUUU UUUC AAUUUUUGGUAAUUCUUGGACAAACAUGAAGAGGAAGAGAGAAAGGGAUAGAC ACAAAGU GAAUGGAGCGGAAUGAAUAGGCA	
>NKXS01002948.1_10647-10513 UCCUUUUUUCUCUCUUUUUUCUUUCUUCUCUUUUCUCUCUUUUUUUCUUCUCUCUU CUUG CAUGACAUGAGGGCUUGAUUUUCUAAAAAAAAAAUUGUUUAAUCAUAUAGAAAAA UGUUAG GAAUUAUUGGUCAUU	
>NKXS01007841.1_5658-5537 GUAUGGUUGUGGUAGGGGGGUGGAAGGGGGAAGGAGAGAGAAAGGUUAGAGGA AGAAGAU GACUAAAAUUUUCUUCUCUUUCCUUUUUUUGUUUUUUUUUUUUUUGUUUUUUGAUU GAGCC UG	miR408
>NKXS01000284.1_16975-16839 AAAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGGAGAGAAUUU UAAUAAAA AAAAAAUAUCACAUUGCCGUUAUUUCUUUCUUUCCUUUCCUUUUCUUUCCUUUUC UUUUU GCCCUAUUUAUGUGUGU	
>NKXS01005832.1_22851-22763 GUAGACAUGGCGUGGUAUUGCUUUGGCUCAGCGCUUUUCAAGUAAACAGUUUGG CGAUGA GCCGAAUCAAUAUCACUCUUGUAUGCUUU	miR479
>NKXS01005832.1_16084-15994 GUAGACAUGGCGUGGUAUUGCUUUGGCUCAGCGCUUUUCAAGUAAGCAGUUUGG CGUUGA GCCGAAUCAAUAUCACUCUUGUAUGCUUUUU	

Table S3. pre-miRNAs identified in the pink ipê genome through hidden Markov models from homolog families. Here is the first page.

Pre-miRNA identified	Pre-miRNA family
>NKXS01002280.1/3680-3837 AAAGUGAGAAACAGACGUGAUGuaguugagagagagagagagagagagauuucuuuuccuuuuuuu uaauuccucuaUAACUCCUUUAAGUUGUGGAUUAGAACUGUUCACUACUUUAAGAUU AAUAAUACCUUCAAUACAAAAUCUAUUGUAUAA >NKXS01002280.1/4617-4774 AAAGUGAGAAACAGACGUGAUGuaguugagagagagagagagagagagauuucuuuuccuuuuuuu uaauuccucuaUAACUCCUUUAAGUUGUGGAUUAGAACUGUUCACUACUUUAAGAUU AAUAAUACCUUCAAUACAAAAUCUAUUGUAUAA >NKXS01002859.1/45801-45949 CCUUUACCUCCACCGUCGguacaaagagagagagagagagagagagagagagagacacuaACGCGCA GUUUUCAACAACCUCUCCUCUGAAACCCUCUUCUUUUUCAAAGUUACGCUCUAA AACUACCUUAGUCGGcuagauauuuauu >NKXS01002859.1/75477-75597 CUUucuuuuagagagagagagagagagagagaauagaUACACUUUGAUUCACCGUUAAUCAUCCA ACGAAUAGAACUACAAAACAAAGAGUCUAAACACUGUCACCAAGUUUGUGUCG UAU >NKXS01000227.1/32742-32881 cuuucuuccuuGUCGCUAAGAGAGAugggagagagggagagggaggagggaaaaaagaaacuucuuuauacu uaaaaaaacuuauuuuaauuaaaacacuuucuUCCAACACUCACACUCUUUGagaaauuaaacauuau >NKXS01000227.1/236750-236902 aaauuaauuuuaagagaagagagagagagagagagaaaaaaggguAUGUGUUUGUGCUUAUCAucucuuau uuuuacuuaccaUCGUAUUCCCCUcccuuauaauacaaucuuuuuuauuccuuuucuucaauCUUGUCUC UCAAAUUAGUCA >NKXS01004632.1/44545-44682 uacauaauuagcaaagagaaagagagagagagagagagagggggggagaGCGAACGCGCGCCCCUCAC CCCCCUACUUAUUUUAAAGUGAUCCACUAUAGAGGUUACUUAAACGUUAUCUU AUGAAAAUUUAGAUC >NKXS01008910.1/10013-10158 GAUuaagagacagagagagagagagagagagagagagagagagagaaggaaaucacaguuuuuuauuauaacuuau uacaAAUACACUACUUAUGUAUACAUGUCCCGAAGUUUCUACACCCUUCUCGACA GCUGAAGCUACCU >NKXS01008910.1/11928-12045 UGUACCGAUUCGUUUaaauugagagagagagacagagagagagagaugugcAUAUAUaccaacacauau auauaauuaucgaUUUAUUUGAAAGAAUUAUUUAACUGGAUUAUCCGUG >NKXS01004191.1/16759-16902 CAACAAACCUAUAuaagagagagagugagagagagagagagagagagauaaggaaccuauuuacuauua aaauaggAAAAAUUUACCGUACACUACGAUCGUACCCAguuaaaaacccguuuuuaacuaacgu uucaacu	miR1023

Table S4. pre-miRNAs non-identical identified in the pink ipê genome through hidden Markov models from homolog families. Here is the first page.

Pre-miRNA identified	Pre-miRNA family
>NKXS01002280.1_3680-3837 AAAGUGAGAAACAGACGUGAUGuaguugagagagagagagagagagagauuucuuuuc cuuuuuuuuaauuccucuaUAACUCCUUUAAGUUGUGGAUUAGAACUGUUCACUACUUUA AGAUUAAUAAUACCUUCAAUACAAAAUCUAUUGUAUAA >NKXS01002859.1_75477-75597 CUUucuuuuagagagagagagagagagagaauagaUACACUUUGAUUCACCGUUAAUC AUCCAACGAAUAGAACUACAAAACAAAGAGUCUAAACACUGUCACCAAGUUUG UGUCGUA U >NKXS01000227.1_32742-32881 cuuucuuccuuGUCGCUAAGAGAGAugggagagagggagagggaggagggaaaaaagaaa cuucuuuauacuuaaaaaaacuuauuuuaauuaaaacacuuucuUCCAACACUCACACUC UUUGagaaauuaaacauuau >NKXS01000227.1_236750-236902 aaauuaauuuuaagagaagagagagagagagagaaaaaaggguAUGUGUUUGUGCUUA UCAucucuuauuuuuacuuaccaUCGUAUUCCCCUcccuuauaauacaaucuuuuuuauu ccuuuucuucaauCUUGUCUCUCAAAUUAGUCA >NKXS01004632.1_44545-44682 uacauaauuagcaaagagaaagagagagagagagagagagggggggagaGCGAACGCG CGCCCCUCACCCCCCUACUUAUUUUAAAGUGAUCCACUAUAGAGGUUACUUAA ACGUUAU CUUAUGAAAAUUUAGAUC >NKXS01008910.1_10013-10158 GAUuaagagacagagagagagagagagagagagagagagagagagaaggaaaucacag uuuuuuauuauaacuuauuacaAAUACACUACUUAUGUAUACAUGUCCCGAAGUUUCUAC ACCCUUCUCGACAGCUGAAGCUACCU >NKXS01008910.1_11928-12045 UGUACCGAUUCGUUUaaauugagagagagagacagagagagagagaugugcAUAUAUacc aacacauauauauaauuaucgaUUUAUUUGAAAGAAUUAUUUAACUGGAUUAUCCGUG >NKXS01004191.1_16759-16902 CAACAAACCUAUAuaagagagagagugagagagagagagagagagagagauaaggaac cuauuuacuauuaaaauaggAAAAAUUUACCGUACACUACGAUCGUACCCAguuaaaaac ccguuuuuaacuaacguuucaacu	miR1023

Conclusões Gerais

No capítulo 2, foi apresentado a ferramenta computacional PmiR-Select (https://github.com/DeborahBambil/PmiRSelect e o seu registro como propriedade intelectual no Instituto Nacional da Propriedade Industrial - BR512022001292) e o pipeline utilizado para sua construção, a partir da mineração das sequências de pre-miRNAs (70 a 300 nt) e miRNAs (20 a 24 nt) de plantas importadas do miRBase (https://www.mirbase.org/). O primeiro filtro de tamanho das sequências, resultou em 8045 pre-miRNAs, de 2623 famílias, de 4 clados de plantas. Dessas 2623 famílias, 38 famílias altamente conservadas ocorrem em comum nos três clados (BRY, GYM e ANG) e nenhuma família de pre-miRNA dos três clados de plantas terrestres ocorreu em comum com as algas. BRY e GYM não tiveram nenhum pre-miRNA em comum. Foram realizados testes com diferentes limites de redundância (95 a 70%, intervalos de 5%) com as ferramentas skipRedundant (https://emboss.sourceforge.net/apps/release/6.1/emboss/apps/skipredundant.html), T-Coffee (https://tcoffee.crg.eu/), Inovtaxon (https://github.com/DeborahBambil/Inovtaxon), Weka (https://www.cs.waikato.ac.nz/ml/weka/) e R Project (https://www.r-project.org/). Estatisticamente, 80% foi o melhor limite para se obter um conjunto de dados representativos sem redundâncias, otimizando o processo de discriminação das cópias. A PmiR-Select foi validada com o genoma do arroz que identificou 8470 novos pre-miRNAs, homólogos a 36 famílias. Dessas, 17 são famílias existentes no miRBase para o arroz e 19 seriam de novas famílias. Isso representa um aumento de 5% de famílias de pre-miRNAs depositados no miRBase que na versão atual tem 341 famílias de pre-miRNAs do arroz. Atualmente no miRBase existem somente 84 espécies de plantas com registros pre-miRNAs. Entre elas o arroz, uma planta com muitos dados genômicos, onde mesmo assim houve um aumento de 5% no número de possíveis novas famílias de pre-miRNAs. A confirmação desses pre-miRNAs no genoma do arroz poderá abrir uma enorme oportunidade para os mais de 1000 genomas de plantas sequenciados (Published Plant Genomes (plabipd.de) e pouco explorados. Este capítulo está sendo finalizado para submissão ao periódico científico 3 BIOTECH, classificação qualis A2.

No capítulo 3, realizamos a identificação inédita de potenciais pre-miRNAs no genoma do ipê rosa (503 Mb) com a PmiR-Select, cuja análise se baseia em modelos de covariância. Foram identificados 305 pre-miRNAs de 22 famílias. Além disso, foram identificados 1293 pre-miRNAs pertencentes a 73 famílias, por meio do pipeline baseado nos modelos ocultos de Markov (hidden Markov models - HMM). Dessas 95 famílias (73 HMM

+ 22 PmiR-Select) somente uma ocorreu em comum entre os dois modelos, fortalecendo a complementaridade do uso de pelo menos dois modelos na identificação de pre-miRNas. Havendo a confirmação total ou parcial desses pre-miRNAs para o ipê rosa, o uso da PmiR-Select pode ser de grande importância para os mais de 1000 genomas sequenciados de plantas (Published Plant Genomes (plabipd.de). Da mesma forma, a análise do RNA-Seq com a PmiR-Select e o HMM irá adicionar informações genômicas importantes para o ipê rosa, juntamente com os preciosos estudos recém publicados do sequenciamento do genoma nuclear, plastidial e do RNA-Seq do ipê rosa; e das diversas análises de metabólitos específicos, filogenéticas e expressão diferencial de genes. Este capítulo está sendo finalizado para submissão em periódico científico a ser definido. Estamos finalizando, a análise do RNA-Seq sob condições de estresse de seca para submissão.

Considerações finais

No capítulo 1, foi apresentado o desenvolvimento da ferramenta computacional PmiR-Select (https://github.com/DeborahBambil/PmiRSelect), especialmente projetada para ser de fácil uso, permitindo que usuários com pouco conhecimento em linguagem de programação possam utilizá-la. Além disso, foi disponibilizado um manual completo no GitHub (https://github.com/DeborahBambil/PmiRSelect/blob/main/Guide.pdf), que fornece explicações detalhadas sobre cada etapa do processo. Nossa perspectiva em relação a este capítulo é manter a PmiR-Select constantemente atualizada, acompanhando as atualizações do miRBase juntamente com a possível integração de dados de pre-miRNAs com a base de dados Rfam (https://rfam.xfam.org/microrna). Adicionalmente, nosso próximo desafio será descrever um pipeline para a identificação de miRNAs maduros a partir dos pre-miRNAs. No entanto, enfrentamos a limitação de não haver características canônicas nas moléculas de miRNAs maduros, o que torna essa identificação mais complexa.

No capítulo 2, foi realizada a identificação de pre-miRNAs no genoma do ipê rosa, utilizando a ferramenta PmiR-Select, bem como o pipeline baseado nos modelos ocultos de Markov. Os resultados dos pre-miRNAs identificados abrirão caminhos para futuras investigações, permitindo uma melhor compreensão da biologia molecular dessa espécie e seu contexto no bioma cerrado. Essas informações podem incentivar pesquisas adicionais relacionadas a outras espécies nativas desse importante bioma. Esperamos colaborar com experimentos de bancada para validar os pre-miRNAs identificados no ipê rosa. E assim abrir possibilidades com outras espécies pouco estudadas do cerrado.

Referências

Achkar NP, Cambiagno DA, Manavella PA (2016) miRNA biogenesis: a dynamic pathway. Trends in Plant Science 21(12):1034–1044, 1

Axtell MJ (2008) Evolution of microRNAs and their targets: are all microRNAs biologically relevant? Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1779(11):725–734, 12

Axtell MJ and Meyers BC (2018) Revisiting criteria for plant microrna annotation in the era of big data. The Plant Cell 30(2):272–284, 1, 2, 9

Backes C, Fehlmann T, Kern F, et al. (2018) miRCarta: a central repository for collecting miRNA candidates. Nucleic Acids Research 46(D1):D160–D167, 7

Barah P, Winge P, Kusnierczyk A, et al. (2013) Molecular signatures in *Arabidopsis thaliana* in response to insect attack and bacterial infection. PLoS One 8(3), 2

Bhogireddy S, Mangrauthia SK, Kumar R, et al. (2021) Regulatory non-coding rnas: a new frontier in regulation of plant biology. Functional & Integrative Genomics 21:313– 330, 1

Brackett M and Earley P. (2009) The DAMA Guide to The Data Management Body of Knowl- edge (DAMA-DMBOK Guide). 2

Chen L, Heikkinen L, Wang C, et al. (2019) Trends in the development of miRNA bioinformatics tools. Briefings in Bioinformatics, 20(5), 1836-1852.

Chitarra W, Pagliarani C, Abbà S Boccacci P et al. (2018) mirvit: a novel mirna database and its application to uncover vitis responses to flavescence dorée infection. Frontiers in Plant Science 9:1034, 7

Dal BG, Franco E, Larran S e Balatti P (2015) First report of alternaria alternata causing black spot on pink lapacho (*Handroanthus impetiginosus*). Australasian Plant Disease Notes 10(1):1–2, 18

Ding D, Zhang L, Wang H et al. (2009) Differential expression of miRNAs in response to salt stress in maize roots. Annals of Botany 103(1):29–38, 2

Dong Z, Han M, Fedoroff N (2008) The RNA-binding proteins hyl1 and SE promote accurate in vitro processing of pri-miRNA by dcl1. Proceedings of the National Academy of Sciences 105(29):9970–9975, 3

Eddy SR (1998) Profile hidden Markov models. Bioinformatics (Oxford, England) 14(9), 755-763.

Edwards EJ, Osborne CP, Strömberg CAE et al. (2010) The origins of c4 grasslands: integrating evolutionary and ecosystem science. Science 328(5978):587–591, 8

Eddy SR and Durbin R (1994) RNA sequence analysis using covariance models. Nucleic Acids Research 22(11), 2079-2088.

El-Nabi SH, Elhiti M e El Sheekh M (2020) A new approach for covid-19 treatment by microRNA. Medical Hypotheses 143:110203, 12

Fei Y, Wang R, Li H et al. (2018) Dpmind: degradome-based plant miRNA–target interaction and network database. Bioinformatics 34(9):1618–1620, 7

Fujita M, Fujita Y, Noutoshi Y et al. (2006) Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signalling networks. Current Opinion in Plant Biology 9(4):436–442, 2

Guo Z, Kuang, Zheng W Ying Z et al. (2020) Pmiren: a comprehensive encyclopedia of plant miRNAs. Nucleic Acids Research 48(D1):D1114–D1121, 1, 7

Hodkinson, TR (2018) Evolution and taxonomy of the grasses (Poaceae): a model family for the study of species-rich groups. Annual Plant Reviews Online 255–294, 8

Kar MM and Raichaudhuri A (2021) Role of microRNAs in mediating biotic and abiotic stress in plants. Plant Gene 26, 100277. doi: 10.1016/j.plgene.2021.100277

Khraiwesh B, Zhu J (2012) Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms 1819(2):137–148, 2012. 2

Kim I, Jung JY, DeLuca TF et al. (2012) Cloud computing for comparative genomics with windows azure platform. Evolutionary Bioinformatics 8:EBO–S9946, 2

Kozomara A, Griffiths JS (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research 42(D1):D68–D73, 9

Lee RC, Feinbaum RL, Ambros V (1993) The c. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75(5):843–854, 1

Li F, Orban R, Baker B (2012) SoMART: a web server for plant miRNA, tasiRNA and target gene analysis. The Plant Journal 70(5):891–901, 5

Li M and Yu B (2021) Recent advances in the regulation of plant miRNA biogenesis. RNA biology 18(12), 2087-2096. doi: 10.1080/15476286.2021.1899491

Liu H, Tian X, Li Y, et al. (2008) Microarray-based analysis of stress-regulated microRNAs in *Arabidopsis thaliana*. RNA 14(5):836–843, 2008. 2

Liu J, Liu X, Zhang S et al. (2021) Tardb: an online database for plant miRNA targets and miRNA-triggered phased siRNAs. BMC Genomics 22(1):1–12, 3, 7

Meng Y, Shao C, Wang H, Chen M (2012) Are all the miRBase-registered microRNAs true? A structure-and expression-based re-examination in plants. RNA Biology 9(3), 249-253.

Millar AA, Lohe A, Wong G (2019) Biology and function of miR159 in plants. Plants 8(8), 255.

Morgado L and Johannes F (2019) Computational tools for plant small RNA detection and categorization. Briefings in Bioinformatics 20(4):1181–1192, 12

Nawrocki EP and Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22), 2933-2935

Olivieira L, Vitor M, Érica ES et al. (2019) *Handroanthus impetiginosus* (Bignoniaceae) as an important floral resource for synanthropic birds in the brazilian semiarid. Oecologia Australis 23(1). 18

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. Cell 136(4):629–641, 1

Raza A, Charagh S, Karikari B, et al. (2023).miRNAs for crop improvement. Plant Physiology and Biochemistry (201) doi: 107857.10.1016/j.plaphy.2023.107857

Reinhart BJ, Weinstein EG, Rhoades MW et al. (2002) microRNAs in plants. Genes & Development 16(13):1616–1626, 1

Sablok G, Milev I, Minkov G et al. (2013) isomiRex: Web-based identification of microRNAs, isomiR variations and differential expression using next-generation sequencing datasets. FEBS Letters 587(16):2629–2634, 12

Satish D and Mukherjee SKGD (2019) Pamirdb: a web resource for plant miRNAs targeting viruses. Scientific Reports 9(1):1–6, 7

Sattar S, Song Y, Anstead JA et al. (2012) *Cucumis melo* microRNA expression profile during aphid herbivory in a resistant and susceptible interaction. Molecular Plant Microbe Interactions 25(6):839–848, 2

Silva JOB, Grattapaglia D, Novaes E, et al. (2018) Genome assembly of the pink ipê (*Handroanthus impetiginosus*, Bignoniaceae), a highly valued, ecologically keystone neotropical timber forest tree. Gigascience 7(1):gix125, 18

Silva-Junior OB, Grattapaglia D, Novaes E, Collevatti RG (2018) Genome assembly of the pink ipê (*Handroanthus impetiginosus*, Bignoniaceae), a highly valued, ecologically keystone neotropical timber forest tree. Gigascience 7(1), gix125. doi: 10.1093/gigascience/gix125

Sobreiro MB, Collevatti RG, Dos Santos YL, et al. (2021) RNA-Seq reveals different responses to drought in Neotropical trees from savannas and seasonally dry forests. BMC Plant Biology 21(1), 1-17. doi: 10.1186/s12870-021-03244-7

Souza JMA, Tarso BS, Paulo D et al. (2020) Longitudinal distribution of lapachol in the stalk of ipê species (*Handroanthus spp*). European Journal of Wood and Wood Products 78(3):609–611, 2020. 18

Stocks MB, Moxon S, Mapleson D et al. (2012) The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. Bioinformatics 28(15):2059–2061, 5

Strassburg BBN, Brooks T, Feltran BR, et al. (2017) Moment of truth for the cerrado hotspot. Nature Ecology & Evolution 1(4):1–3, 2017. 18

Sun X, Zhang Y, Zhu X et al. (2014) Advances in identification and validation of plant microRNAs and their target genes. Physiologia Plantarum 152(2):203–218, 12

Teng Y, Xu F, Zhang X (2021) Plant-derived exosomal microRNAs inhibit lung inflammation induced by exosomes sars-cov-2 nsp12. Molecular Therapy 29(8):2424–2440, 12

Triguero I, García-Gil D, Maillo J et al. (2019) Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(2):e1289, 2

Wang W, Vinocur B, Altman A (2003) Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. Planta, 218(1):1–14, 2, 2

Wang N, Zhang J, Xiao B, et al. (2022) Recent advances in the rapid detection of microRNA with lateral flow assays. Biosensors and Bioelectronics, 211, 114345.

Wink M (2013) Evolution of secondary metabolites in legumes (Fabaceae). South African Journal of Botany 89:164–175, 8

Yang-Turner F, Gripper L, Swann J et al. (2018) An open-source Azure solution for scalable genomics workflows. IEEE World Congress on Services 39–40, 2

Yu D, Lu J, Shao W et al. (2019) Mepmirdb: a medicinal plant microRNA database. Database 7

Zhang B, Pan X, Cannon CH, et al. (2006) Conservation and divergence of plant microRNA genes. The Plant Journal 46(2), 243-259. doi: 10.1111/j.1365-313X.2006.02697.x

Zhang P, Meng XC, Hongjun L et al. (2017) Plantcircnet: a database for plant circRNA–miRNA–mRNA regulatory networks. Database 7

Yu T, Xu N, Haque N et al. (2020) Popular computational tools used for miRNA prediction and their future development prospects. Interdisciplinary Sciences: Computational Life Sciences, 12, 395-413.

Zhang Z, Yu J, Li D, et al. (2010) PMRD: plant microRNA database. Nucleic Acids Research 38(suppl_1), D806-D813.