Universidade Estadual de Londrina

ALAN PÉRICLES RODRIGUES LORENZETTI

ANÁLISE EM LARGA ESCALA DE PEQUENOS RNAs DERIVADOS DE ELEMENTOS TRANSPONÍVEIS EM GENOMAS VEGETAIS

Londrina
2016

ALAN PÉRICLES RODRIGUES LORENZETTI

ANÁLISE EM LARGA ESCALA DE PEQUENOS RNAs DERIVADOS DE ELEMENTOS TRANSPONÍVEIS EM GENOMAS VEGETAIS

Dissertação apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito à obtenção do título de Mestre em Genética e Biologia Molecular.

Orientador: Prof. Dr. Douglas Silva Domingues

Coorientador: Prof. Dr. Alexandre Rossi Paschoal

Londrina
2016

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

Lorenzetti, Alan Péricles Rodrigues.
Análise em larga escala de pequenos RNAs derivados de elementos transponíveis em genomas vegetais / Alan Péricles Rodrigues Lorenzetti. - Londrina, 2016.
98 f. : il.

Orientador: Douglas Silva Domingues.
Coorientador: Alexandre Rossi Paschoal.
Dissertação (Mestrado em Genética e Biologia Molecular) - Universidade Estadual de Londrina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Genética e Biologia Molecular, 2016.
Inclui bibliografia.

1. Elementos transponíveis - Teses. 2. Pequenos RNAs - Teses. 3. Banco de dados - Teses. I. Domingues, Douglas Silva. II. Paschoal, Alexandre Rossi. III. Universidade Estadual de Londrina. Centro de Ciências Biológicas. Programa de Pós-Graduação em Genética e Biologia Molecular. IV. Título.

ALAN PÉRICLES RODRIGUES LORENZETTI

ANÁLISE EM LARGA ESCALA DE PEQUENOS RNAs DERIVADOS DE ELEMENTOS TRANSPONÍVEIS EM GENOMAS VEGETAIS

Dissertação apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular, da Universidade Estadual de Londrina, como requisito à obtenção do título de Mestre em Genética e Biologia Molecular.

BANCA EXAMINADORA

_____

Orientador: Prof. Dr. Douglas Silva Domingues
Universidade Estadual de Londrina – UEL

_____

Coorientador: Prof. Dr. Alexandre Rossi Paschoal
Universidade Tecnológica Federal do Paraná - UTFPR

_____

Prof. Dr. André Luís Laforga Vanzela
Universidade Estadual de Londrina – UEL

_____

Profa. Dra. Elaine Silva Dias
Universidade Estadual Paulista Júlio de Mesquita Filho – UNESP
São José do Rio Preto

Londrina, 29 de Fevereiro de 2016.

Dedico este trabalho aos cientistas não lembrados pela história e àqueles que nos tempos antigos e recentes tiveram suas vidas perturbadas, ameaçadas, e até mesmo ceifadas, simplesmente por buscar a verdade através do método científico.

Agradecimentos

Os agradecimentos principais são direcionados aos meus pais, Renato e Izolina, sem os quais eu não poderia ter chegado até aqui. Também agradeço a minha namorada, Katiuska, por estar presente nos meus dias durante esses últimos anos.

Além disso, gostaria de agradecer ao meu orientador Dr. Douglas Silva Domingues, que não poupou esforços para o desenvolvimento de minha pesquisa, e ainda, por disponibilizar seu apartamento para locação durante o último ano. Agradeço também ao Dr. Alexandre Rossi Paschoal, que me coorientou durante esta pesquisa e me forneceu transporte para Cornélio Procópio diversas vezes ao longo do mestrado.

Uma parte do trabalho também não seria realizada sem o empenho e as centenas de linhas de código escritas pelo Gabriel Yuri Alves de Antonio, que desenvolveu o *front-end* e o *back-end* do PlanTE-MIR DB. Outro grande contribuinte para o desenvolvimento deste trabalho foi o Dr. Romain Guyot, pesquisador do *Institute de recherche pour le développement*, localizado em Montpellier, na França. Os ensinamentos de Romain foram essenciais para o desenvolvimento deste trabalho.

Agradeço ainda à comunidade GNU/Linux e a todos os desenvolvedores que contribuem para a *Open Source Initiative*. Grande parte dos programas utilizados neste trabalho são frutos desses projetos. Por fim, agradeço à Equipe abnTEX2[1], desenvolvedora da suíte homônima que facilita a formatação de textos acadêmicos de acordo com as normas da ABNT.

[1] <http://www.abntex.net.br/>.

"Somewhere,
something incredible
is waiting to be known".
(Carl Sagan)

LORENZETTI, Alan Péricles Rodrigues. **Análise em larga escala de pequenos RNAs derivados de elementos transponíveis em genomas vegetais.** 2016. 98f. Dissertação (Mestrado em Genética e Biologia Molecular) – Universidade Estadual de Londrina, Londrina, 2016.

RESUMO

Os elementos transponíveis (TEs) constituem uma grande parte dos genomas eucariotos e desempenham grande importância em sua evolução. Esse componente do genoma pode atuar de diversas maneiras, seja inativando genes, alterando a expressão deles, aumentando seu número de cópias, ou até mesmo criando novas sequências nucleotídicas capazes de influenciar no funcionamento do organismo. Nesse último caso, sabe-se que os TEs podem originar *loci* responsáveis pela transcrição de RNAs não codificantes funcionais (ncRNAs), como microRNAs (miRNAs), os quais podem atuar na regulação pós-transcricional de genes. Com base nessas informações, os objetivos da primeira parte do trabalho foram a anotação de TEs de 15 genomas vegetais utilizando métodos de busca baseados em similaridade, e a procura por interseções posicionais com precursores de miRNAs anotados. Os resultados obtidos permitiram a criação do *Plant Transposable Element-related microRNAs Database* (PlanTE-MIR DB, disponível em <http://bioinfo-tool.cp.utfpr.edu.br/plantemirdb/>), que agrega 152 TE-MIRs para 10 espécies vegetais e facilita o acesso às anotações por meio de uma interface amigável, disponibilizando múltiplos formatos de arquivo. Boa parte desses *loci* produtores de pequenos ncRNAs estão relacionados a um tipo específico de TEs: os elementos transponíveis de repetição invertida em miniatura (MITEs, do inglês *Miniature Inverted-Repeat Transposable Elements*). Visando explorar em maior profundidade essa relação, o objetivo da segunda parte do trabalho foi realizar a busca por pequenos RNAs relacionados aos MITEs no genoma de *Coffea canephora*. A análise revelou que aproximadamente 1,5% do genoma da espécie é composto por MITEs e que a maior parte das inserções está localizada em regiões ricas em genes. Além disso, foram encontradas e classificadas 44 famílias relacionadas a pequenos RNAs, sendo pelo menos uma delas representada em ESTs. A maior parte dos pequenos RNAs associados a essas famílias são de 24 nt, sugerindo a participação desses elementos na biogênese de siRNAs. Esses dados trazem importantes contribuições para a compreensão da evolução genômica em plantas, fornecendo informações sobre a inter-relação entre os seus diversos componentes.

Palavras-chave: Elementos transponíveis. Pequenos RNAs. Genomas vegetais. Banco de dados.

LORENZETTI, Alan Péricles Rodrigues. **Large scale analysis of small RNAs derived from transposable elements in plant genomes.** 2016. 98p. Dissertation (Master's degree in Genetics and Molecular Biology) – Universidade Estadual de Londrina, Londrina, 2016.

ABSTRACT

Transposable elements (TEs) comprise a major fraction of eukaryotic genomes and they are known to drive their evolution by several mechanisms. They can act inactivating genes, changing the expression levels, raising their copy number, or even creating new nucleotide sequences. In this context, sometimes they are responsible for shaping new functional non-coding RNA loci (e.g. microRNAs), which may participate in post-transcriptional gene regulation process. The first part of this work had as main objective the similarity search-based TE annotation for 15 plant genomes in order to find positional intersections with annotated miRNAs. We assembled these findings in the Plant Transposable Elementrelated miRNAs Database (PlanTE-MIR DB), hosted at <http://bioinfo-tool.cp.utfpr.edu.br/plantemirdb/>. This database has 152 TE-MIR annotations for 10 plant species in a user-friendly web interface, providing annotation data using several file formats. Many ncRNA loci producing small RNAs are related to a specific TE type: the Miniature Inverted-Repeat Transposable Elements (MITEs). Considering this, the second part of this work had as main objective the investigation of MITE-associated small RNAs in the *Coffea canephora* genome. We observed that 1,5% of this genome is composed by MITEs. We also found genome-wide association between the MITE and exon densities. Moreover, 44 MITE families were associated to small RNAs, and at least one family is represented in EST contig data. Most small RNAs are 24 nt, suggesting participation of MITEs in the siRNA biogenesis pathway. These data bring important insights for the comprehension of plant genomes evolution, providing information about the relationship of their different components.

Keywords: Transposable elements. Small RNAs. Plant genomes. Database. MITEs.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

Parte I

Introdução

1 Elementos transponíveis

Os trabalhos pioneiros de Barbara McClintock na década de 40 sugeriram a existência de elementos genéticos móveis capazes de controlar outros *loci*, mas suas descobertas foram consideradas controversas e misteriosas pela comunidade científica da época (JONES, 2005). Atualmente, os mesmos elementos móveis, que outrora foram considerados parte do "DNA lixo" (BIÉMONT, 2010), estão amplamente difundidos como elementos transponíveis (TEs, do inglês *Transposable Elements*) e são matéria de interesse em seus mais diversos aspectos como origem, função, regulação e demais particularidades (BENNETZEN; WANG, 2014).

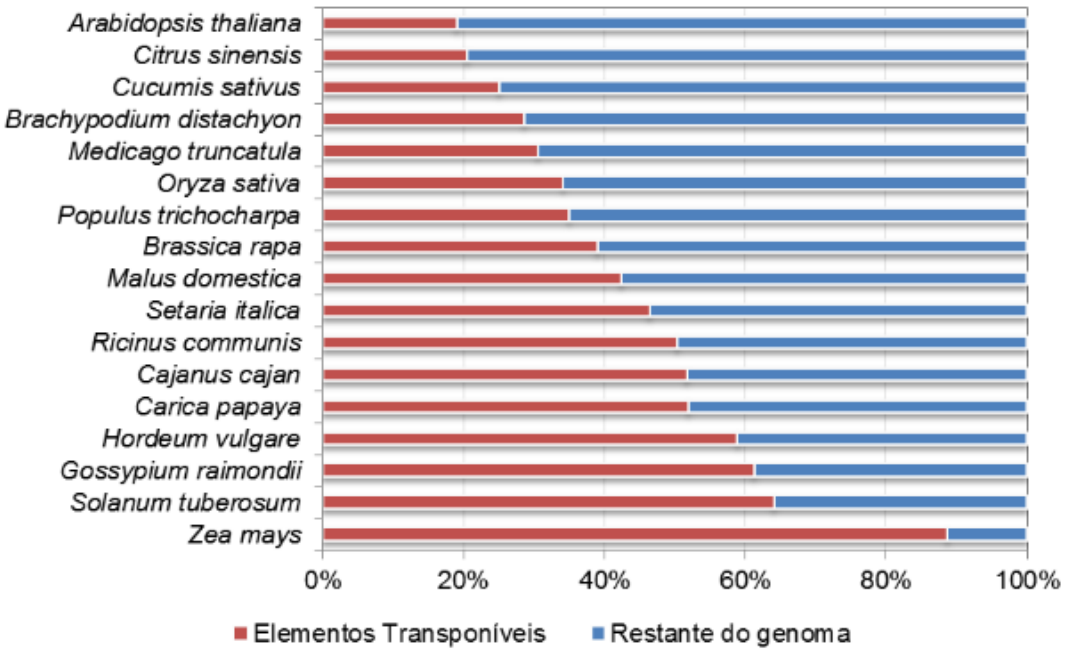
Os TEs estão presentes nos genomas de organismos eucariotos (WICKER et al., 2007) e nos vegetais podem constituir grande parte do material genético de um indivíduo (Figura 1). Além disso, nota-se sua importância em processos relacionados à inativação, criação e movimentação de genes, rearranjos cromossômicos, modulação da expressão gênica e silenciamento epigenético (LISCH, 2013; BENNETZEN; WANG, 2014). Eventos como a poliploidização, e os estresses bióticos e abióticos têm sido relacionados com a ativação desses elementos, induzindo grandes alterações no genoma do hospedeiro (PARISOD et al., 2010; LISCH, 2013). TEs também são conhecidos por proporcionar o surgimento de *loci* responsáveis pela transcrição de pequenos RNAs não codificantes, como microRNAs (miRNAs), pequenos RNAs de interferência RNAs (siRNAs), além de longos RNAs não codificantes (lncRNAs) (PIRIYAPONGSA; JORDAN, 2008; LI et al., 2011; HADJIARGYROU; DELIHAS, 2013; GIM et al., 2014).

Devido à sua relevância genômica, houve um grande crescimento do volume de dados de identificação e anotação de TEs gerados para diversos genomas, trazendo a necessidade de criar uma classificação unificada para esses elementos repetitivos. Assimilou-se princípios e nomenclaturas inicialmente propostos por Finnegan (1989) às descobertas mais recentes, como detalhes de seu mecanismo de transposição e características estruturais, e assim novos critérios foram estabelecidos (WICKER et al., 2007). De maneira ampla, os TEs são divididos em elementos de Classe I e II.

1.1 Elementos de Classe I

Os elementos de Classe I, ou retrotransposons, são capazes de copiar-se para outras regiões genômicas por intermédio de transcrição em ácido ribonucleico (RNA) e uma posterior transcrição reversa que se antecede à integração do novo fragmento (Figura 2). Esse evento é usualmente referido como mecanismo "copia-e-cola". Geralmente, as enzimas envolvidas no processo são codificadas pelo próprio retrotransposon (SCHULMAN, 2013).

Figura 1 – Proporção do conteúdo de TEs nos genomas de algumas espécies vegetais.



Fonte: Adaptado de Ragupathy, You e Cloutier (2013).

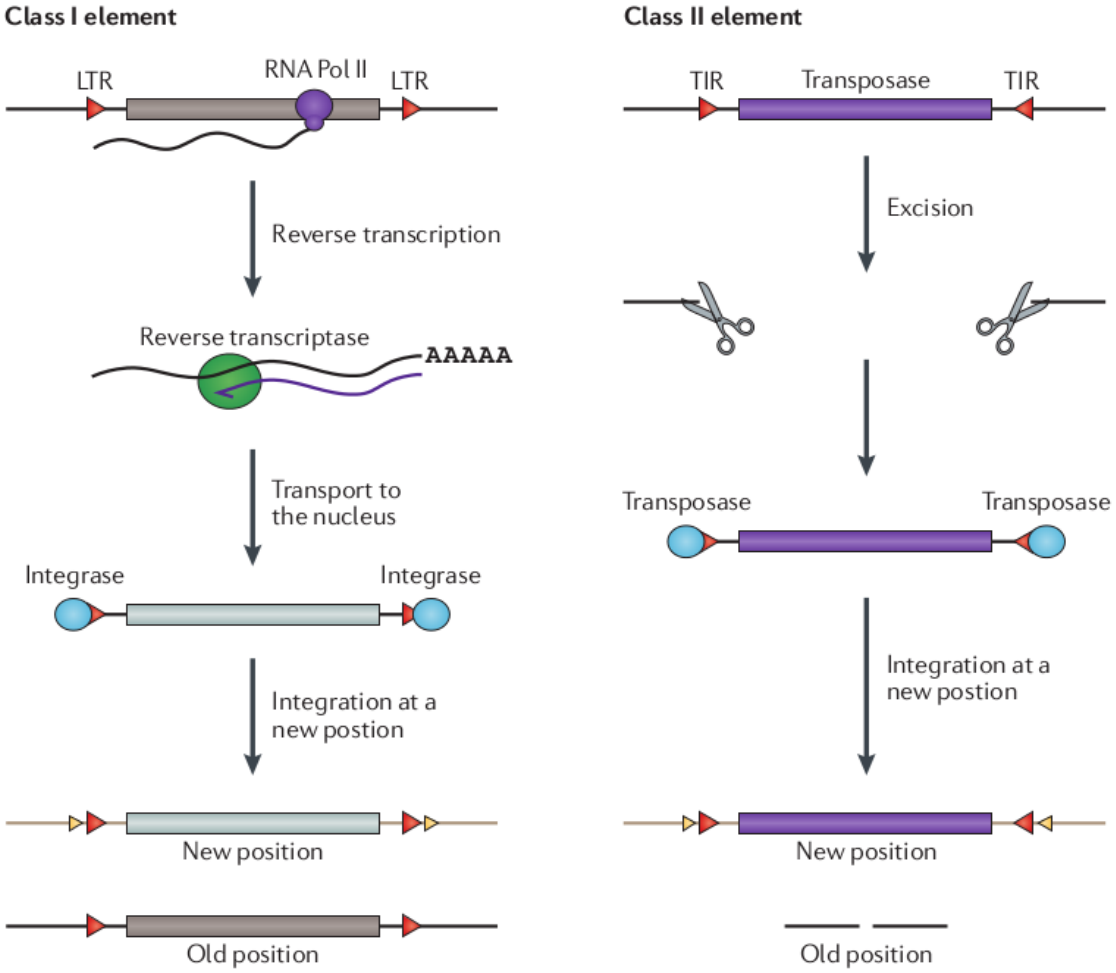
A Classe I é dividida em cinco ordens principais: *Long Terminal Repeat* (LTR), *Dictyostelium Intermediate Repeat Sequence* (DIRS), *Penelope-like Elements* (PLE), *Long Interspersed Nuclear Elements* (LINE) e *Short Interspersed Nuclear Elements* (SINE). Cada uma delas apresenta especificidades não só quanto a presença, tipo e tamanho de suas regiões terminais, mas também de acordo com a sequência codificante que carregam. O mesmo vale, em muitos casos, para suas subdivisões denominadas superfamílias (WICKER et al., 2007).

Os retrotransposons da ordem LTR (LTR-RTs) estão presentes abundantemente em plantas e podem chegar a milhões de cópias em um único indivíduo haploide (KUMAR; BENNETZEN, 1999). As principais superfamílias dessa ordem, no contexto dos eucariotos não metazoários, são *Gypsy* e *Copia* (Figura 3). A diferença mais notável entre elas é a alteração da disposição dos domínios da transcriptase reversa e da integrase (WICKER et al., 2007).

LTR-RTs são transcritos pela RNA polimerase II (Pol II), que reconhece um sítio promotor na LTR 5'. A LTR 3' carrega sinais para a terminação da transcrição e também para poliadenilação. Após ser transcrito, os LTR-RTs são exportados para o citoplasma onde suas regiões codificantes são traduzidas para as proteínas necessárias para o seu processo de integração. Esse processo é altamente complexo e envolve a internalização do transcrito em uma partícula semelhante a vírus onde ocorre sua transcrição reversa.

O cDNA gerado é então importado para o núcleo celular e integrado a um novo sítio (SCHULMAN, 2013).

Figura 2 – Mecanismos de transposição de retrotransposons e transposons.



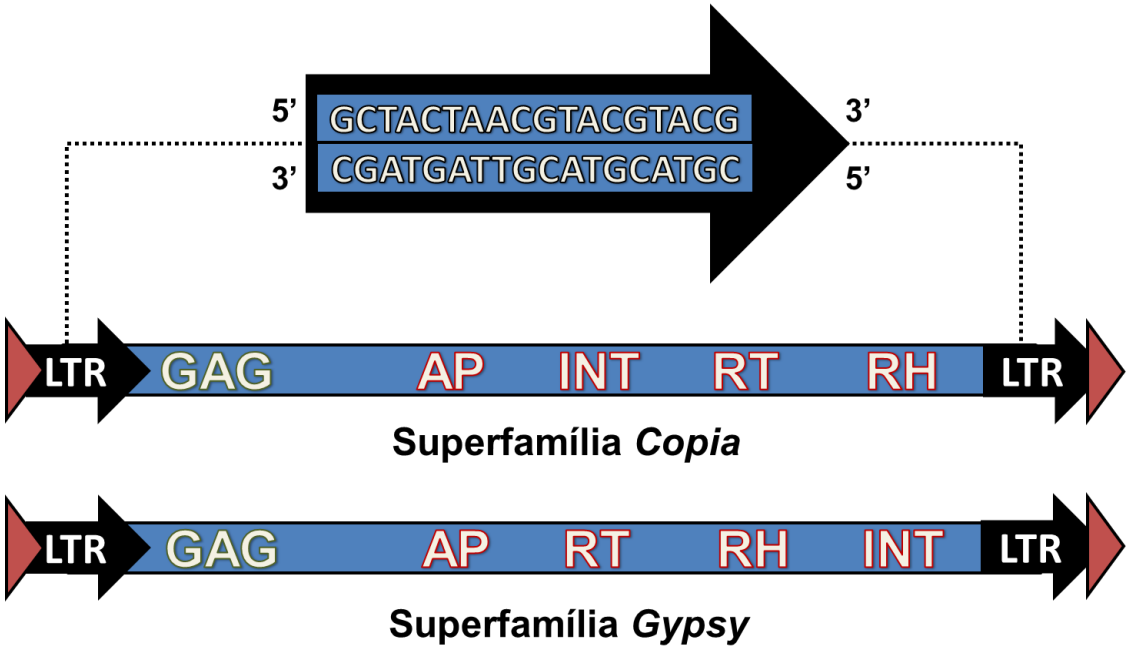
Fonte: Lisch (2013).

1.2 Elementos de Classe II

Os elementos de Classe II, ou transposons, diferenciam-se dos retrotransposons principalmente pelo seu mecanismo de transposição. Esse mecanismo consiste basicamente em sua autoexcisão com auxílio de enzimas, e inserção em uma nova região do genoma, fenômeno chamado de "corta-e-cola" (Figura 2). De maneira geral, carregam a capacidade de produção da proteína transposase, mediadora do processo (LISCH, 2013).

Essa classe se divide em duas subclasses principais, de acordo com a classificação de Wicker et al. (2007). Cada uma delas difere pela quantidade de fitas clivadas durante a transposição. Na Subclasse I, as duas fitas são clivadas e na Subclasse II, apenas uma

Figura 3 – Representação dos LTR-RTs das superfamílias *Copia* e *Gypsy*. O quadro de leitura aberto GAG (Grupo antígenos, contorno em verde) codifica proteínas estruturais de partículas semelhantes a vírus e o quadro de leitura aberto POL (Poliproteína, contorno em vermelho) contém as enzimas proteinase aspártica (AP), integrase (INT), transcriptase reversa (RT) e RNase H (RH). Cada seta preta apresentada nas regiões terminais significa a presença de uma repetição terminal longa (LTR, do inglês *Long Terminal Repeat*) e cada ponta de seta vermelha representa uma duplicação do sítio alvo (TSD, do inglês *Target Site Duplication*).



Fonte: Adaptado de Wicker et al. (2007) e Pierce (2010).

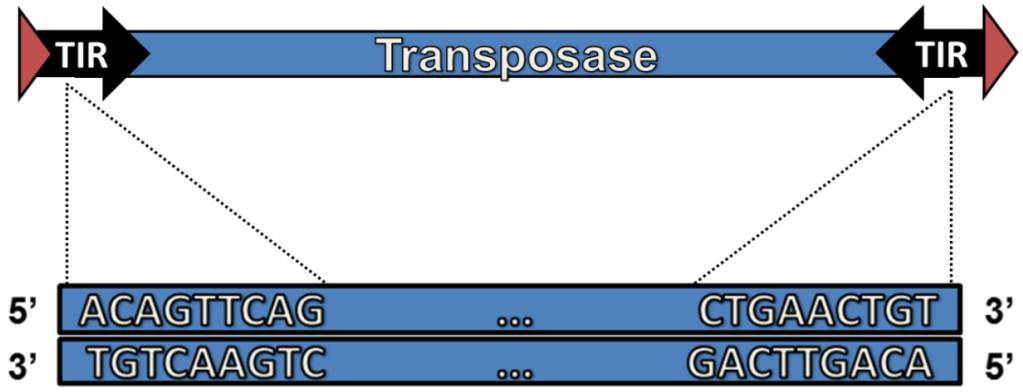
sofre o corte. A primeira contém as ordens *Terminal Inverted Repeat* (TIR) e *Crypton*, e a segunda tem como representantes as ordens *Helitron* e *Maverick*.

As superfamílias da ordem TIR são *Tc1-Mariner*, *hAT*, *Mutator*, *Merlin*, *Transib*, *P*, *PiggyBac*, *PIF-Harbinger* e *CACTA*, e são discriminadas não só de acordo com suas duplicações do sítio alvo, geradas durante sua inserção, mas também pelo padrão encontrado em suas repetições invertidas terminais. Além disso, podem ter domínios proteicos característicos (Figura 4) (WICKER et al., 2007).

A ordem Crypton possui apenas uma superfamília de mesmo nome. Os elementos dessa ordem possuem a enzima tirosina recombinase e produzem duplicações do sítio alvo quando se inserem em um novo sítio (WICKER et al., 2007).

A ordem Helitron também possui apenas uma única superfamília homônima. Os elementos dessa ordem são replicados por meio da circularização de uma das fitas que é

Figura 4 – Representação das repetições terminais invertidas (TIRs, do inglês *Terminal Inverted Repeats*) flanqueando o quadro de leitura aberto da transposase. Setas pretas representam TIRs e as pontas de seta vermelhas representam TSDs.



Fonte: Adaptado de Pierce (2010).

clivada, e codificam a enzima tirosina recombinase. Essa enzima contém o domínio helicase e a capacidade de engatilhar a atividade de replicação. Helitrons frequentemente levam genes adjacentes consigo durante a transposição (WICKER et al., 2007).

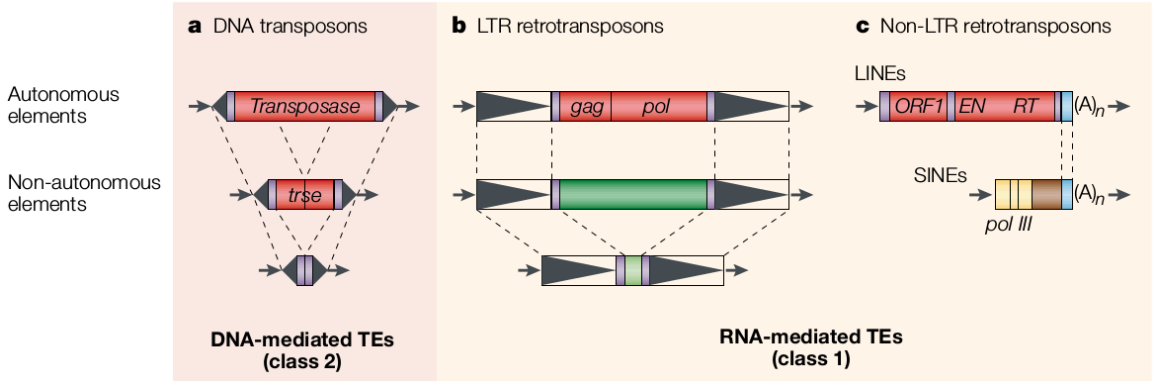
Embora flanqueados por TIRs, alguns elementos de até 20 Kb são incluídos na ordem Maverick (também chamada de Polinton). Elementos dessa ordem podem codificar até 11 proteínas, entre as quais estão a Polimerase B e integrase semelhante àquelas encontradas nos retrotransposons. Entretanto, por não codificarem transcriptase reversa, seu mecanismo de transposição se dá sem intermediários de RNA (WICKER et al., 2007).

1.3 Elementos não autônomos

Alternativamente, elementos das duas classes podem ainda perder domínios codificantes ou terem suas sequências alteradas (Figura 5), implicando no surgimento de códons de terminação prematura. Isso causa a incapacidade de transposição autônoma da sequência, que fica dependente de outros elementos intactos capazes de codificar as enzimas necessárias para o processo (WICKER et al., 2007). Exemplos de elementos não autônomos são demonstrados na Tabela 1.

No caso dos transposons, elementos não autônomos podem ser mobilizados pelas enzimas codificadas por elementos autônomos relacionados. O número de cópias de elementos não autônomos ultrapassa rapidamente o número de elementos autônomos, havendo assim competição dos dois tipos pelas mesmas enzimas de transposição. Com o aumento dessas sequências, mecanismos de defesa inatos podem ser induzidos a inativar as cópias autônomas. Como consequência acontece a "morte" de determinada família

Figura 5 – Origem de elementos não autônomos a partir de seus ancestrais autônomos. Nota-se a perda gradativa dos domínios codificantes para cada categoria de TEs.



Fonte: Feschotte, Jiang e Wessler (2002).

Tabela 1 – Exemplos de elementos não autônomos.

Classe	Acrônimo	Descrição
Classe I	LARD	Derivativo de grande retrotransposon
	TRIM	Miniatura de retrotransposons com repetição terminal
Classe II	MITE	Elemento transponível de repetição invertida em miniatura
	SNAC	Pequeno transposon não autônomo CACTA

Fonte: Adaptado de Wicker et al. (2007).

de transposons (FESCHOTTE; PRITHAM, 2007). Por outro lado, MITEs (do inglês *Miniature Inverted-Repeat Transposable Elements*) podem continuar aumentando em número de cópias mesmo sem ter algum elemento relacionado ativo. Isso ocorre por meio do mecanismo de mobilização cruzada, como no caso dos *Stowaways* de *Oryza sativa*, que utilizam transposases de elementos *Osmar* da superfamília *Tc1-Mariner* para se transpor (YANG et al., 2009).

1.4 Papel dos TEs na evolução dos genomas

TEs estão envolvidos em modificações que incluem a inativação gênica, introdução de novas funções, mudanças na estrutura gênica, mobilização e rearranjo de fragmentos gênicos, e silenciamento epigenético (LISCH, 2013). São também responsáveis pela grande variação no tamanho dos genomas ao lado das variações promovidas por ploidias (BENNETZEN; WANG, 2014).

A inativação de genes pode ser promovida por inserções de TEs em um gene

codificante. O fenótipo das ervilhas rugosas de Mendel é o exemplo mais clássico. A inserção de um transposon em meio ao gene codificador da enzima ramificadora de amido é a responsável por sua inativação. A falta dessa enzima faz com que os açúcares não sejam polimerizados normalmente, alterando o fenótipo das ervilhas (BHATTACHARYYA et al., 1990).

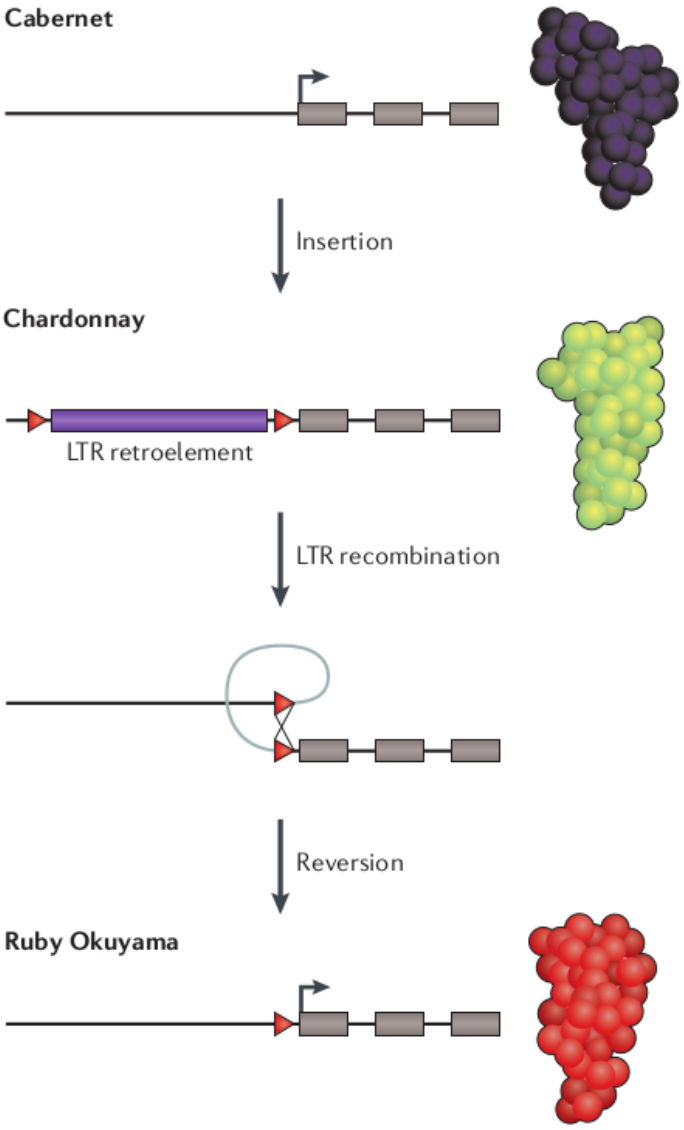
Além disso, a inserção de TEs em regiões regulatórias de genes pode ocasionar a alteração do seu padrão transcricional. O fenótipo das uvas *Cabernet*, *Chardonnay* e *Ruby Okuyama* foi influenciado por modificações desse tipo. A inserção de um retrotransposon próximo à região promotora do gene Vvmby1A em uvas *Cabernet* causou a perda de sua cor, originando a variedade *Chardonnay*. A recombinação deste mesmo retrotransposon, deixando apenas sua região LTR, promoveu a alteração do fenótipo novamente, originando a variedade *Ruby Okuyama* (Figura 6). Os TEs também podem ser inseridos em acentuassomos (do inglês, *enhancers*), repressores e trazer novos sítios promotores, modificando assim o padrão transcricional de determinado gene (LISCH, 2013).

Inversões e translocações de trechos cromossômicos são apontados como possíveis modificações induzidas por TEs. Elas podem ocorrer quando as regiões terminais de dois TEs adjacentes, separados por trechos de até 100 Kb, são reconhecidas pelas enzimas de transposição como se fossem de um único elemento e levadas para um novo sítio de inserção (FESCHOTTE; PRITHAM, 2007). Caso similar acontece na translocação de pequenos trechos em solanáceas. Um trabalho identificou diversos membros de uma mesma família de MITEs (*MiS5*) formando complexos e movendo as sequências localizadas entre dois elementos para outras regiões do genoma (KUANG et al., 2009).

Muitos estudos demonstraram a "domesticação" de regiões codificantes de transposons por meio de sua fusão com regiões codificantes de um gene hospedeiro, como é o caso dos fatores de transcrição FAR1 e FHY3, aparentemente envolvidos em processos de resposta à luz. Esses fatores de transcrição são derivados de transposases de elementos de Classe II, que possivelmente contribuem para o reconhecimento de sítios promotores no DNA (BENNETZEN; WANG, 2014). Embora seja muito pequena a chance de algo do tipo acontecer, essas modificações induzidas por TEs podem ser de grande valor para evolução dos genes. A analogia do relógio mostra que uma mutação em um gene é como enfiar um prego em um relógio. Há grande chance de o relógio ser destruído, mas em algum momento isso pode promover a origem de uma nova função (Figura 7a). Uma outra forma desta analogia mostra que diversos trechos funcionais de TEs podem ser assimilados, produzindo algo totalmente novo e funcional, como se fossem aproveitadas peças de diversos relógios para construir um novo dispositivo. Da mesma forma que no primeiro caso, há grandes chances de isso dar errado, produzindo regiões não funcionais (Figura 7b).

TEs geralmente apresentam-se metilados. O metiloma de *Arabidopsis thaliana* mostra que regiões repetitivas (compostas principalmente por TEs) correspondem às

Figura 6 – Influência da inserção de TEs em região regulatória de fator de transcrição em *Vitis vinifera*.

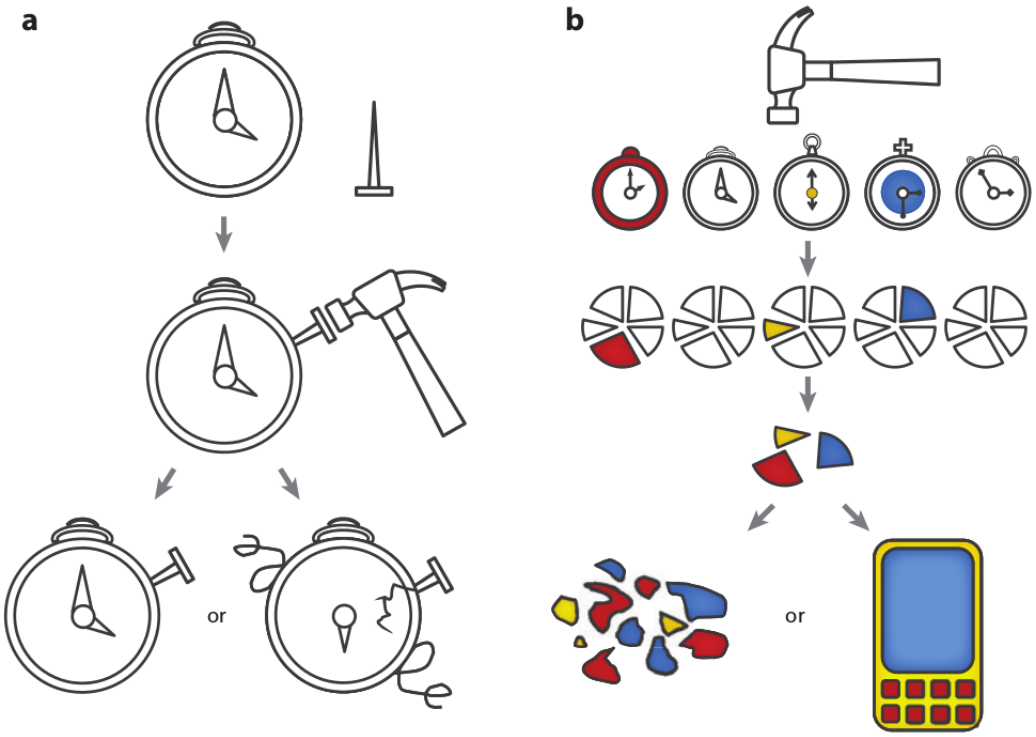


Fonte: Lisch (2013).

regiões de metilação de sítios CG, CHG e CHH (onde H pode ser C, T ou A) (Figura 8) (MIROUZE; VITTE, 2014). Com a metilação *de novo* de TEs pode ocorrer o silenciamento epigenético de genes adjacentes, o que contribui para alteração do seu padrão transcricional (LISCH, 2013).

Além de estresses bióticos e abióticos (LISCH, 2013), eventos de alopoliploidização podem provocar a ativação de TEs no genoma (PARISOD et al., 2010). Em vista das grandes modificações provocadas por esses elementos, pode haver o surgimento de novas características nas populações submetidas a grandes mudanças ambientais, e consequentemente proporcionar sua sobrevivência.

Figura 7 – Analogia do relógio. Genes podem adquirir novas funções a partir de mutações ou inserções de TEs.



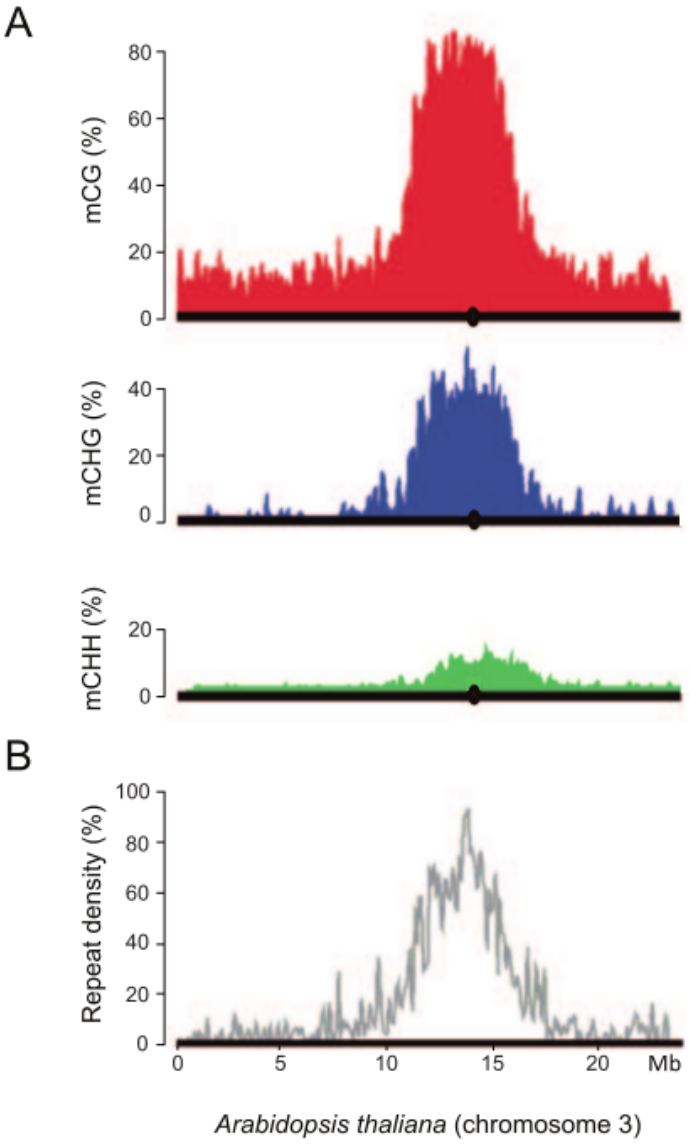
Fonte: Bennetzen e Wang (2014).

1.5 Ferramentas utilizadas para anotação de TEs

No início das análises genômicas, para facilitar o estudo das regiões codificadoras, foi adotado o processo de mascaramento de regiões repetitivas, o qual ignorava repetições simples, de baixa complexidade e também TEs. Esse processo sempre foi dependente de bancos de dados - e.g. Repbase Update (JURKA et al., 2005) - que continham esses tipos de sequências. Com o tempo esses repositórios foram aprimorados, permitindo que a informação lá armazenada servisse não apenas para o mascaramento, mas também para a identificação e anotação de elementos repetitivos (JANICKI; ROOKE; YANG, 2011). Pode-se citar aqui dois programas baseados em busca por similaridade que surgiram para esse fim: RepeatMasker (SMIT; HUBLEY; GREEN, 1996) e CENSOR (JURKA et al., 2005).

Esses métodos são eficientes para anotação de TEs codificadores de proteínas, mesmo que sejam evolutivamente distantes. Entretanto, existem elementos não autônomos que podem não ser detectados devido à falta de domínios codificadores (JANICKI; ROOKE; YANG, 2011). Ademais, esta abordagem é capaz de identificar somente sequências parecidas com aquelas que já são conhecidas, subestimando assim o conteúdo total de TEs em um

Figura 8 – Perfil de metilação do cromossomo 3 de *A. thaliana*. São mostrados os níveis de metilação em sítios CG (mCG, em vermelho), CHG (mCHG, em azul) e CHH (mCHH, em verde) (**A**). É mostrada também a densidade de repetições para o mesmo cromossomo (**B**). As regiões repetitivas são evidentemente mais metiladas do que as outras regiões. Um círculo preto aponta a região do centrômero.

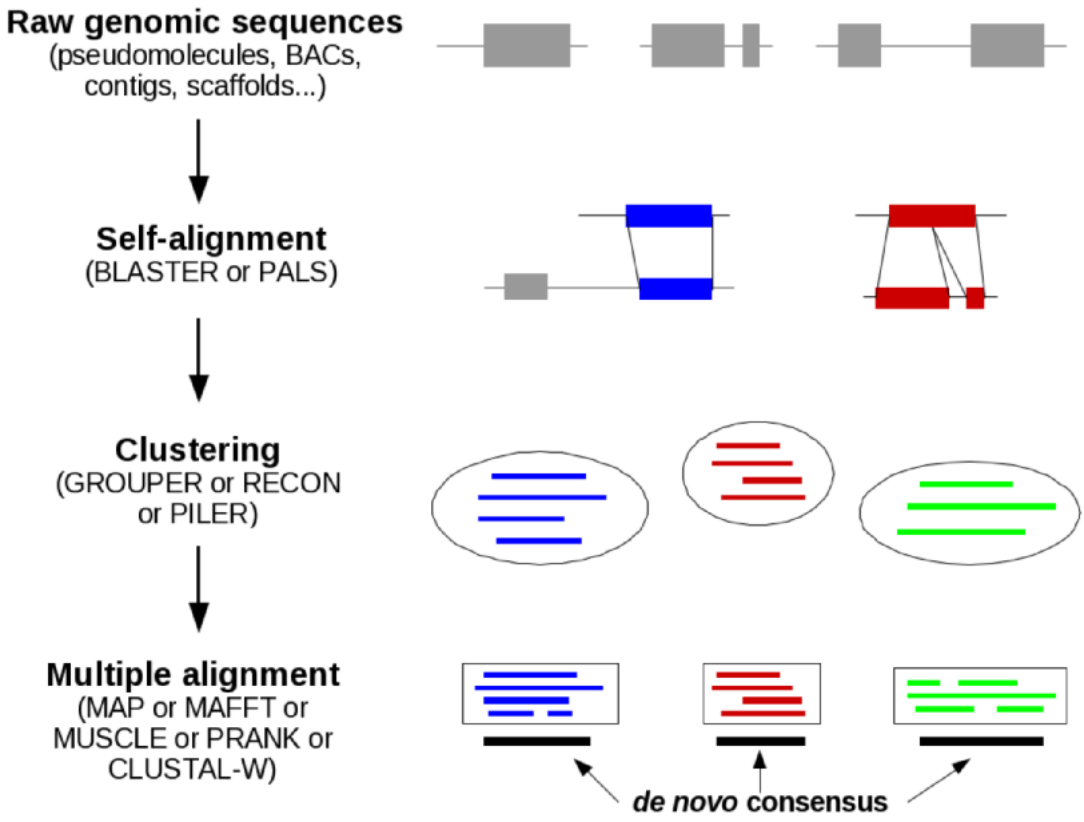


Fonte: Mirouze e Vitte (2014).

organismo. Outra armadilha ocasionada pelas análises usando ferramentas de similaridade é a possibilidade de atribuir-se funções erroneamente para proteínas. Esse tipo de análise leva em conta apenas a similaridade da estrutura primária entre duas sequências, sem saber se os seus domínios funcionais são realmente parecidos. Funções inferidas por similaridade de sequência (ISS) no *Gene Ontology*, por exemplo, mostraram erro estimado de 49% (FREITAS; WIESER; APWEILER, 2010).

Protocolos baseados no agrupamento de sequências repetitivas são utilizados para predição *ab initio* de TEs. Pode-se citar como exemplo o RepeatExplorer (NOVAK et al., 2013), ferramenta que pode ser utilizada com leituras de sequenciamento e que é capaz de identificar regiões moderadamente e altamente repetitivas. Outro exemplo é o pacote REPET (FLUTRE et al., 2011), que tem como base dois *pipelines*: (1) TEdenovo e (2) TEannot. O primeiro é responsável pela predição e classificação, e o segundo pela anotação propriamente dita. Para predição, o conjunto de dados deve ser primeiramente fragmentado em grandes porções, utilizando janelas deslizantes para geração de extremidades sobrepostas. Cada uma dessas sequências deve então ser comparada com todas as outras a fim de encontrar repetições. As sequências mais similares são então agrupadas e em seguida cada grupo é submetido a um alinhamento global para inferência de consensos (Figura 9). Esses *pipelines* permitem a identificação de novos TEs, pois não dependem só de sequências conhecidas. Entretanto, o REPET e outros pipelines para predição *ab initio* demandam um esforço computacional intenso (FLUTRE et al., 2011), podendo demorar até uma semana para analisar um pequeno genoma em uma máquina de pequeno porte[1].

Figura 9 – *Pipeline* TEdenovo para predição *ab initio* de TEs.



Fonte: Adaptado de Flutre (2009).

[1] Informação fornecida por Romain Guyot (IRD, Montpellier, França) em sua estadia no Brasil como pesquisador visitante especial do programa Ciência sem Fronteiras.

2 miRNAs e siRNAs

A existência de RNAs não codificadores para proteínas (ncRNAs) é conhecida desde a década de 50, quando já se falava dos RNAs ribossomais e RNAs transportadores. A partir daí, seguiu-se a descoberta de muitos outros tipos, como os pequenos RNAs nucleares (snRNAs) e pequenos RNAs nucleolares (snoRNAs), envolvidos nos processos de *splicing* e maturação de grandes RNAs. Além disso, foram descobertos outros RNAs não codificantes, como por exemplo os miRNAs e os siRNAs, que serão o escopo desta seção (KREBS; GOLDSTEIN; KILPATRICK, 2014).

2.1 miRNAs

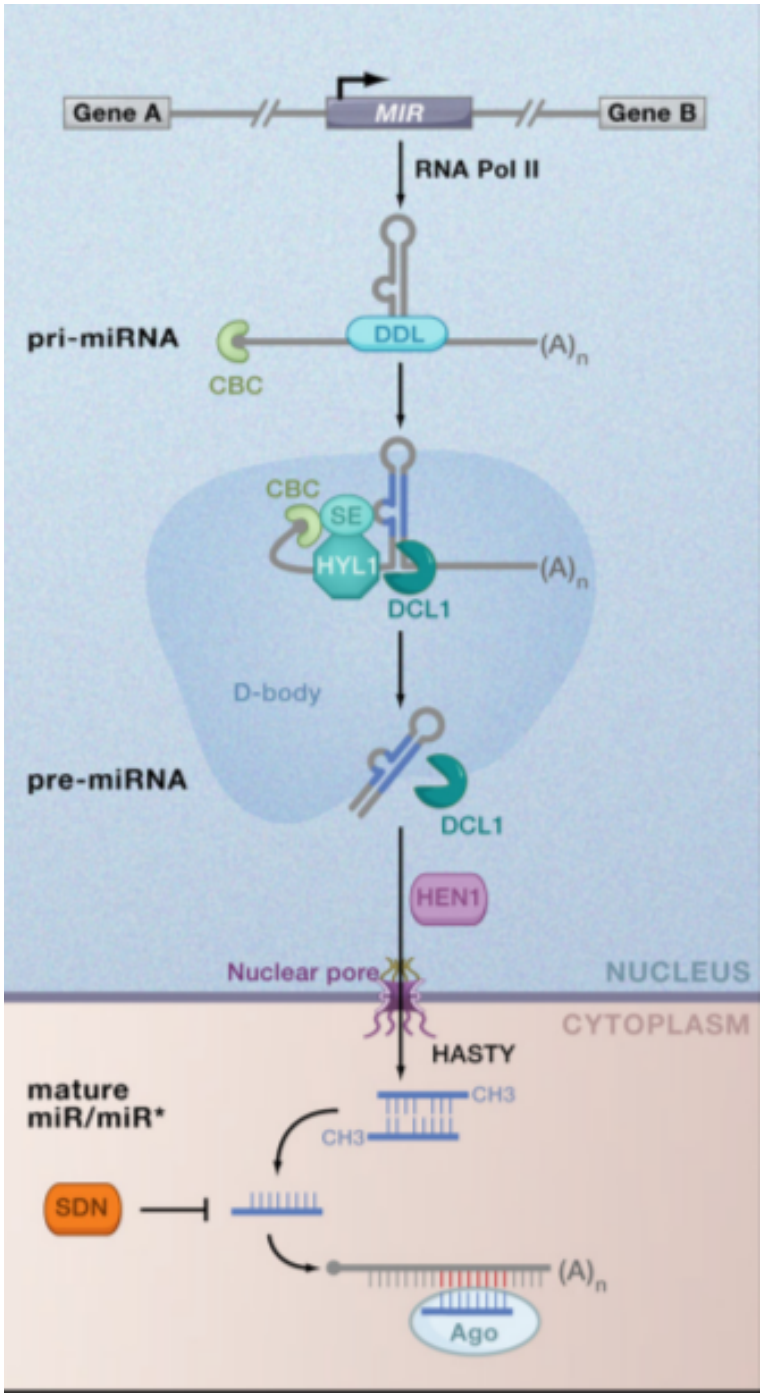
Os miRNAs são pequenas cadeias de RNA derivadas de estruturas secundárias denominadas grampos. Possuem usualmente 21 nucleotídeos (nt) de extensão e estão envolvidos no mecanismo de regulação da expressão gênica, ao lado de outros pequenos RNAs não codificantes como os pequenos RNAs de interferência (siRNAs) (AXTELL, 2013; ERSON-BENSAN, 2014).

Há evidência do surgimento das vias de biogênese de miRNAs nas algas antes mesmo de sua multicelularidade. O atual modelo de surgimento dessas vias sugere ainda que as estruturas precursoras tenham sido originadas a partir de duplicações invertidas dos seus genes alvos (DEBAT; DUCASSE, 2014).

O mecanismo de biogênese canônica em plantas (Figura 10) consiste inicialmente na transcrição de um *locus* pela enzima RNA Polimerase II (Pol II). O produto desta reação (microRNA primário, ou pri-miRNA) é então reconhecido pela proteína DAWDLE (DDL) de ligação ao RNA e ao complexo nuclear de ligação ao *cap* 5' (CBC, do inglês *Cap-Binding Complex*) para estabilização do grampo. Com o auxílio da proteína SERRATE (SE) e da proteína de ligação ao RNA de dupla fita HYPONASTIC LEAVES1 (HYL1), o complexo DICER-LIKE1 (DCL1) cliva o pri-miRNA para a formação do microRNA precursor (pre-miRNA), caracterizado por sua estrutura secundária em forma de grampo e uma extremidade saliente de dois nucleotídeos na porção 3'. Os pre-miRNAs são posteriormente processados para formação do dúplex miRNA/miRNA*. O dúplex passa por metilação promovida pela metiltransferase HEN1 (do inglês *Hua Enhancer*) para evitar degradação, e o produto maduro é exportado possivelmente por exportinas HASTY. Fora do núcleo, o miRNA é acoplado a proteínas ARGONAUTE (AGO) (VOINNET, 2009). Há a formação do complexo de silenciamento induzido por RNA (RISC, do inglês *RNA-induced silencing complex*), agente promotor do fenômeno do RNA interferente (RNAi).

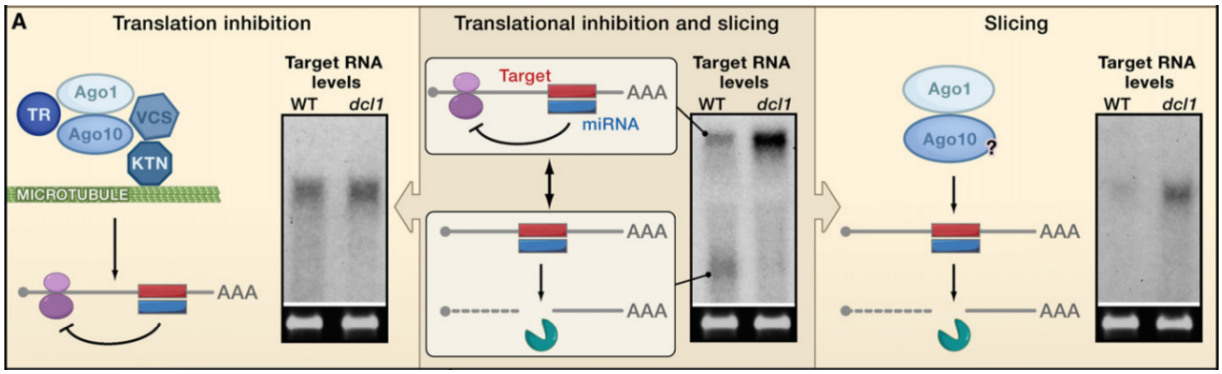
De acordo com Debat e Ducasse (2014), a partir da formação do RISC, o miRNA maduro se liga a RNAs mensageiros (mRNAs) na região não traduzida 3' (UTR 3') e os silencia por desestabilização, repressão traducional, ou por clivagem direta, com alto nível de complementaridade entre os RNAs pareados (Figura 11).

Figura 10 – Biogênese canônica de miRNAs.



Fonte: Voinnet (2009).

Figura 11 – Mecanismos de ação de miRNAs. A caixa da esquerda mostra o mecanismo de repressão traducional. Após o acoplamento do complexo RISC ao transcrito, o ribossomo é impedido de promover a tradução. A caixa da direita mostra como ocorre a clivagem direta do transcrito reconhecido pelo RISC. Na caixa central, é mostrado um caso onde ocorrem ambos os mecanismos descritos. Cada um dos quadros mostra um experimento de *northern-blot* em *Arabdopsis thaliana* para confirmar o tamanho dos mRNAs encontrados em cada situação. WT: tipo selvagem; dcl1: mutante sem DCL1.



Fonte: Voinnet (2009).

O papel dos miRNAs em plantas já foi demonstrado de diversas formas. Por exemplo, mutantes com alterações severas nas proteínas DCL1 são abortados ainda enquanto embriões, ou podem apresentar anomalias na organogênese floral, morfologia foliar, e iniciação dos meristemas axilares. A superexpressão do miR159, conservado entre espécies vegetais e que tem como alvo os fatores de transcrição da família MYB, pode promover esterilidade das flores masculinas e também atrasar a época de floração (JONES-RHOADES; BARTEL; BARTEL, 2006).

Aplicar esse conhecimento para o melhoramento vegetal não é algo fora da realidade. Desde o início da década de 90, abordagens baseadas em mecanismos de RNA de interferência têm sido empregadas para modificar organismos e obter melhores condições para sua produção e comercialização. Pode-se citar como exemplo a inibição da produção de poligalacturonase durante o amadurecimento do tomate, resultando no retardo da degradação de suas paredes celulares. Outro exemplo, é a resistência a vírus conferida por essa técnica biotecnológica, já aplicada a diversas culturas como mamão, beterraba, ameixa, feijão e tabaco. Com a elucidação de fenômenos relacionados ao funcionamento e regulação das redes de ncRNAs, novas estratégias de melhoramento poderão ser traçadas (LIU; ZHU, 2014).

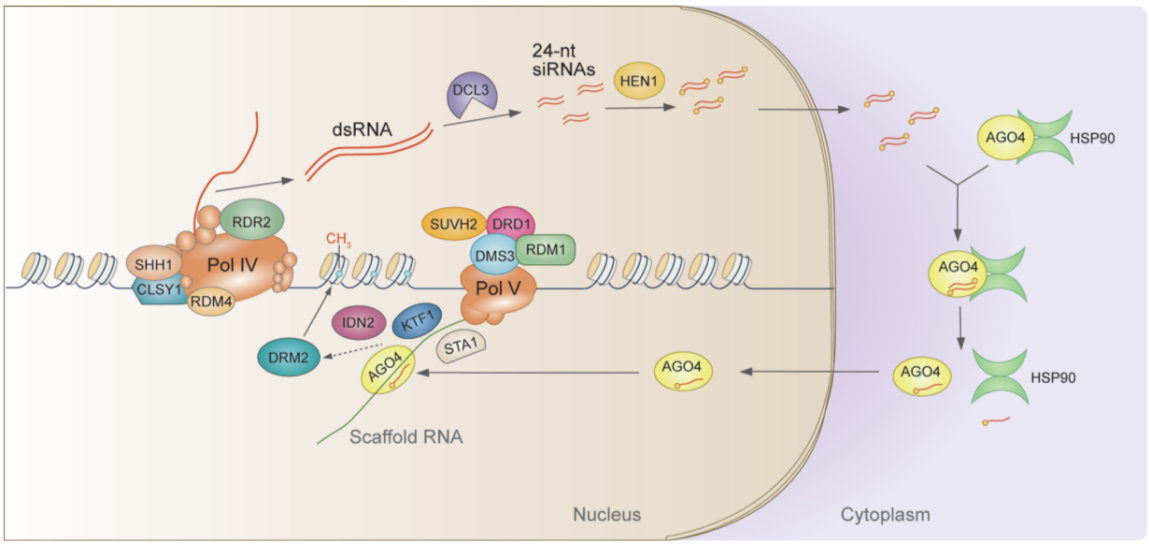
2.2 siRNAs

Os sequenciamentos de nova geração permitiram mostrar que o repertório de pequenos RNAs em plantas é vastamente dominado por um "oceano" de siRNAs, agindo principalmente ao nível de cromatina e tendo como alvo *loci* de transposons e outras repetições (VOINNET, 2009).

SiRNAs são pequenos RNAs em sua maioria derivados de RNAs precursores de dupla fita (dsRNAs, do inglês *double-stranded* RNA). Possuem 23 ou 24 nt de extensão e participam principalmente do mecanismo de metilação de DNA direcionada por RNA (RdDM, do inglês RNA-*directed* DNA *methylation*), promovendo modificações *de novo* 5-metilcitosina no DNA, e H3K9 em histonas (AXTELL, 2013).

Sua via canônica de biogênese depende da RNA polimerase dependente de DNA IV (Pol IV), RNA polimerase 2 dependente de RNA (RDR2, do inglês RNA-*dependent* RNA *polymerase 2*), DICER-LIKE 3 (DCL3) e ARGONAUTE 4 (AGO4). O mecanismo efetor da metilação tem como protagonistas as proteínas DNA metil-transferase (DRM2) e RNA Polimerase dependente de DNA V (Pol V) (XIE; YU, 2015). A biogênese é iniciada quando a Pol IV transcreve um trecho que será complementado pela RDR2 e depois clivado pela DCL3. O dúplex siRNA gerado é exportado para o citoplasma onde será ligado à proteína AGO4 e em seguida direcionado de volta ao núcleo. A partir da transcrição do *locus* alvo pela Pol V, a AGO4 é atraída para o complexo ao mesmo tempo em que direciona a DRM2 para o sítio de metilação, que pode ser RNA, DNA ou histonas. Muitas outras proteínas acessórias atuam em conjunto para que o processo seja efetuado (Figura 12).

Figura 12 – Biogênese de siRNAs e mecanismo de metilação do DNA direcionada por RNA.



Fonte: Zhao e Chen (2014).

O mecanismo RdDM impede a transposição de TEs autônomos que na maioria das vezes promovem alterações deletérias ou neutras nos genomas (LISCH, 2013), e uma vez que age em determinado sítio, reforça o processo de manutenção da metilação executado por outros mecanismos (BOND; BAULCOMBE, 2014). Mutantes de *A. thaliana* com DRM2 defectivas não demonstraram grandes alterações nos padrões de transcrição, enquanto mutantes para MET1 (proteína envolvida na manutenção da metilação) tiveram aumento massivo nos níveis de transcrição de transposons e pseudogenes (ZHANG et al., 2006). Isso mostra a efetividade do mecanismo RdDM, pois após a metilação de um TE, dificilmente ele voltará a ser ativo.

Devido a natureza da metilação, que acontece diretamente nos desoxirribonucleotídeos, essas alterações podem ser passadas para as gerações seguintes (ZHAO; CHEN, 2014), permitindo assim a acumulação de TEs nos genomas, mesmo que inativos (FEDOROFF, 2012).

3 Relação TE-miRNA

Diversos estudos relacionam a origem de miRNAs às inserções de TEs. Piriyapongsa e Jordan (2008) sugerem um modelo em que MITEs transcritos são processados pela maquinaria de biogênese de pequenos RNAs, devido à sua natureza estrutural que favorece a formação de grampos. Outros trabalhos levantaram evidências para esse fenômeno, mostrando que a produção de miRNAs derivados de MITEs são dependentes de proteínas DCL1 e AGO1 em *O. sativa* (OU-YANG et al., 2013), e que pequenos RNAs de solanáceas são dependentes de proteínas DCL3 e DCL4 (KUANG et al., 2009).

Roberts, Cardin e Borchert (2014) demonstram como a amplificação das cópias de TEs em um genoma pode favorecer o surgimento de grampos e gerar novos alvos com a inserção aleatória desses TEs em genes codificadores. Em seu exemplo, a inserção de um mesmo LINE em direções opostas proporciona o surgimento de um *loci* formador de grampo. Este grampo após ser processado pela via de biogênese de miRNAs terá um produto que poderá agir na regulação pós-transcricional de mRNAs complementares. É possível ainda que este mesmo LINE se insira aleatoriamente em algum gene sem afetá-lo estruturalmente, mas marcando-o como alvo para atuação do mecanismo de silenciamento (Figura 13). Debat e Ducasse (2014) afirmam ainda que TEs também podem ter sido responsáveis pela amplificação do número de cópias de genes de miRNAs ao longo do processo evolutivo. Deve-se isto ao acúmulo de mutações em LTR-RTs, favorecendo a formação aleatória de estruturas em grampo, e à consequente seleção positiva dos miRNAs a partir deles gerados, já que silenciam sua atividade replicativa, que muitas vezes é deletéria.

Segundo Bennetzen e Wang (2014), levando-se em conta que os miRNAs são abundantes e muitas vezes espécie-específicos, pode-se sugerir um modelo simples, no qual a maior parte deles - senão todos - possam ter sido criados a partir de TEs. Porém, somente aqueles gerados mais recentemente mostram vestígios que possam relacioná-los.

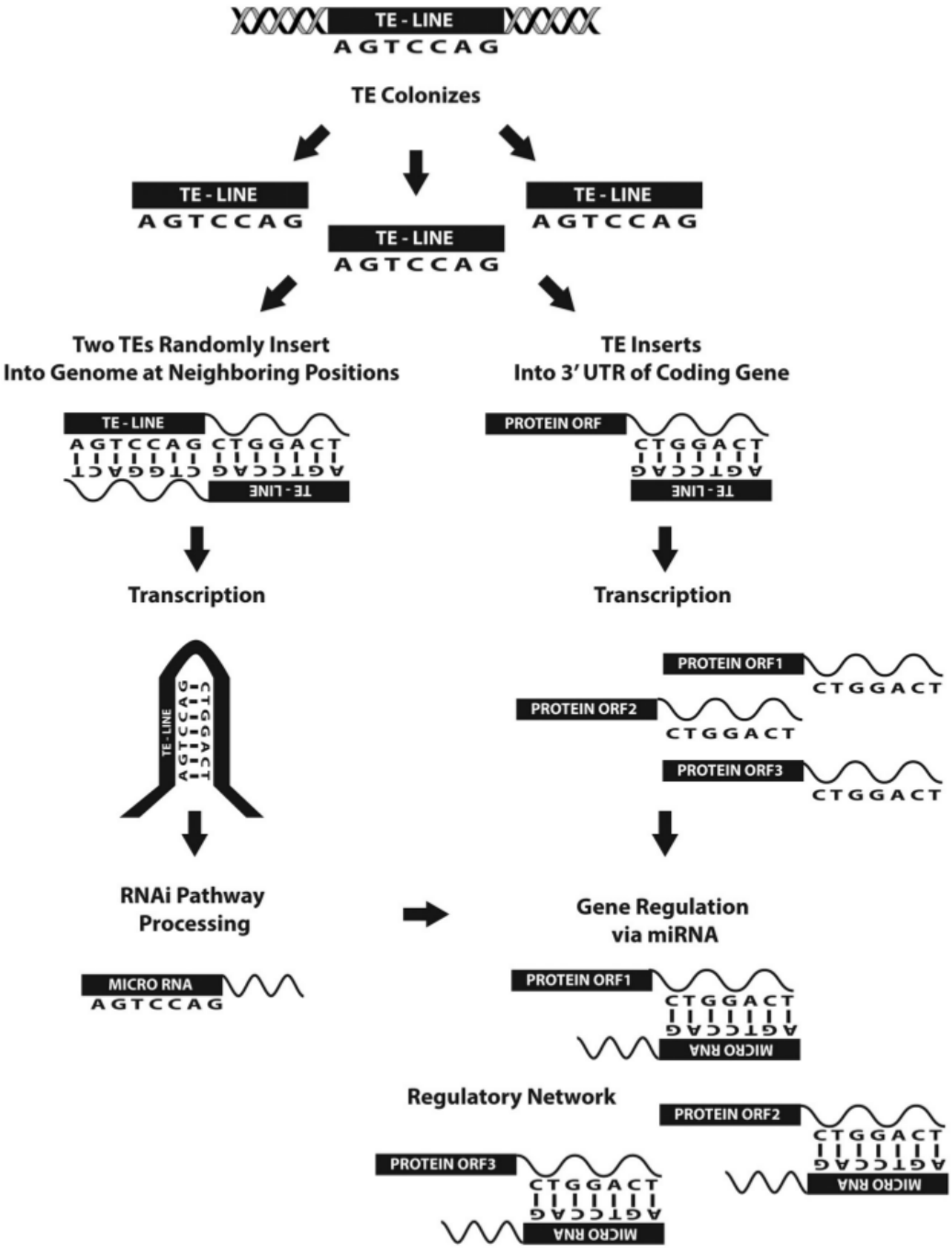
A fim de identificar pequenos RNAs que possam ter sido originados a partir de TEs em genomas vegetais, diversos autores buscaram similaridade entre estes dois tipos de sequências (PIRIYAPONGSA; JORDAN, 2008; LI et al., 2011; ZHANG; JIANG; GAO, 2011; SUN et al., 2012; ROBERTS et al., 2013). Entre eles, Li et al. (2011) analisaram esta relação em sete espécies vegetais: *A. thaliana*, *Brassica napus*, *Glycine max*, *Medicago truncatula*, *O. sativa*, *Solanum lycopersicum* e *Triticum aestivum*. Esses autores encontraram resultados em três delas, mas só os achados em *O. sativa* foram suficientemente abundantes para serem investigados a partir de dados provenientes do sequenciamento de pequenos RNAs. Neste trabalho, os miRNAs derivados de TEs foram denominados TE-MIRs, e

somente aqueles que passaram pelos critérios de anotação de Meyers et al. (2008) foram endossados.

Em metazoários essa relação já foi melhor explorada. Tem-se como exemplo o trabalho de Levy, Sela e Ast (2008), no qual analisou-se em escala genômica as interseções posicionais entre anotações de TEs e pre-miRNAs em sete espécies de vertebrados e invertebrados. Essa análise permitiu que fosse criado o primeiro banco de dados para armazenar anotações de TE-MIRs provenientes desses animais: o microTranspoGene[1]. Outro exemplo é o trabalho de Tempel e Tahi (2012), que lançou a ferramenta ncRNAclassifier e identificou TE-MIRs em seis espécies de cordados.

[1] <http://transpogene.tau.ac.il/microTranspoGene.html>.

Figura 13 – Origem de grampos pela inserção de TEs e a geração de novos alvos.



Fonte: Roberts, Cardin e Borchert (2014).

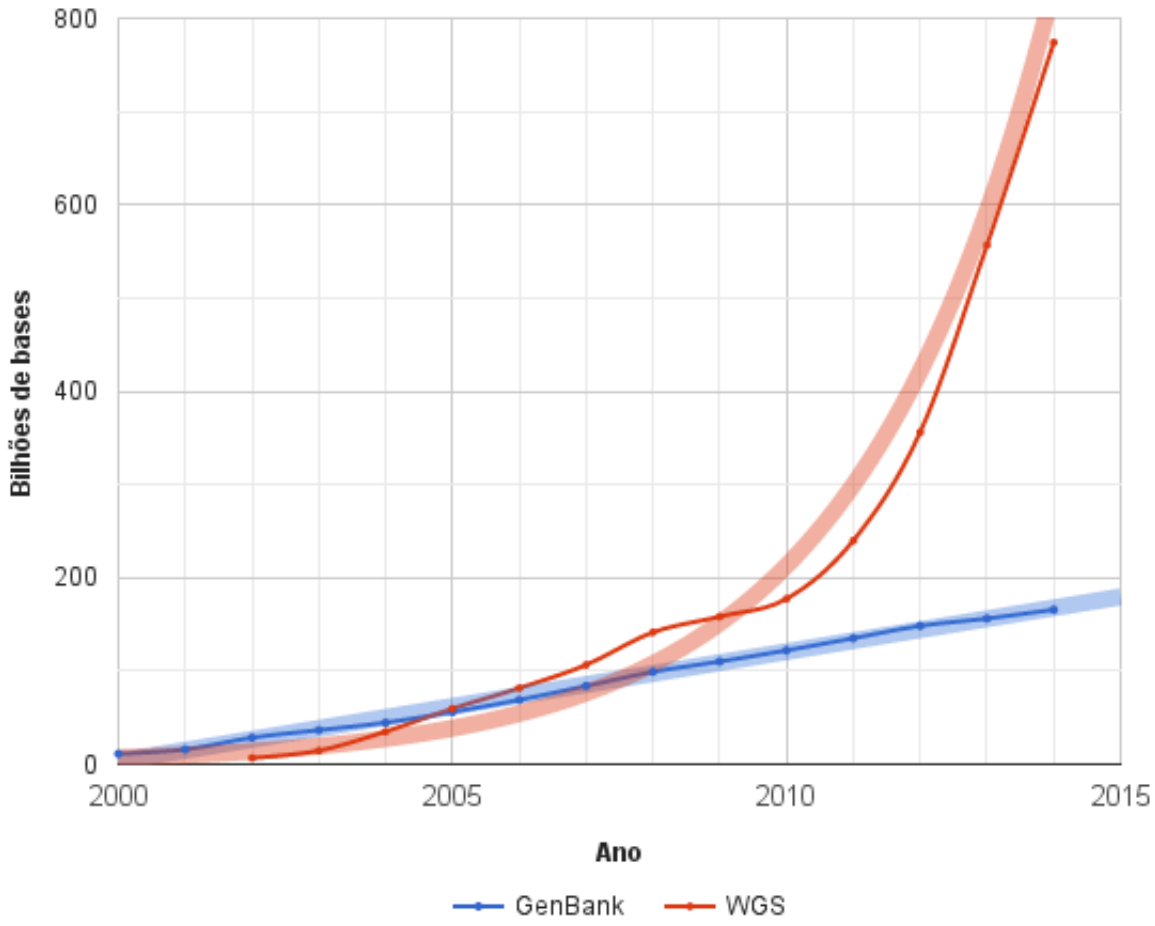
4 Bancos de dados para informação biológica

Com a queda dos preços dos equipamentos e protocolos utilizados no sequenciamento, e a redução do tempo de obtenção das sequências, mesmo pequenos laboratórios estão gerando grandes quantidades de dados biológicos. Os pesquisadores da área biológica estão entrando agora para o "clube do *big-data*" (MARX, 2013). Pode-se ilustrar este fato com a Figura 14, que mostra o crescimento do volume de dados armazenados no *GenBank*[1] do *National Center for Biotechnology Information* (NCBI), repositório público agregador de toda a informação da *International Nucleotide Sequence Database Collaboration* (INSDC). Somente em 2013 foram sequenciados de 50 a 80 genomas de plantas, trazendo à tona assim grandes desafios no que se diz respeito ao armazenamento e compartilhamento de dados biológicos (BRAUER; SINGH; POPESCU, 2014).

Muitas outras bases de dados biológicos têm se estabelecido nos últimos anos para organizar e armazenar informações cada vez mais específicas. O Phytozome[2] (GOODSTEIN et al., 2012) e o PLAZA[3] (PROOST et al., 2015) são bons exemplos de repositórios que contém sequências genômicas completas e suas anotações. Outro, já citado aqui neste trabalho, é o Repbase Update[4] (JURKA et al., 2005), uma das maiores coleções curadas de elementos repetitivos. É possível encontrar ainda, repositórios para tipos específicos de TEs, como é o caso do P-MITE[5] (CHEN et al., 2014), que guarda apenas sequências de MITEs anotados em genomas vegetais. Paschoal et al. (2012) relatam a existência de mais de 100 bancos de dados que armazenam algum tipo de informação relacionada a ncRNAs, entre os quais está incluído o miRBase[6] (KOZOMARA; GRIFFITHS-JONES, 2014). Vale ainda enfatizar a existência de um banco de dados denominado microTranspoGene (LEVY; SELA; AST, 2008) já citado na seção anterior, que agrega anotações relativas aos TE-MIRs de algumas espécies de animais.

[1] <http://www.ncbi.nlm.nih.gov/genbank/>
[2] <http://phytozome.jgi.doe.gov/pz/portal.html>
[3] <http://bioinformatics.psb.ugent.be/plaza/>
[4] <http://www.girinst.org/repbase/>
[5] <http://pmite.hzau.edu.cn/>
[6] <http://mirbase.org/>

Figura 14 – Crescimento do *GenBank* nos últimos 15 anos. Pode-se observar que os dados genômicos (WGS, *Shotgun* de genoma inteiro) apresentam tendência ao crescimento exponencial, enquanto os dados do *Genbank*, propriamente dito, crescem com tendência linear.



Fonte: Adaptado de Benson et al. (2015).

5 Objetivos

Considerando a importância dos TEs para a elucidação do funcionamento dos genomas e sua relação com pequenos RNAs, bem como a necessidade de organizar os dados em plataformas de fácil acesso para a comunidade científica, o objetivo geral da primeira parte foi investigar a possível relação entre TEs e miRNAs nos mais diversos genomas vegetais e tornar os resultados acessíveis. Os objetivos específicos desse trabalho foram: anotar TEs e miRNAs para as espécies selecionadas, realizar a busca por associações entre as duas categorias, e disponibilizar os resultados em um banco de dados público.

Em vista da grande quantidade de MITEs relacionados a pequenos RNAs encontrados na primeira parte do trabalho, o objetivo geral da segunda parte foi identificar famílias desses elementos no genoma de *Coffea canephora* e mostrar quais delas podem ser precursoras de pequenos RNAs. Para isso, os objetivos específicos foram: realizar a anotação *de novo* de MITEs para esse genoma, identificar as famílias que possivelmente estão associadas a pequenos RNAs, e classificar cada uma delas de acordo com critérios pré-estabelecidos.

Parte II

Artigo publicado no periódico
Functional & Integrative Genomics:
PlanTE-MIR DB: a Database for Transposable Element-related microRNAs in Plant Genomes

Abstract

Transposable elements (TEs) comprise a major fraction of many plant genomes and are known to drive their organization and evolution. Several studies show that these repetitive elements have a prominent role in shaping noncoding regions of the genome such as microRNA (miRNA) loci, which are components of post-transcriptional regulation mechanisms. Although some studies have reported initial formation of miRNA loci from TE sequences, especially in model plants, the approaches that were used did not employ systems that would allow results to be delivered by a user-friendly database. In this study, we identified 152 precursor miRNAs overlapping TEs in 10 plant species. PlanTE-MIR DB was designed to assemble this data and deliver it to the scientific community interested in miRNA origin, evolution, and regulation pathways. Users can browse the database through a web interface and search for entries using various parameters. This resource is cross-referenced with repetitive element (Repbase Update) and miRNA (miRBase) repositories, where sequences can be checked for further analysis. All data in PlanTE-MIR DB are publicly available for download in several file formats to facilitate their understanding and use. The database is hosted at <http://bioinfo-tool.cp.utfpr.edu.br/plantemirdb/>.
Keywords: Transposable elements. microRNAs. plant genomes. database.

6 Introduction

Transposable elements (TEs) are present in almost all living organisms and comprise a significant fraction of most plant genomes (BAIDOURI; PANAUD, 2013; RAGUPATHY; YOU; CLOUTIER, 2013). They are known to drive important modifications in host genomes, including the inactivation, creation, and mobilization of genes, chromosomal rearrangement, gene expression modulation, and epigenetic silencing (LISCH, 2013; BENNETZEN; WANG, 2014). TEs are also known to have an important role in shaping long noncoding RNAs (lncRNAs) and small ncRNAs, (e.g., piwi-interacting RNAs [piRNAs], small interfering RNAs [siRNAs], and microRNAs [miRNAs]) (HADJIARGYROU; DELIHAS, 2013; PIRIYAPONGSA; JORDAN, 2008; LI et al., 2011; GIM et al., 2014).

Mature plant miRNAs are usually 21-nucleotide-long hairpin RNA-derived (hpRNA) sequences that can bind to target mRNAs through Watson-Crick base pairing on the 3' UTR. This pairing results in mRNA destabilization or translational repression, which are effective mechanisms for gene regulation. In plants, RNA polymerase II enzymatic complexes are generally committed to transcribe miRNA loci. Inside the cell nucleus, cleavage of foldback transcribed structures through DICER-LIKE 1 (DCL1) is executed. This step promotes conversion of primary miRNAs (pri-miRNAs) into precursor miRNAs (pre-miRNAs), which is followed by further processing to transform them into miRNA/miRNA* duplex. Finally, one of the dissociated miRNAs is loaded into ARGONAUTE (AGO) proteins to assemble the functional RNA-induced silencing complex (RISC) (AXTELL, 2013; ERSON-BENSAN, 2014).

Hairpin structures are supposed to arise either by inverted duplication of the target gene locus (ALLEN et al., 2004) or juxtaposed TEs (ROBERTS et al., 2013). Piriyapongsa and Jordan (2008) also describe a model in which folded expressed Miniature Inverted-repeat Transposable Elements (MITEs) may be processed by the miRNA biogenesis pathway. In another report, Ou-Yang et al. (2013) used AGO1 immunoprecipitation and DCL mutants to find three MITE-associated bona fide miRNAs depending on those proteins, pointing to their functional activity.

Most miRNA loci originate from intergenic genomic sequences, but there is considerable evidence that they were initially formed from TE sequences (HADJIARGYROU; DELIHAS, 2013; ROBERTS; CARDIN; BORCHERT, 2014; BUDAK; AKPINAR, 2015). The "domestication" of TEs to form miRNA genes has been demonstrated by high throughput sequencing in rice (LI et al., 2011; BARRERA-FIGUEROA et al., 2012), and other plant species were checked for TE-MIRs at the genomic level (ZHANG; JIANG; GAO, 2011; SUN et al., 2012; KURTOGLU; KANTAR; BUDAK, 2014). Similar findings have

been reported for the human genome and other metazoan genomes (LEVY; SELA; AST, 2008; TEMPEL; POLLET; TAHI, 2012), for which there are publicly available resources showing matched overlaps between TE and miRNA loci. However, the comparable plant data has not yet been compiled into a user-friendly database enabling search and retrieval of this information.

In this study, we present PlanTE-MIR DB[1], which provides a user-friendly database for investigation of overlaps between TE and pre-miRNA loci in 10 plant genomes.

[1] Available at <http://bioinfo-tool.cp.utfpr.edu.br/plantemirdb/>.

7 Material and methods

7.1 Pre-miRNA annotation and curation

Our analysis relied on the annotated pre-miRNAs from miRBase (version 21) (KOZOMARA; GRIFFITHS-JONES, 2014) within 15 genome assemblies. Genome assemblies were retrieved from several repositories based on reference versions indicated by miRBase (Apêndice A, Table S1). However, due to divergences between miRBase annotation file accession names and assembly headers, a checking step was executed. For that, miRBase pre-miRNA sequences in FASTA format were obtained for each of the studied plant species and BLAST (BLASTN, version 2.2.28+) (CAMACHO et al., 2009) searched against their respective genomes. An in-house bash script was made to run the program and perform the tasks. Only hits with 100% query coverage and identity were maintained. In the case that some of the annotated pre-miRNAs from the same family showed indistinguishable sequences (e.g., mtr-MIR2669a, mtr-MIR2669b), they were aligned with more than one position. In these cases, we split single position hits from repeated ones and transformed them into GFF3 annotation files. Manual inspection was done using Artemis (version 15.1.1) (RUTHERFORD et al., 2000) through the comparison of previously cited GFF3 files with an accession name corrected miRBase annotation file, all loaded on source genomes. New manually inspected GFF3 annotation files were created to match the accession names to source assembly headers.

7.2 Reference TEs

Plant TE libraries were obtained from Repbase Update (version 19.04 and 19.06 REPET edition) (JURKA et al., 2005). We used CENSOR (version 4.2.28 and 4.2.29) (JURKA et al., 2005) implemented with WU-BLAST (version 2.0 04-May-2006) as a search engine, using BLASTN and BLASTX algorithms according to well established criteria (WICKER et al., 2007) to stringently remap reference TEs to genome assemblies. Initially, we used early versions of software and libraries with BLASTN, and then later versions with BLASTX. A bash script was written to filter the results, according to the 80-80-80 rule proposed by Wicker and colleagues, and to parse TE coordinates to GFF3 annotation files.

7.3 TE-MIR relationship

We found positional overlaps between TEs and pre-miRNAs using the BEDTools (version 2.17.0) (QUINLAN; HALL, 2010) intersection function. Only pre-miRNAs having at least 36% of their extension covered by a TE were maintained. Intersections were manually checked using Artemis. Whole sequences were captured from source assemblies through an in-house bash script running EMBOSS tools (version 6.6.0.0) (RICE; LONGDEN; BLEASBY, 2000).

7.4 Evolutionary conservation between TE-MIRs across taxa

We used a sequence-based similarity search method (Reciprocal Best BLAST Hit - RBH), following the rationale of Sun et al. (2012) to track evolutionary conservation. Thus, we BLAST (BLASTN, version 2.2.28+) searched our pre-miRNAs against all miRBase (version 21) hairpin sequences. Only hits with E-values $\leq$ 1e-06 with at least 90% query alignment and minimum of 80% identity were maintained. We employed the same criteria used by Zhou et al. (2011) and Sun et al. (2012) to classify TE-MIRs at three levels: highly conserved (when TE-MIR homologs are present in both monocots and eudicots), low conserved (when TE-MIR homologs are present only in monocots or eudicots) and nonconserved (when a TE-MIR has no homologs outside a single species).

7.5 Transcriptional evidence for miRNAs

We checked for transcriptional evidence through browsing only high confidence miRNAs for each one of the plant species in the miRBase (version 21). Our pre-miRNAs were also used as queries and BLAST (BLASTN, version 2.2.28+) searched against the miRNEST (version 2.0) deep sequencing prediction file (SZCZEŚNIAK; MAKAŁOWSKA, 2014). Only hits presenting full identity and coverage of queries were maintained.

7.6 Database and web interface implementation

Annotation files were parsed to table using a Perl in-house script. Additional information was manually introduced using the Kingsoft Office Spreadsheet software. The data were then exported to a comma-separated values table and automatically inserted in MySQL Database Server (version 5.6) relational tables using a PHP script (version 5.3.10).

PlanTE-MIR DB was built on a 64-bit Windows (version 8.1) workstation. XAMPP (version 3.2.1) was executed to integrate Apache HTTP Server (version 2.2) with PHP and MySQL. The back-end was encoded in PHP and HTML5, using JavaScript jQuery library

(version 1.11.2), with the plugins jQuery Vegas (version 1.3.5) and Ajax. For website design and structure customizing, we used Cascading Style Sheets (CSS3) as front-end. Except for the Windows operating system, only open source and cross-platform software were used for database and web interface implementation. The complete system is hosted in the Information Office of the Federal University of Technology - Paraná, Brazil (UTFPR) and available at <http://bioinfo-tool.cp.utfpr.edu.br/plantemirdb/>.

8 Results and discussion

8.1 PlanTE-MIR DB: system and database overview

PlanTE-MIR DB (Plant Transposable Element-related miRNA Database, Figura 15) was built as a resource for researchers interested in the evolution of TE and miRNA and their relationship. In this section, we detail the website and its functionalities.

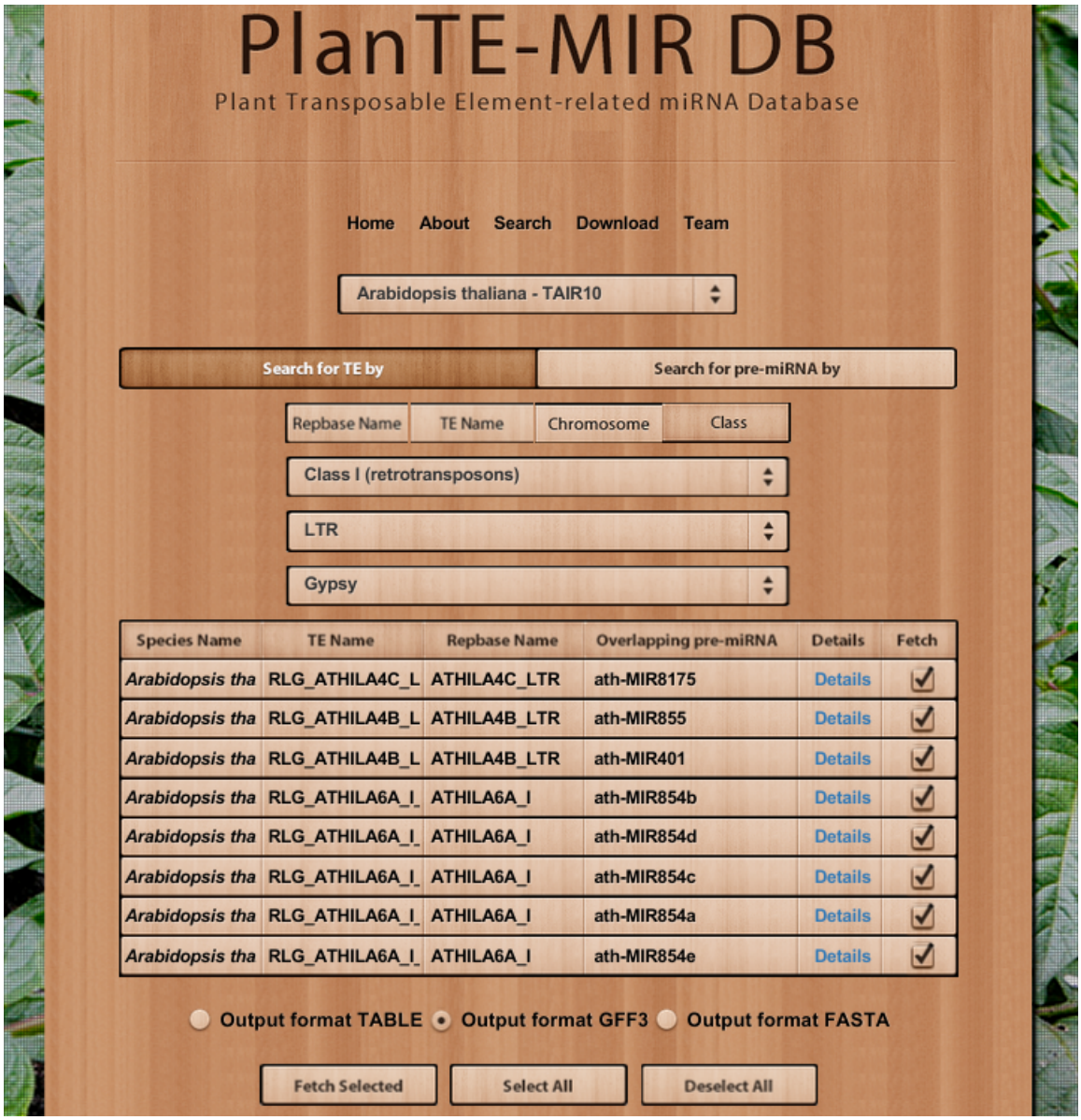
The web interface was designed to be user-friendly, prioritizing easy ways of finding desired data through the use of filters, and providing alternative file formats when downloading entries. Accordingly, the website is divided in five sections: Home, About, Search, Download, and Team.

The Home and About sections concisely describe the purpose of the repository and its methods, and briefly instruct the user on how to interact with the search and download tools. They also contain information about assembly versions and reference libraries employed in the analysis.

In the Search section, users have a web interface for searching TE-MIR entries. The page was designed as an intuitive step-by-step form where the user can select options by name of the organism and TE or pre-miRNA attributes (Figura 15). TEs can be found (1) by selecting the reference name (as supplied by the Repbase Update), (2) by TE name according to the nomenclature of Wicker and colleagues, (3) by TE position in the genome assembly, and (4) by TE class (WICKER et al., 2007). The last option is a hierarchical filter that allows the user to choose among TE classes, orders and superfamilies. Similarly, pre-miRNAs may be found by miRBase ID, miRBase name, or position in the assembly. Next, a list of hits is shown to the user, allowing him or her to download search results by selecting Table file format, GFF3 file format, or FASTA file format. Furthermore, the user may access a detailed page containing information about the organism as well as annotations and cross-references obtained for each result. The description table shows species name, common name, assembly version, TE name, TE classification, Repbase Name, TE annotation details (such as Repbase version, CENSOR coverage, CENSOR similarity, start position, end position, and strand), overlapping pre-miRNA, pre-miRNA ID, and pre-miRNA annotation details. Further information relative to these items can be found in electronic supplementary material.

Bulk data for each species are provided through the Download section. All data are available in three formats: table, GFF3 and FASTA. GFF3 annotation files for TEs and pre-miRNAs can be directly loaded into publicly available assemblies using a genome browser tool (e.g., Artemis).

Figura 15 – Search section overview. Entries can be searched either by TE or pre-miRNA. Here we present a search example for TE by Class I (retrotransposons), LTR order, and Gypsy superfamily.



Source: Produced by the authors.

8.2 Identification of TE-MIRs

Since we found some inconsistencies between genome versions in miRBase and those in TE curated databases, we started our analyses by remapping reference TEs to standardized versions of plant genomes (Apêndice A, Table S1). Figura 16 summarizes our approach to finding TE-MIR associations; in brief, we searched for positional intersections between pre-miRNAs and TEs. Among 10 species, we obtained a total of 152 pre-miRNAs that overlapped at least one TE (Tabela 2, Figura 18). Most pre-miRNAs were associated

to DNA transposons and MITEs (Figura 17). Our analysis provided 60 new cases of high confidence repetition-related miRNAs (Apêndice A, Fig. S1).

Tabela 2 – Overall numbers of TEs, miRNAs, and TE-MIRs for the plant genomes analyzed in this study. Ten species presented at least one TE-MIR in their genome.

Species	TEs[a]	TEs[b]	pre-miRNAs[c]	TE-MIRs[d]
Arabidopsis thaliana	6837	314	325	22
Brachypodium distachyon	3991	1402	317	2
Glycine max	18035	18290	573	4
Medicago truncatula	7481	1128	672	20
Oryza sativa	66794	2930	592	56
Physcomitrella patens	23744	3855	229	1
Populus trichocarpa	14587	2143	352	10
Sorghum bicolor	178106	29885	205	35
Solanum tuberosum	5530	14814	224	1
Vitis vinifera	16240	4787	163	1

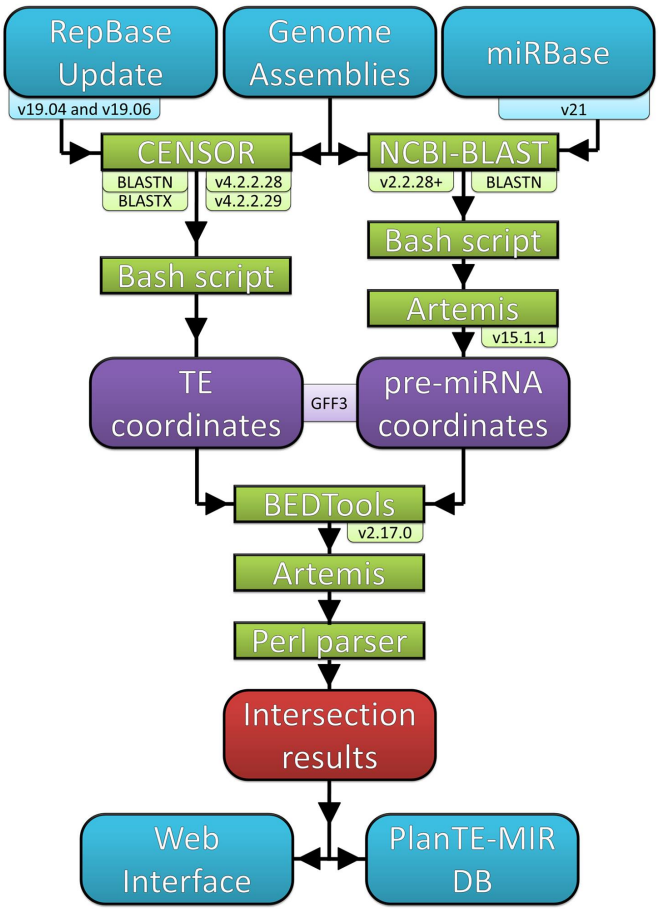
[a]Annotated TEs using CENSOR with BLASTN. [b]Annotated TEs using CENSOR with BLASTX. [c]Pre-miRNAs retrieved from miRBase (version 21). [d]MiRNAs intersecting at least one TE in each species.

Source: Produced by the authors.

Nine pre-miRNAs (ath-MIR401, ath-MIR854a, ath-MIR854b, ath-MIR854c, ath-MIR854d, ath-MIR855, osa-MIR812b, osa-MIR814a, and osa-MIR814b) indicated by Piriyapongsa e Jordan (2008) as the products of siRNA-miRNA dual coding TEs were confirmed by our analyses. Four members in a rice miRNA family (osa-MIR812f, osa-MIR812h, osa-MIR812i, and osa-MIR812j) and osa-MIR1850 were formerly classified as typical TE-MIRs (LI et al., 2011), since they are in conformity to a standardized protocol of miRNA annotation rules (MEYERS et al., 2008). Ten miRNAs are indicated by miRBase to have transcriptional evidence, and 10 were found in miRNEST deep-seq predictions. Only two of them are present in both repositories.

To our knowledge, few small RNA precursors have been reported to be formed by juxtaposed TE insertions in plant species (KUANG et al., 2009; LI et al., 2011; ZHANG; JIANG; GAO, 2011). One of these cases is osa-MIR1879, which was classified as a bona-fide miRNA spanning two short non-autonomous retrotransposons. Other two pre-miRNAs (osa-MIR815b and osa-MIR815c) have similar structures, but were suggested as potential pre-evolved miRNAs (LI et al., 2011). Our analysis detected these pre-miRNAs, but they intersected only one TE. However, we found three cases of stem-loop structures formed

Figura 16 – Workflow diagram for the identification of TE-MIRs. CENSOR and BLAST programs were used to map TEs and pre-miRNAs. Bash script was used to filter and parse results to GFF3 file format. Using Artemis, pre-miRNA files were checked to confirm names and positions according to miRBase. Positional intersection analysis between TE-miRNA was run using BEDtools and manually checked with Artemis. These results were modelled to build PlanTE-MIR DB.

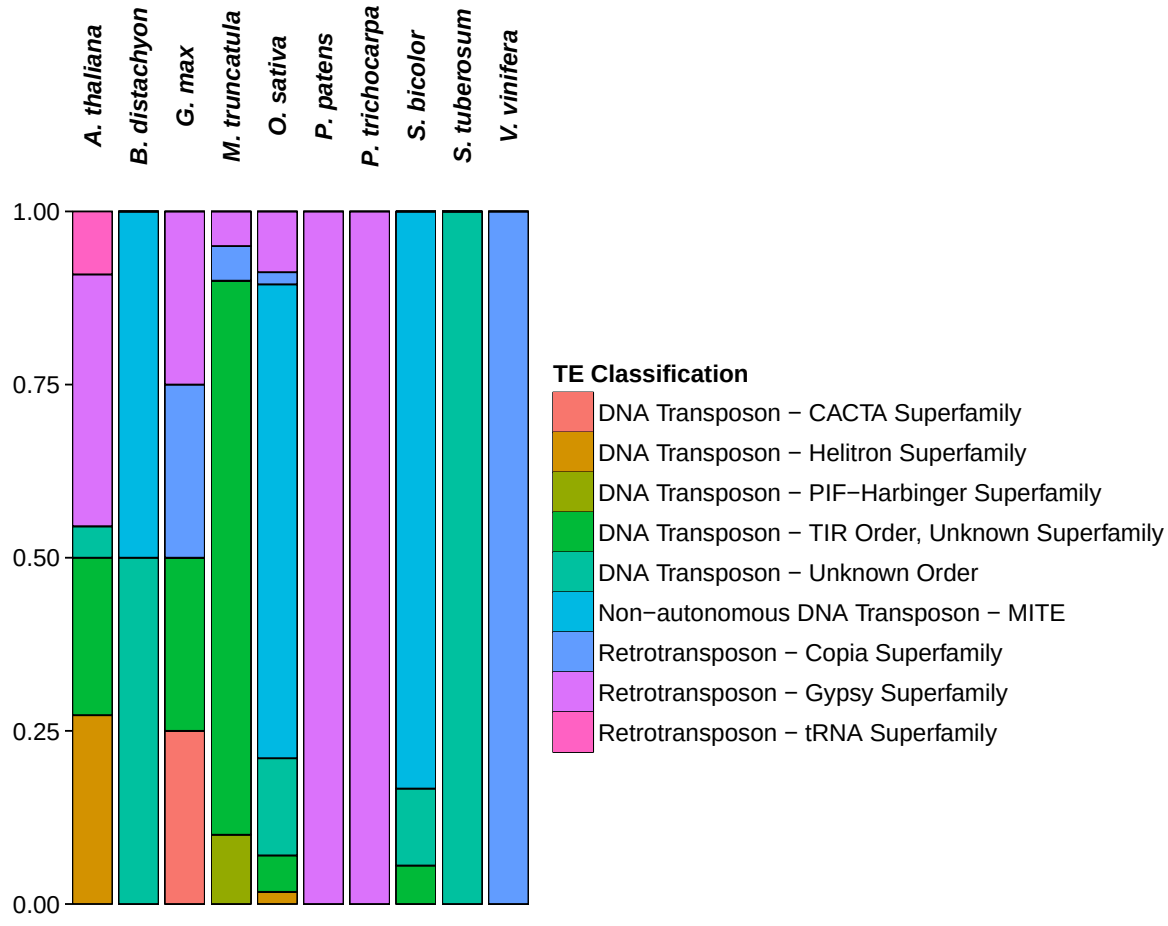


Source: Produced by the authors.

by TE juxtaposition in potato, sorghum, and rice. In *Solanum tuberosum*, two SONATA2 non-autonomous DNA transposons on the same strand compose the stu-MIR8019 foldback structure (Figura 19). In *Sorghum bicolor*, two DNA-3-2N_Sbi non-autonomous DNA transposons give rise to sbi-MIR6227 (Figura 20a). Within *Oryza sativa* the insertion of two OLO24 non-autonomous DNA transposons on opposite strands probably gave rise to osa-MIR1441 (Figura 20b).

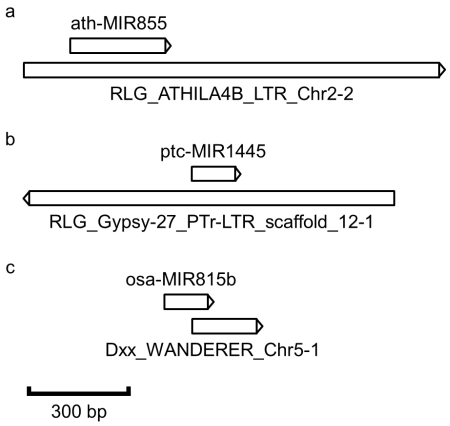
Using an adapted Reciprocal Best BLAST Hit method, we found that 92.11% of the matches were species-specific. This result emphasizes that the repetitive element-related miRNAs tend to be species-specific (SUN et al., 2012).

Figura 17 – Database composition by plant species and TE classification. MITEs are the most frequent type of repetition associated to miRNAs. DNA transposons were also highly related to this phenomenon. Classification data were collected from the Repbase Update.



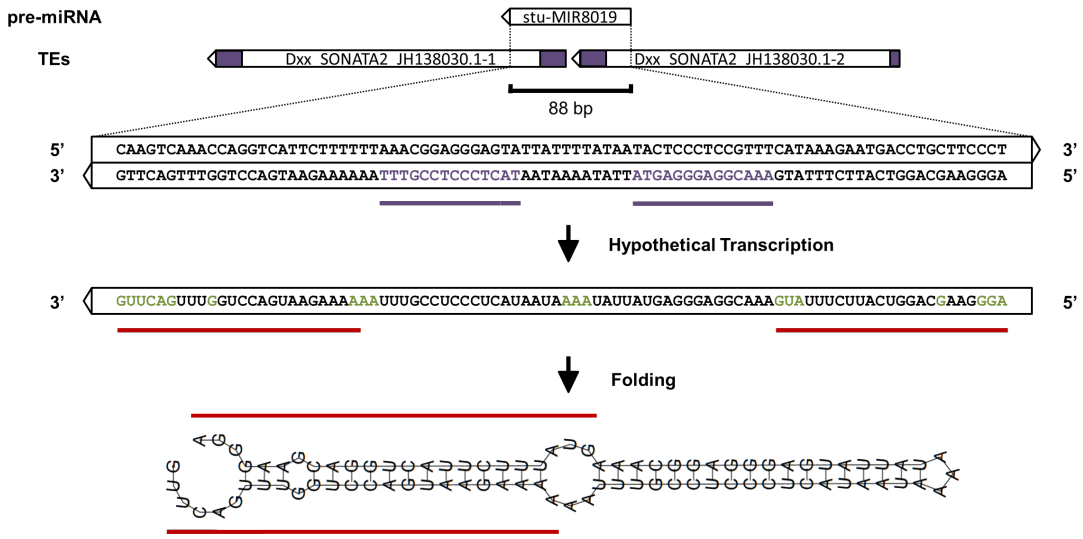
Source: Produced by the authors.

Figura 18 – Representative example of intersection patterns found by our analysis. Pre-miRNAs may intersect long terminal repeat (LTR) regions (a, b) or terminal regions of DNA transposons (c).



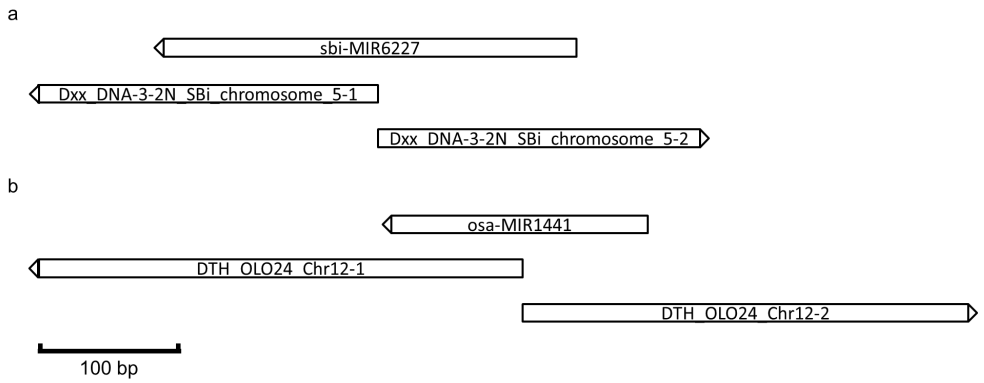
Source: Produced by the authors.

Figura 19 – Example of two juxtaposed non-autonomous DNA transposons overlapping a pre-miRNA in *Solanum tuberosum*. Terminal inverted repeats (TIRs) are highlighted in purple. Dxx_SONATA2_JH138030.1-1 is an intact element and Dxx_SONATA2_JH138030.1-2 lacks part of the 5' TIR. A hairpin structure may emerge due to transcript complementarity. Light green letters show loop regions, and red lines emphasize mature miRNA regions. The secondary structure was plotted using RNAfold Webserver with the Minimum Free Energy (MFE) prediction method (LORENZ et al., 2011).



Source: Produced by the authors.

Figura 20 – Juxtaposed TEs possibly structuring pre-miRNAs in grasses. Both cases show an inverted insertion of the same TE. In *Sorghum bicolor*, two DNA-3-2N_SBi non-autonomous DNA transposons span sbi-MIR6227 locus (a). In *Oryza sativa*, two OLO24 non-autonomous DNA transposons intersect osa-MIR1441 (b).



Source: Produced by the authors.

9 Conclusions

To our knowledge, PlanTE-MIR DB is the first resource storing the putative relationship between TEs and miRNAs in plants. The database delivers, through a user-friendly web interface, several file formats to facilitate understanding and use of the available data. Future versions will update the database to support data provided by other studies. The discovery of new TE-MIRs strongly relies on comprehensive TE annotation, which is still a drawback for several species. Thus, *de novo* TE annotation for organisms for which there is available data in miRBase would be a valuable resource that would promote future discoveries. Also, new releases of Repbase Update, miRBase, and plant species genome assemblies should be considered in the next versions of PlanTE-MIR DB. In conclusion, we present a new resource, PlanTE-MIR DB, which allowed us to find new TE-MIR overlaps. We believe that PlanTE-MIR DB can supply insight into the evolution of TEs and miRNAs that will be of great value to the scientific community interested in this subject.

Parte III

Anotação de MITEs no genoma de *Coffea canephora*

10 Introdução

10.1 Características gerais dos MITEs

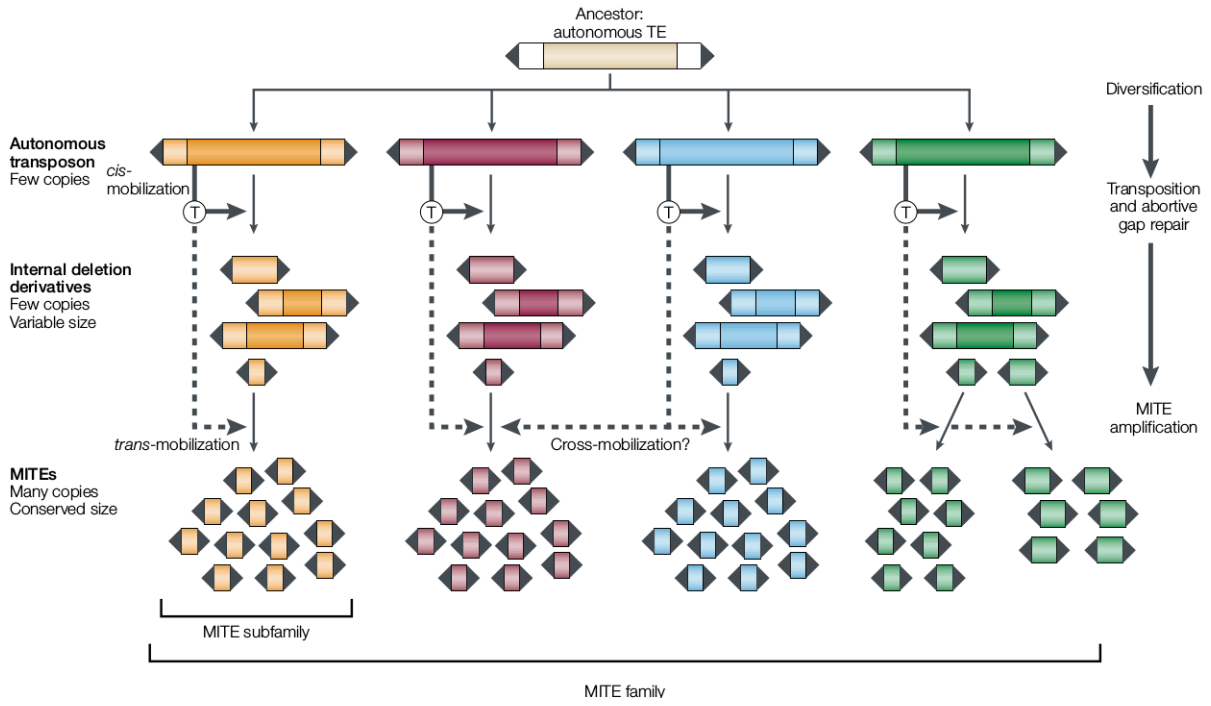
Alguns TEs são conhecidos por não apresentarem a capacidade de se transpor, já que lhes faltam as enzimas necessárias para este processo (ver seção 1.3). É o caso dos MITEs, que são pequenas sequências que variam de poucas dezenas até algumas centenas de nucleotídeos em extensão (100-600 pb) (FESCHOTTE; PRITHAM, 2007), e são compostas por alto conteúdo AT (GUYOT et al., 2009). De acordo com Lisch (2013), são os elementos não autônomos mais comuns nos genomas vegetais, sendo encontrados inseridos preferencialmente em regiões próximas a genes codificadores, como aponta a revisão de Bennetzen e Wang (2014).

O modelo de origem e amplificação de MITEs mais aceito é mostrado na Figura 21. Segundo este modelo, elementos autônomos ancestrais podem dar origem a diversas famílias de elementos, similares apenas por suas TIRs e transposases. Cada uma dessas famílias pode originar elementos não autônomos pela deleção de sua região codificante. Assim são formados os MITEs, que continuam sendo mobilizados, seja por meio das enzimas dos elementos que lhes deram origem, ou de famílias relacionadas (mobilização cruzada) (FESCHOTTE; JIANG; WESSLER, 2002).

Embora possuam TIRs e TSDs muitas vezes idênticas àquelas encontradas em TEs de Classe II (Tabela 3), evidências apontam que em arroz, só algumas famílias de MITEs tenham sido originadas diretamente a partir de deleções de domínios codificantes. Um exemplo é a família *mPing*, derivada da família autônoma *Ping* em *O. sativa* (YANG et al., 2009). Mesmo sem codificar transposase, encontra-se em maior número de cópias em relação aos elementos que lhes deram origem. As enzimas utilizadas para sua transposição são provenientes de elementos *Ping*, e o mesmo fenômeno foi demonstrado para famílias *Stowaway* que podem "emprestar" transposases de elementos não relacionados (mobilização cruzada) (YANG et al., 2009). Por esse motivo, González e Petrov (2009) atribuem a estes elementos a alcunha de "parasitas derradeiros", pois são parasitas de TEs, que outrora foram considerados - indevidamente (FEDOROFF, 2012) - como parasitas do genoma. Vale ressaltar que esta alcunha já não é mais aceita, uma vez que aspectos funcionais do genoma foram atribuídos a esta categoria de TEs.

MITEs, como todo elemento de Classe II, podem ser copiados durante o processo de transposição, aumentando seu número de cópias. Isto pode acontecer durante a replicação do material genético, quando um elemento do DNA molde se transpõe para uma parte adiante do garfo de replicação que ainda não foi copiada. Pode ocorrer também quando uma

Figura 21 – Representação do modelo de origem e amplificação de MITEs. TIRs são representadas por triângulos flanqueadores e transposases representadas por caixas de cor escura.



Fonte: Feschotte, Jiang e Wessler (2002).

Tabela 3 – Características estruturais das principais superfamílias de transposons.

Superfamília	TSD	TIR	Padrão 5' da TIR
Mutator	9-11 pb	longa	GGGTTAAAAACAAAAA
hAT	8 pb	11 pb	TAGGGGTGCAAA
Tc1-Mariner	2 pb (TA)	10-30 pb	CTCCCTCCGTCCCA
CACTA	3 pb	15-100 pb	CACTA
PIF-Harbinger	3 pb	14-25 pb	GGGTTTGTTTGGATA

Fonte: Informação ministrada em curso por Romain Guyot.

lacuna gerada pela excisão de um elemento é reparada, promovendo assim a duplicação do elemento (WICKER et al., 2007).

Em plantas, os MITEs podem contribuir com até cerca de 10% do genoma, como é o caso de *O. sativa*. Entretanto, pode haver grande variação na quantidade desses TEs entre diferentes espécies. Por exemplo, *M. truncatula* apresenta alto conteúdo de MITEs (8,21%), enquanto *S. lycopersicum* (3,44%) e *V. vinifera* (3,02%) apresentam valores intermediários. *Carica papaya* (0,21%) apresenta baixo conteúdo com apenas 538 elementos identificados

(CHEN et al., 2014).

Os polimorfismos causados pela inserção de MITEs são apontados como potenciais marcadores de diversidade genética, devido principalmente ao seu pequeno tamanho, grande quantidade de cópias, e associação com genes (SAMPATH et al., 2014). No gênero *Coffea*, MITEs ajudaram na análise de especiação e na inferência de relação entre espécies próximas (DUBREUIL-TRANCHANT et al., 2011).

10.2 MITEs e pequenos RNAs

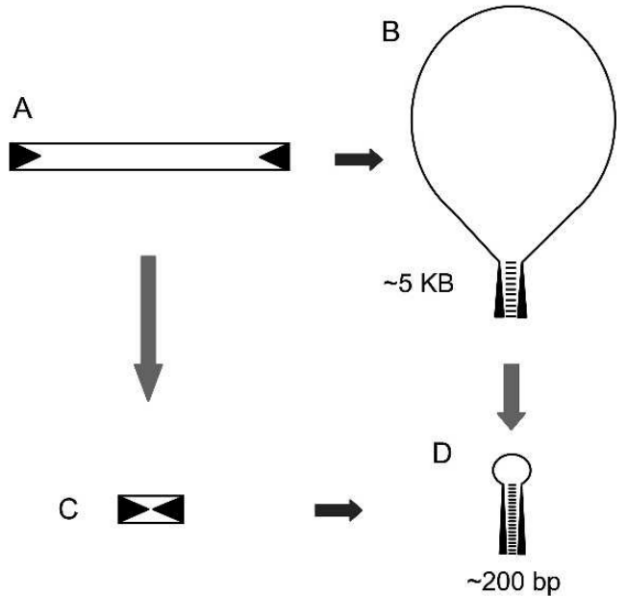
MITEs tipicamente formam estruturas secundárias em haste e alça (do inglês *stem loop*) devido ao caráter complementar de suas extremidades (DUBREUIL-TRANCHANT et al., 2011; LU et al., 2012). Diversos estudos mostram que essas estruturas em grampo podem ser processadas pela mesma maquinaria de biogênese de pequenos RNAs, produzindo agentes funcionais (ver Capítulo 3). Em arroz, sequências relacionadas a MITEs são responsáveis por gerar 23,5% do total de pequenos RNAs, dos quais 70,1% possuem 24 nt de extensão (LU et al., 2012). Similarmente, foi observado que 74% dos pequenos RNAs de *Triticum* sp. relacionados a elementos de Classe II mapeavam-se a MITEs. Entretanto, diferentemente do caso anterior, sequências de 21 nt foram mais representativas (CANTU et al., 2010).

Um estudo aponta que a biogênese de pequenos RNAs de 24 nt a partir de MITEs, referidos como siRNAs associados a repetições (ra-siRNAs, do inglês *Repeat-associated* siRNAs), são mediadas pelas proteínas DICER-LIKE 3 (DCL3) e RNA polimerase 2 dependente de RNA (RDR2, do inglês RNA-*dependent* RNA *polymerase 2*) (KUANG et al., 2009). Esta última é necessária para a geração de precursores compostos de RNAs em dupla fita (dsRNAs, do inglês *double-stranded RNAs*) (XIE; YU, 2015), pré-requisito para a classificação de um pequeno RNA como siRNA (AXTELL, 2013). Diferentemente, outro estudo evidenciou a produção de pequenos RNAs de 24 nt associados a MITEs que são dependentes apenas de DCL3, sugerindo precursão a partir de RNAs em grampo (hpRNAs, do inglês *hairpin* RNAs) (YAN et al., 2011). Este último fenômeno já foi sugerido como modelo para a origem de microRNAs (miRNAs) em plantas, onde um hpRNA poderia produzir tanto miRNAs quanto siRNAs, dependendo da via de biogênese na qual fosse processado (Figura 22). Com o tempo os hpRNAs poderiam ser integrados nas vias de biogênese exclusivas de miRNAs formando produtos funcionais que agiriam no processo de silenciamento gênico pós-transcricional (PTGS, do inglês *Post-transcriptional gene silencing*) (PIRIYAPONGSA; JORDAN, 2008).

Como descrito na seção 2.2, pequenos RNAs de 24 nt estão envolvidos no mecanismo efetor de RdDM, e os sítios alvo de metilação podem ser os próprios MITEs que os geram. Após a metilação desses elementos pode haver também a metilação de regiões adjacentes

(LISCH, 2013). Uma vez que MITEs são encontrados em regiões ricas em genes, é provável que eles influenciem na regulação de diversos genes a eles associados.

Figura 22 – Representação do modelo que propõe o surgimento de hpRNAs derivados de MITEs. Transposon autônomo com TIRs (triângulos) flanqueando a região interna (**A**). Estrutura secundária formada pela complementaridade das TIRs de um transposon autônomo (**B**). MITE derivado da deleção da região interna de um transposon autônomo (**C**). hpRNA derivado do dobramento de um MITE transcrito (**D**).



Fonte: Piriyapongsa e Jordan (2008).

10.3 Métodos para identificação de MITEs

Como já descrito na seção 1.5, existem basicamente duas abordagens para identificação de TEs: identificação por similaridade e identificação *ab initio*. A identificação por similaridade é dependente de sequências já identificadas, e portanto limitada a descoberta de sequências conservadas entre *taxa*. Em contrapartida, a identificação *ab initio* promove a descoberta de novos TEs, mas requer tempo e grande esforço para a filtragem e classificação manual dos dados gerados (HAN; WESSLER, 2010).

Há quatro principais programas capazes de realizar a identificação *de novo* de MITEs: (1) MITE-Hunter (HAN; WESSLER, 2010), (2) Sequências repetitivas com extremidades precisas (RSPB, do inglês *Repetitive sequence with precise boundaries* (LU et al., 2012), (3) MITE Digger (YANG, 2013) e (4) *detect*MITE (YE; JI; LIANG, 2016). O MITE-Hunter é um programa baseado na identificação estrutural de TIRs e TSDs que elimina falsos positivos principalmente por meio da análise das regiões flanqueadoras. Sua

taxa de falsos positivos (4,4%-8,3%) é baixa, mas sua execução é computacionalmente custosa em termos de CPU e RAM. Para analisar o genoma completo de *O. sativa* (≈380 Mb), por exemplo, foram necessárias aproximadamente 44 horas, utilizando cinco CPUs (HAN; WESSLER, 2010). O RSPB consiste em uma coleção de *scripts* escritos em linguagem Perl, que realizam quase os mesmos passos executados pelo MITE-Hunter. Porém, o método identifica primeiramente diversas sequências que possuam alta similaridade em suas bordas, e somente depois verifica a presença de TIRs e TSDs. Essa abordagem permitiu descobrir 37 novas famílias que não foram identificadas pelo MITE-Hunter no mesmo conjunto de dados (LU et al., 2012).

Motivado pela falta de um método eficaz em termos de tempo de processamento e que apresentasse robustez nos resultados, Yang (2013) lançou mais recentemente um novo programa. O MITE Digger procura otimizar o tempo da análise evitando e reduzindo redundâncias. Baseia-se principalmente na pressuposição de que os MITEs estão distribuídos aleatoriamente no genoma, e de que uma única cópia é capaz de representar a maior parte dos indivíduos contidos em uma família. Deste modo, é alta a probabilidade de encontrar um elemento nos primeiros trechos analisados, dada a presença de um alto número de cópias deste mesmo elemento no genoma. Assim, as famílias encontradas são utilizadas para o mascaramento do resto do genoma, economizando tempo. Este programa foi capaz de rodar o mesmo conjunto de dados anteriormente citado em um terço do tempo (≈15 horas) levado pelo MITE-Hunter, com uma taxa de falsos positivos de 1,8%. A máquina utilizada para a aferição possuía apenas 4 CPUs e seu uso de RAM não ultrapassou 150 MB.

A análise em larga escala realizada por Chen et al. (2014) utilizou os três métodos citados, empregando os dois primeiros para os menores genomas, e o último para genomas maiores que 800 Mb. Em todos os casos, a anotação manual das famílias identificadas se fez necessária.

Cada um dos métodos anteriormente citados tem seus pontos fortes, mas somente providenciam ótimos resultados quando utilizados em conjunto. Para resolver este problema, Ye, Ji e Liang (2016) desenvolveram o *detect*MITE. Quando comparado aos seus antecessores (MITE-Hunter, RSPB e MITE Digger), foi capaz de identificar mais famílias, com menor taxa de falsos positivos, e em menor tempo no genoma de *O. sativa*. Além disso, utiliza métodos mais recentes em termos de filtragem de sequências de baixa complexidade (Lempel-Ziv) e clusterização (CD-HIT), facilitando a anotação manual.

Considerando a capacidade do MITE Digger de identificar grande quantidade de MITEs confiáveis, juntamente com sua baixa demanda de recursos computacionais, optou-se por utilizá-lo neste trabalho.

10.4 O genoma de *Coffea canephora*

O café é um dos produtos agrícolas mais consumidos no mundo. É classificado dentro da família Rubiaceae e tem como principais representantes as espécies *C. arabica* ($2n = 4x = 44$ cromossomos) e *C. canephora* ($2n = 2x = 22$ cromossomos), conhecidos popularmente como café arábica e café robusta, respectivamente (SOUZA et al., 2004; DENOEUD et al., 2014). Recentemente foi publicado o sequenciamento do genoma de *C. canephora*, com cobertura de ≈90x, correspondendo a 568,6 Mb (80%) de um total de 710 Mb (DENOEUD et al., 2014). O genoma montado e sua anotação gênica estão disponíveis no *Coffee Genome Hub*[1] (DEREEPER et al., 2014).

Aspectos evolutivos do genoma de *C. canephora* já foram explorados em diversos trabalhos, dentre os quais pode-se citar os seguintes temas: evidências citogenéticas e transcricionais de TEs (LOPES et al., 2008; DUBREUIL-TRANCHANT et al., 2011; YUYAMA et al., 2012; LOPES et al., 2013; DIAS et al., 2015), e anotação de miRNAS (REBIJITH et al., 2013; LOSS-MORAIS et al., 2014; CHAVES et al., 2015). Até o momento, não foi encontrada análise que centrasse seus esforços na identificação em larga escala de MITEs e seus produtos não codificantes. Nesse sentido, teve-se como objetivo realizar a anotação sistemática de MITEs no genoma sequenciado de *C. canephora* e apontar quais famílias podem contribuir para produção de pequenos RNAs.

[1] <http://coffee-genome.org/>

11 Material e Métodos

11.1 Identificação *de novo* de MITEs

Para identificação de novos MITEs no genoma de *C. canephora*, obteve-se a montagem do sequenciamento de seu genoma por meio do *Coffee Genome Hub* (DEREEPER et al., 2014). Foi utilizado o programa MITE Digger (versão 1.4E9) (YANG, 2013) com parâmetros padrão para predizer MITEs em todos os cromossomos (incluindo-se chr0). As sequências obtidas foram então agrupadas utilizando o programa BLASTClust (versão 2.2.26) (DONDOSHANSKY; WOLF, 2002). Sequências que compartilhavam identidade $\geq$ 80% e cobertura $\geq$ 80% foram incluídas em um mesmo conjunto.

Cada um dos conjuntos foi avaliado manualmente para eleger uma sequência representante. Foi dada preferência às sequências que satisfaziam dois critérios principais: (1) TIRs com complementaridade invertida quando observadas no programa Dotter (versão 4.36-17-g677f) (SONNHAMMER; DURBIN, 1995); (2) TIRs conservadas quando observadas no programa Jalview (versão 2.8.0b1) (WATERHOUSE et al., 2009) após passarem por alinhamento múltiplo no MUSCLE (versão 3.8.31) (EDGAR, 2004). Estas sequências representantes foram então mapeadas no genoma de *C. canephora* por meio do programa CENSOR (versão 4.2.28) (JURKA et al., 2005) utilizando como mecanismo de busca o BLASTN (versão 2.2.28+) (CAMACHO et al., 2009). Os resultados deste mapeamento foram filtrados de acordo a regra 80-80-80 proposta por Wicker et al. (2007), e foram extraídas de todas as sequências mais 25 nucleotídeos flanqueadores utilizando a ferramenta *blastdbcmd* do NCBI-BLAST+ *Toolkit* (versão 2.2.28+) (CAMACHO et al., 2009). As últimas etapas descritas foram automatizadas por um *shell script* de autoria própria.

11.2 Estimativa de cobertura e densidade

Foi utilizada a função *genomecoverage* do programa BEDTools (versão 2.17.0) (QUINLAN; HALL, 2010) para realizar os cálculos de cobertura dos MITEs em cada um dos cromossomos e para o genoma como um todo (excluindo-se o chr0). Os arquivos de anotação no formato GFF3 utilizados nesta etapa foram gerados por meio de um *shell script* executando comandos da linguagem AWK. Foram consideradas apenas famílias com no mínimo 10 cópias.

Foram gerados arquivos para plotar a densidade de MITEs e éxons no programa Circos (versão 0.66) (CONNORS et al., 2009) usando as funções *makewindows* e *coverage*

do BEDTools (versão 2.17.0) (QUINLAN; HALL, 2010). Os dados de anotação gênica foram obtidos a partir do *Coffee Genome Hub* (DEREEPER et al., 2014). A densidade foi calculada para janelas deslizantes de 500 Kb com passos de 100 Kb. Para o cálculo do coeficiente de correlação de Spearman (ρ) foram utilizados estes mesmos dados normalizados pelo método *feature scaling* no programa RStudio (TEAM, 2014).

11.3 Limpeza e filtragem da biblioteca de pequenos RNAs

Foram obtidos dados públicos de sequenciamento de pequenos RNAs de *C. canephora* depositados no *Sequence Read Archive* (SRA) do NCBI (acesso SRX273683), gerados pelo estudo de Loss-Morais et al. (2014). Os dados da corrida foram inspecionados para avaliação de qualidade utilizando o programa FastQC (versão v0.10.1) (ANDREWS, 2010). Já que as sequências de adaptadores não foram disponibilizadas, foi necessário identificá-las por meio deste mesmo programa. Após dedução das sequências de adaptadores (ver Apêndice B), utilizou-se o programa Trimmomatic (versão 0.35) (BOLGER; LOHSE; USADEL, 2014) para removê-las e selecionar apenas leituras com valor médio de Phred ≥ 30. Utilizando a linguagem AWK, foi possível filtrá-los por tamanho, sendo mantidas apenas leituras de 16 a 26 nt.

As leituras remanescentes foram então colapsadas utilizando a ferramenta *fastx_collapser* do FASTX-Toolkit (versão 0.0.6) (GORDON; HANNON, 2010) e mapeadas contra sequências de DNA cloroplastídico, DNA mitocondrial, rRNAs, tRNAs e snoRNAs (ver Apêndice B) utilizando Bowtie2 (versão 2.1.0) (LANGMEAD; SALZBERG, 2012). As leituras mapeadas contra essas sequências sem *gaps* e *mismatches* foram então eliminadas da análise.

11.4 Identificação das famílias relacionadas a pequenos RNAs

Todas as sequências representantes de MITEs com pelo menos 10 cópias no genoma foram utilizadas como referência para o mapeamento de leituras de pequenos RNAs. Para o mapeamento utilizou-se o programa Bowtie2 (versão 2.1.0) (LANGMEAD; SALZBERG, 2012) com parâmetros padrão. O arquivo de saída desta etapa foi filtrado para manter apenas leituras mapeadas com até 1 *mismatch* e sem *gaps*. As sequências representantes que tiveram pelo menos 10 leituras alinhadas foram então inspecionadas e classificadas manualmente. Uma nova rodada de mapeamento foi então executada utilizando como referência todos os membros das famílias que foram classificadas com êxito. Nesta etapa, foram utilizadas as mesmas leituras. O arquivo de saída foi então filtrado para manter apenas alinhamentos com até dois *mismatches* e sem *gaps*, e analisado para estimar o número de leituras alinhadas por família.

Para o cálculo do coeficiente de correlação de Spearman (ρ) foram utilizados os dados de contagem de leituras e cópias normalizados pelo método *feature scaling* no programa RStudio (TEAM, 2014).

11.5 Classificação dos MITEs

As sequências representantes que tiveram ao menos 10 leituras de pequenos RNAs alinhadas foram selecionadas para inspeção manual. A partir dos 50 membros mais similares à sequência representante de cada família, mais 25 nt flanqueadores, foi realizado alinhamento múltiplo utilizando MUSCLE (versão 3.8.31) (EDGAR, 2004). Os alinhamentos foram observados na interface gráfica do Jalview (versão 2.8.0b1) (WATERHOUSE et al., 2009), elegendo novas sequências representantes da mesma forma descrita na seção 11.1, com o auxílio do programa Dotter (versão 4.36-17-g677f) (SONNHAMMER; DURBIN, 1995). Procurou-se padrões de TIRs e TSDs que caracterizassem cada uma das superfamílias de elementos de Classe II (transposons), com critérios baseados no trabalho de Wicker et al. (2007) e no protocolo fornecido por Romain Guyot[1] (Tabela 3). As sequências representantes de cada família foram novamente mapeadas pelo mesmo método descrito na seção 11.1. O número de cópias para as famílias anotadas foi então atualizado.

O nome das famílias foi dado seguindo o modelo de Chen et al. (2014). Faz-se referência ao nome da espécie (*Coffea canephora*) seguido pelas iniciais do nome de suas superfamílias (HAT para *hAT* e Mu para *Mutator*), ou pelas iniciais das famílias relacionadas (Sto para *Stowaway* e To para *Tourist*). Também foi adicionado um número cardinal para diferenciar cada uma delas (*e.g. CocHAT1*).

11.6 Inferência de similaridade para famílias anotadas

Os três membros mais similares às sequências representantes de cada família foram utilizados para a dedução de fenogramas. Para isso, as sequências foram separadas em superfamílias (DTA, DTH, DTM e DTT) e alinhadas globalmente utilizando MUSCLE (versão 3.8.31) (EDGAR, 2004) com parâmetros padrão. Os alinhamentos gerados foram usados para inferência de fenogramas pelo método de máxima verossimilhança aleatória acelerada do programa RAxML (versão 8.2.4) (STAMATAKIS, 2014). A análise foi executada com *bootstrap* de 2500 replicatas e modelo de substituição GTRGAMMA. Os fenogramas foram plotados no RStudio (TEAM, 2014) utilizando o pacote ggtree, disponível no repositório Bioconductor (GENTLEMAN et al., 2004).

[1] Pesquisador do *Institute de recherche pour le développement* - IRD, Montpellier, França.

11.7 Evidências transcricionais e análises comparativas

Os membros das famílias anotadas foram utilizados como sondas eletrônicas contra uma montagem de etiquetas de sequências expressas (ESTs, do inglês *Expressed Sequence Tags*) de *C. canephora* gerada pelo trabalho de Mondego et al. (2011), para identificação de membros com evidências transcricionais. As sequências foram comparadas utilizando o programa CENSOR (versão 4.2.28) (JURKA et al., 2005) por meio do mecanismo de busca BLASTN (versão 2.2.28+) (CAMACHO et al., 2009). Os resultados deste mapeamento foram filtrados de acordo a regra 80-80-80 proposta por Wicker et al. (2007).

As famílias também foram comparadas com sequências de MITEs anteriormente anotadas por Guyot et al. (2009) (acesso EU164537.1, no NCBI). Para isso utilizou-se BLASTN (versão 2.2.28+) (CAMACHO et al., 2009) com parâmetros padrão e o arquivo de saída foi filtrado para mostrar apenas o melhor resultado para cada MITE previamente anotado.

11.8 Análise da região de inserção dos MITEs anotados

Para estimar se as regiões de inserção dos MITEs tinham relação de distância física com genes anotados, foram utilizados arquivos de anotação no formato GFF3 contendo a anotação geral gerada usando as sequências representantes (ver seção 11.2). A anotação gênica foi a mesma apontada pela mesma seção.

Os dois arquivos foram confrontados por meio do programa BEDTools (versão v2.25.0) (QUINLAN, 2014). A função *intersect* foi utilizada para encontrar MITEs inseridos em meio a genes ou nas regiões flanqueadoras com no mínimo 1 pb de sobreposição. Então as entradas correspondentes a este último passo foram removidas. A função *window* foi utilizada para encontrar MITEs inseridos em regiões flanqueadoras de até 1 Kb a montante ou a jusante dos genes.

12 Resultados e discussão

12.1 Anotação de MITEs

Foram identificadas 163 sequências de MITEs e, após a remoção de redundâncias, restaram 150 sequências que serviram de sementes para a anotação em larga escala. Ao todo, 107 famílias com pelo menos 10 cópias foram identificadas nos 11 cromossomos de *Coffea canephora*. Foi obtido um total de 16.477 sequências de MITEs putativos, correspondendo a 1,48% do genoma (Tabela 4). As famílias apresentaram grande variação no número de cópias, chegando até 1.786 para a mais numerosa (Figura 23).

Tabela 4 – Estimativa da cobertura dos MITEs no genoma de *C. canephora*. Nota-se que a cobertura dos MITEs no genoma é homogênea, sendo próxima de 1,5% em cada cromossomo.

Cromossomo	Bases MITEs[a]	Bases genoma[b]	MITEs/Genoma[c]
chr1	553.665	38.193.400	0,0144964
chr2	900.785	54.522.928	0,0165212
chr3	458.081	32.030.951	0,0143012
chr4	427.825	28.191.985	0,0151754
chr5	392.505	29.137.935	0,0134706
chr6	580.788	37.293.965	0,0155732
chr7	434.929	29.833.120	0,0145787
chr8	445.298	31.585.744	0,0140981
chr9	294.153	22.352.177	0,0131599
chr10	423.639	27.624.748	0,0153355
chr11	463.376	33.540.656	0,0138154
Total	5.375.044	364.307.609	0,0147541

[a]Cobertura dos MITEs no genoma em número de pares de bases. [b]Tamanho do genoma em número de pares de bases. [c]Razão entre o número de pares de bases ocupadas por MITEs e o número total de pares de bases no genoma.

Fonte: Produzido pelo autor.

Para as sequências obtidas, encontrou-se conteúdo AT médio de 71,06 ± 6,00% (desvio padrão), valor semelhante aos observados para elementos da família *Monkey King* em *Arabidopsis thaliana* (67,0 ± 1,4%), *A. lyrata* (65,7 ± 2,3%) e *Brassica rapa* (67,0 ± 5,0%) (DAI et al., 2015). Além disso, o presente trabalho encontrou associação entre a

densidade de MITEs e éxons (ρ = 0,8457; P < 0,01) em escala genômica (Figura 24), confirmando a correlação entre MITEs e genes evidenciada por Guyot et al. (2009) em um clone BAC sequenciado da mesma espécie.

Foram identificadas cinco famílias previamente anotadas pelo trabalho de Guyot et al. (2009). As três famílias com maior número de cópias (*CocTo1*, *CocTo2* e *CocSto1*) são similares a *Gerard1* (E-value = 2e-74; Identidade = 98,24%), *Jose2* (E-value = 1e-71; Identidade = 97,42%) e *Anis1* (E-value = 5e-164; Identidade = 97,89%), respectivamente. *CocTo4* e *CocSto3* também já foram anotadas anteriormente pelos mesmos autores, sendo similares a *Jose1* (E-value = 3e-63; Identidade = 97,83%) e *Anis3* (E-value = 4e-175; Identidade = 97,23%), respectivamente.

A família *Alex-1* identificada nos estudos de Dubreuil-Tranchant et al. (2011), amplamente distribuída no gênero *Coffea*, também foi identificada neste trabalho. Ela é equivalente a família *CocTo5* identificada neste trabalho (E-value = 1e-77; Identidade = 95,51%), com um total de 309 cópias (Figura 23). As sequências representantes de MITEs identificadas em *C. canephora* também foram submetidas a análise de similaridade por BLAST com o banco de dados P-MITE, mas não obteve-se resultados significativos.

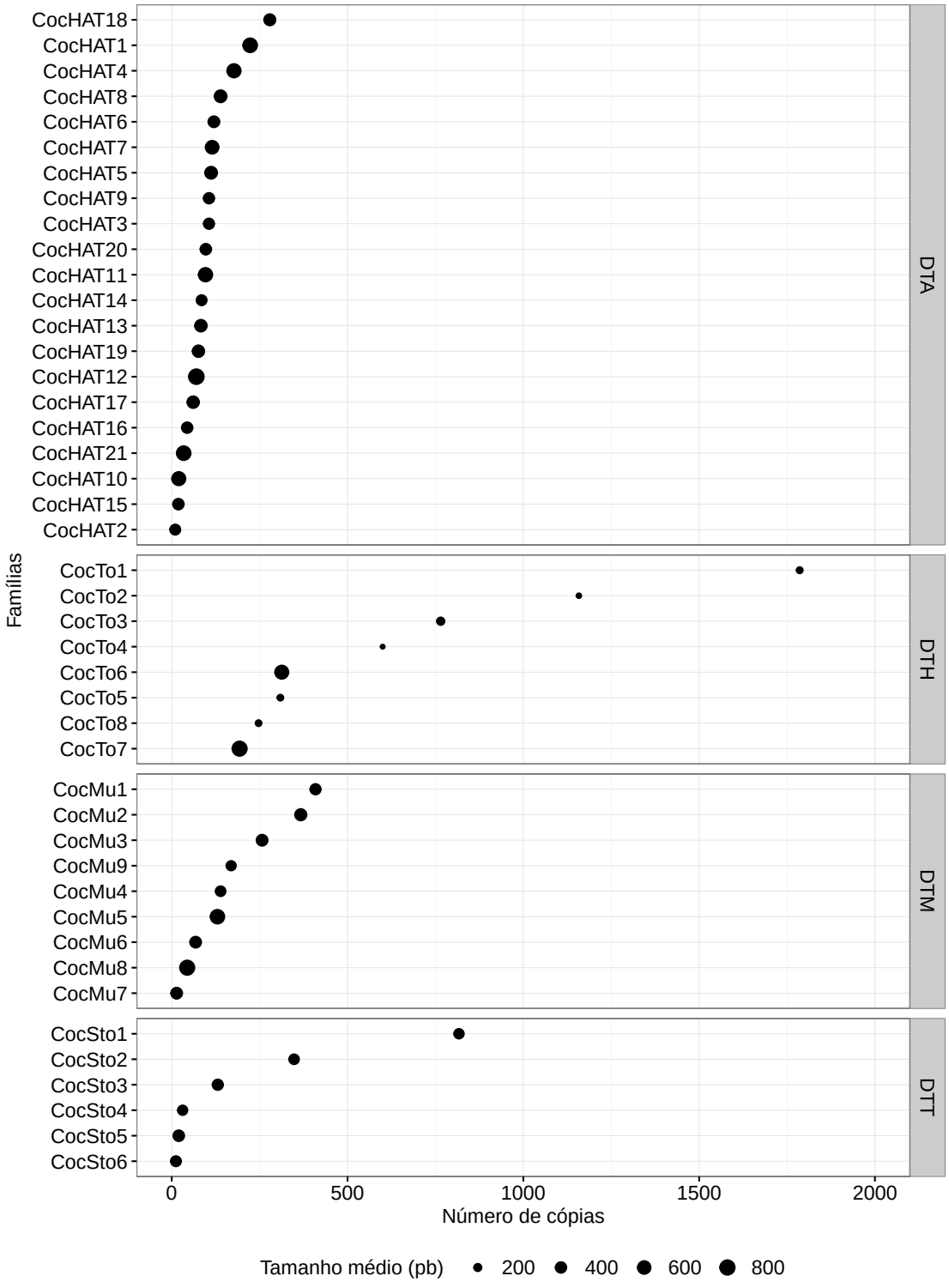
12.2 MITEs podem dar origem a pequenos RNAs

Foram encontradas 52 famílias de MITEs com pelo menos 10 leituras de pequenos RNAs mapeados. Foi possível classificar 44 delas dentro das superfamílias *hAT*, *PIF-Harbinger*, *Mutator* e *Tc1-Mariner*. Do total de leituras pré-processadas (6.074.810), 204.731 (3,37%) foram alinhadas aos membros das famílias classificadas.

Considerando o total de leituras mapeadas, a família *CocTo2* foi a que apresentou o maior número sequências alinhadas (9,49%), seguida por *CocSto1* (9,31%), *CocMu1* (7,30%), *CocHAT9* (6,14%) e *CocTo1* (5,48%). Os valores normalizados pelo número de cópias de cada família mostram que a correspondência entre o número de leituras alinhadas e o número de cópias não é obrigatória para todas elas (Figura 25). Há moderada correlação entre as duas variáveis (ρ = 0,5257; P < 0,01).

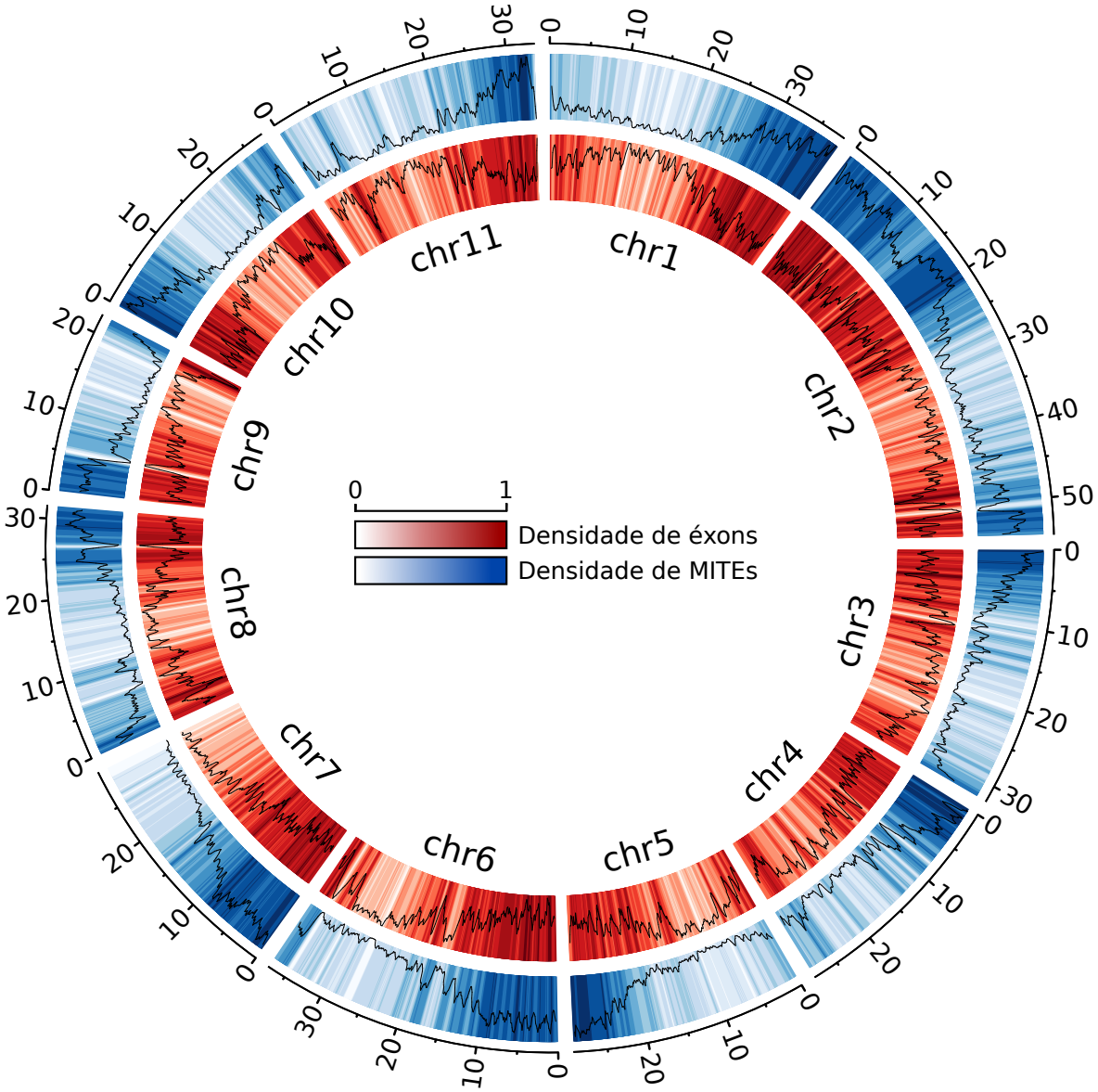
Pode-se notar a predominância de leituras de 24 nt mapeadas às famílias (Figura 25), resultado que corrobora com trabalho anterior na família Solanaceae (KUANG et al., 2009), amplamente utilizada nos estudos comparativos com espécies de café. Além disso, outros trabalhos evidenciaram predominância de leituras de 24 nt associados aos MITEs em *O. sativa* (LU et al., 2012; WEI et al., 2014). Por outro lado, em *Triticum* spp. foi evidenciada predominância de leituras de 21 nt associadas aos MITEs (CANTU et al., 2010). Portanto, mesmo que em menor número de leituras, a possibilidade de MITEs gerarem miRNAs não é descartada.

Figura 23 – Número de cópias e tamanho médio das famílias de MITEs que possivelmente dão origem a pequenos RNAs no genoma de *C. canephora*. São mostradas apenas as famílias que puderam ser classificadas. Percebe-se que a maior parte dessas famílias possui menos de 500 cópias. O tamanho dos pontos indica o tamanho médio das sequências das famílias. DTA: superfamília *hAT*; DTH: superfamília *PIF-Harbinger*; DTM: superfamília *Mutator*; DTT: superfamília *Tc1-Mariner*. Gráfico plotado utilizando RStudio (TEAM, 2014).



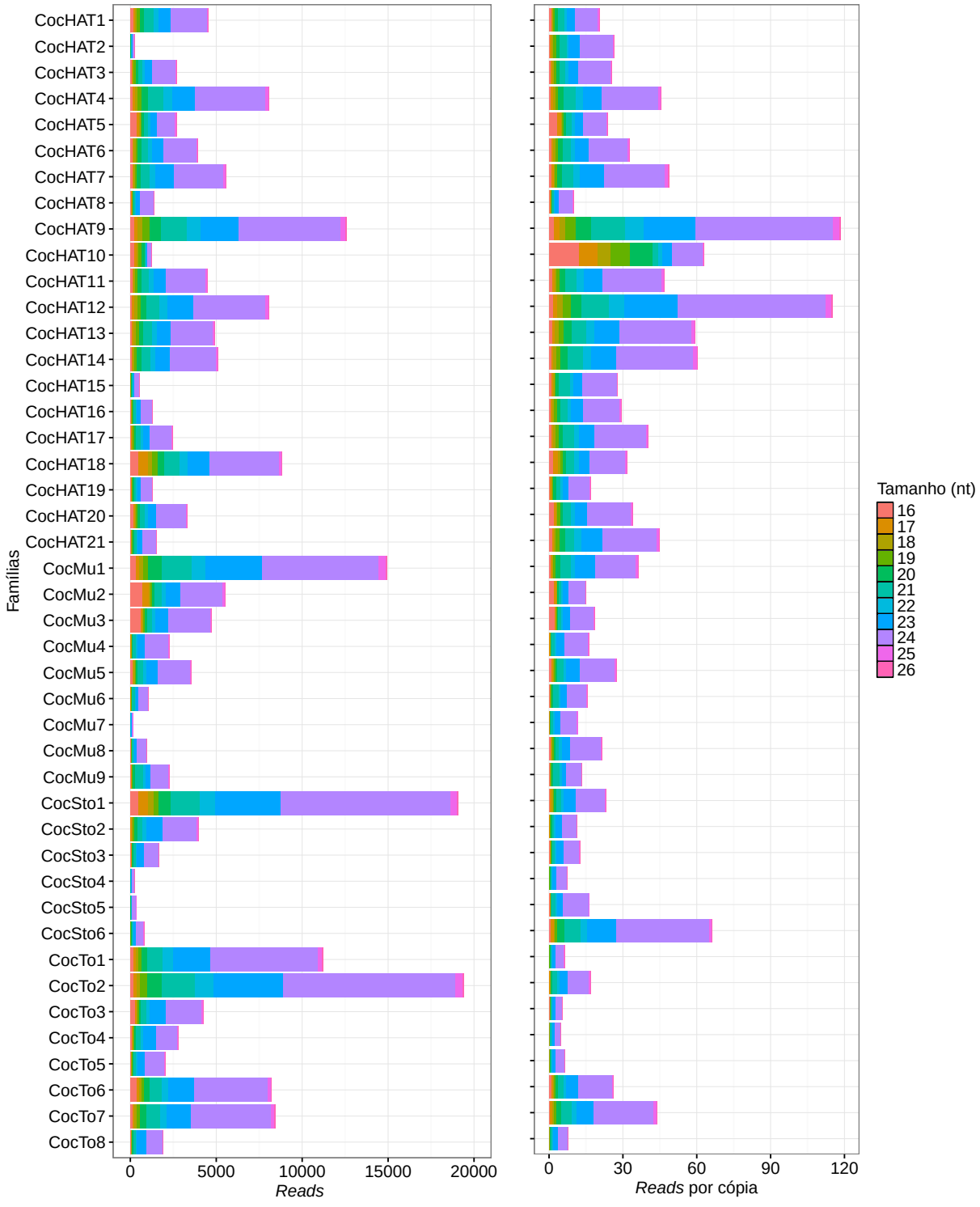
Fonte: Produzido pelo autor.

Figura 24 – Densidade de MITEs e éxons no genoma de *C. canephora*. No anel externo, vê-se um *heatmap* que indica as regiões com maior quantidade de éxons em azul. No anel interno, o *heatmap* indica as regiões com maior quantidade de MITEs em vermelho. As linhas acompanham os níveis demonstrados pelo *heatmap*, tendo os seus maiores valores no lado externo e interno, para cada um dos anéis, respectivamente. Percebe-se que há correspondência entre as regiões ricas em éxons e as regiões altamente colonizadas por MITEs.



Fonte: Produzido pelo autor.

Figura 25 – Número de leituras de pequenos RNA contabilizados por família de MITEs. À esquerda, observa-se o número absoluto de leituras alinhadas por família. À direita, vê-se a quantidade de leituras alinhadas normalizada pelo número de cópias de cada família. Ressalta-se a predominância de leituras com 24 nt de extensão.



Fonte: Produzido pelo autor.

Algumas famílias tiveram membros representados em ESTs do trabalho de Mondego et al. (2011). São elas *CocTo1* (2), *CocTo2* (3), *CocTo3* (1), *CocTo4* (3), *CocTo5* (1), *CocSto1* (1) e *CocSto3* (1). Em vista do alto número de sequências de MITEs associadas a regiões gênicas (Tabela 5), é bem provável que estes elementos tenham sido transcritos juntamente a genes. MITEs transcritos junto a genes podem influenciar na regulação pós-transcricional se a inserção tiver ocorrido na porção 3' não traduzida do gene, de acordo com o fenômeno relatado pela revisão de Feschotte (2008). Nota-se que além de leituras de 24 nt, também há quantidade relevante de leituras de 21 nt nas famílias dos elementos transcritos, as quais podem participar ativamente do mecanismo PTGS.

Durante a inspeção manual das sequências montadas de ESTs, foi possível verificar que um elemento da família *CocTo4* encontra-se inserido na fita oposta na região 3' da sequência montada Contig2865. Uma busca rápida contra o Pfam (FINN et al., 2015) mostrou que esta sequência codifica uma subunidade do fator de transcrição TFIID (acesso PF02291; E-value = 1,1e-44). Entretanto, um trabalho anterior no gênero *Coffea* aponta que transcritos codificantes contendo TEs podem não ter produtos funcionais (LOPES et al., 2008).

As demais sequências de ESTs também foram comparadas ao Pfam, mas não apresentaram proteínas conhecidas. A posição de inserção de MITEs nessas sequências é variável e elas apresentaram uma grande quantidade de códons de parada. É possível que essas sequências sejam RNAs não codificantes. Na família *CocSto3*, um único membro parece ser de fato transcrito, ocupando a maior parte da sequência Contig4485.

Tabela 5 – Quantidade de MITEs inseridos em posições relativas a genes. Nota-se que nas superfamílias anotadas há grande representatividade das inserções na posição a jusante aos genes.

Superfamília	Inserções em genes[a]	Inserções a montante[b]	Inserções a jusante[c]
hAT	151 (7,32%)	318 (15,42%)	220 (10,67%)
PIF-Harbinger	477 (8,88%)	1095 (20,39%)	609 (11,34%)
Mutator	153 (9,58%)	241 (15,09%)	156 (9,77%)
Tc1-Mariner	67 (4,97%)	260 (19,30%)	148 (10,99%)

[a]Inserções dentro de genes ou em sobreposição às suas extremidades. [b]Inserções localizadas na região de 1 Kb a montante dos genes. [c]Inserções localizadas na região de 1 Kb a jusante dos genes. Os valores percentuais são relativos ao número total de cópias de cada superfamília.

Fonte: Produzido pelo autor.

12.3 Análise de similaridade das famílias anotadas

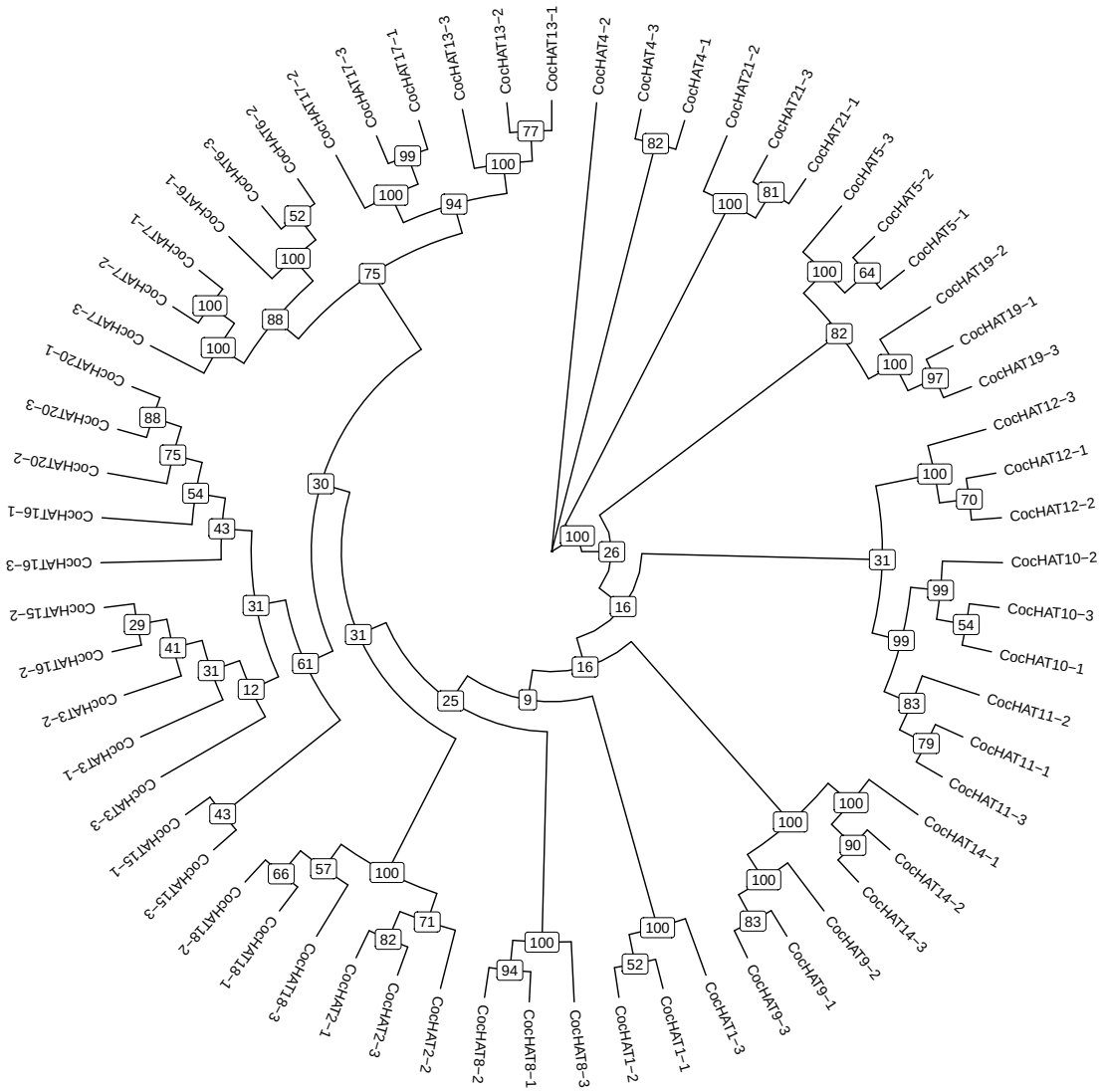
Fenogramas foram construídos para as quatro superfamílias identificadas: *hAT*, *PIF-Harbinger*, *Mutator* e *Tc1-Mariner*. Três membros representam cada família nos fenogramas. A maneira esperada de agrupamento das três sequências é em grupos que contenham só elas, separadas das sequências de outras famílias.

Na superfamília *hAT*, as famílias *CocHAT13*, *CocHAT15*, *CocHAT16* foram as únicas que não se agruparam como esperado (Figura 26). CocHAT16-1 e CocHAT16-3 agruparam-se com membros da família *CocHAT20*. CocHAT15-2 e CocHAT16-2, diferentemente, agruparam-se com membros da família *CocHAT13*. Os membros remanescentes da família *CocHAT15* agruparam-se em um ramo separado. Essas famílias podem ser derivadas de uma mesma família de elementos autônomos e talvez a classificação delas como uma mesma família possa ser aplicada.

Para a superfamília *PIF-Harbinger* também foi evidenciada separação de membros de uma mesma família. CocTo8-2 separou-se dos membros de sua família na base do fenograma, posicionando-se como grupo vizinho do ramo que reúne as demais famílias. CocTo4-2 foi agrupada com CocTo6 e os remanescentes da família se uniram em um ramo como grupo vizinho (Figura 27). Possivelmente, *CocTo4* e *CocTo6* deveriam ser também agrupados em uma mesma família. A mesma topologia foi identificada no ramo que reúne as famílias *CocMu1* e *CocMu3* na superfamília *Mutator* (Figura 28), onde novamente poderia ser proposta uma única família.

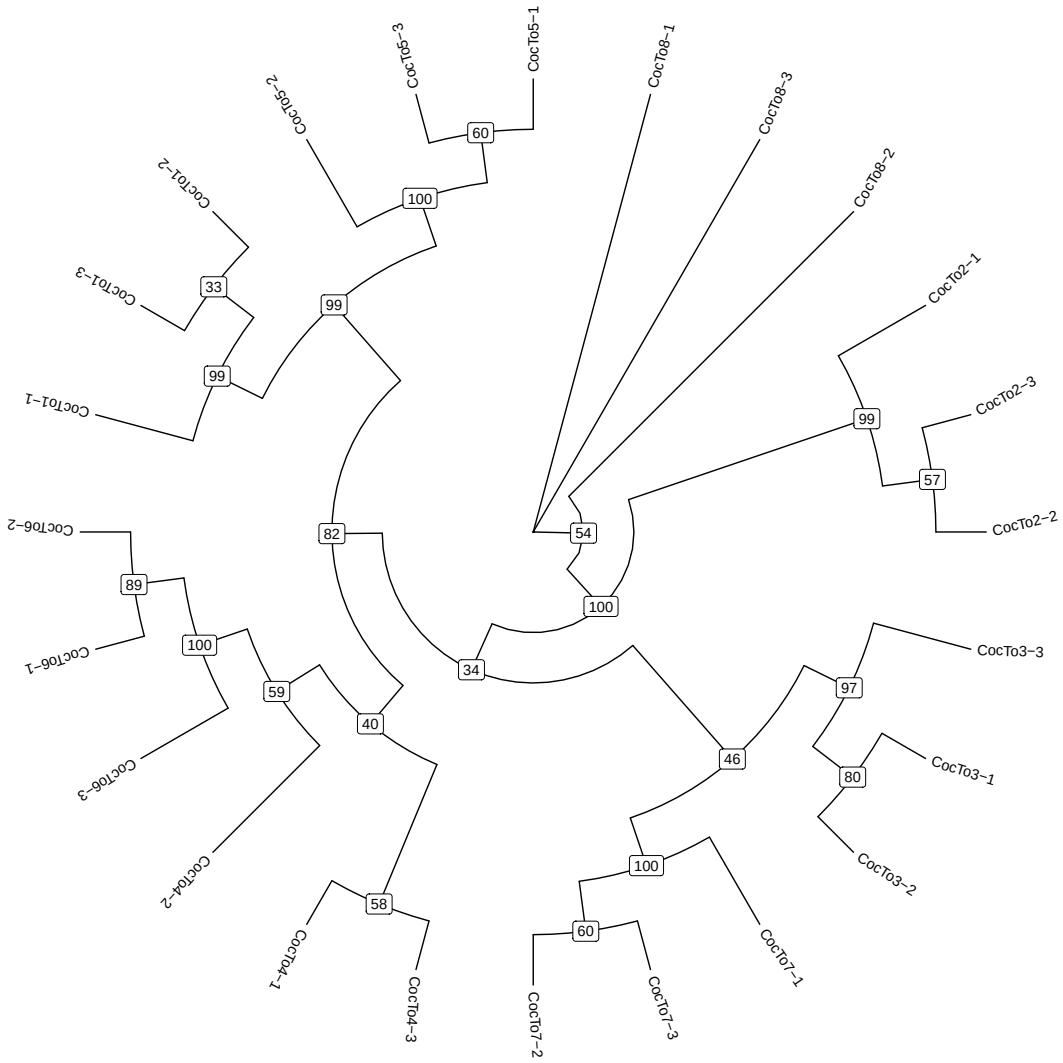
O fenograma da superfamília *Tc1-Mariner* não apresentou conflitos, pois os membros selecionados de cada família reuniram-se da maneira esperada (Figura 29).

Figura 26 – Fenograma para membros da superfamília *hAT*.



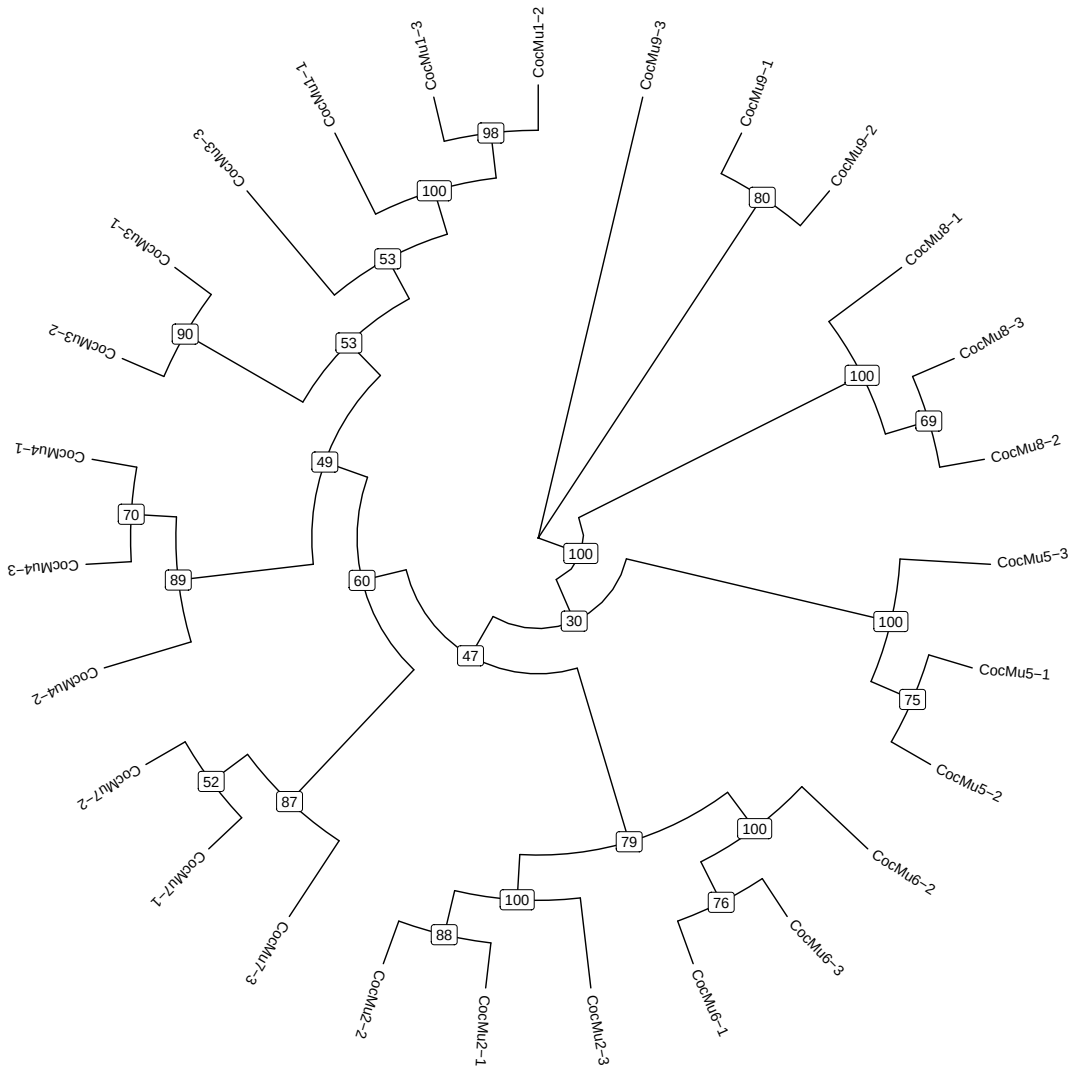
Fonte: Produzido pelo autor.

Figura 27 – Fenograma para membros da superfamília *PIF-Harbinger*.



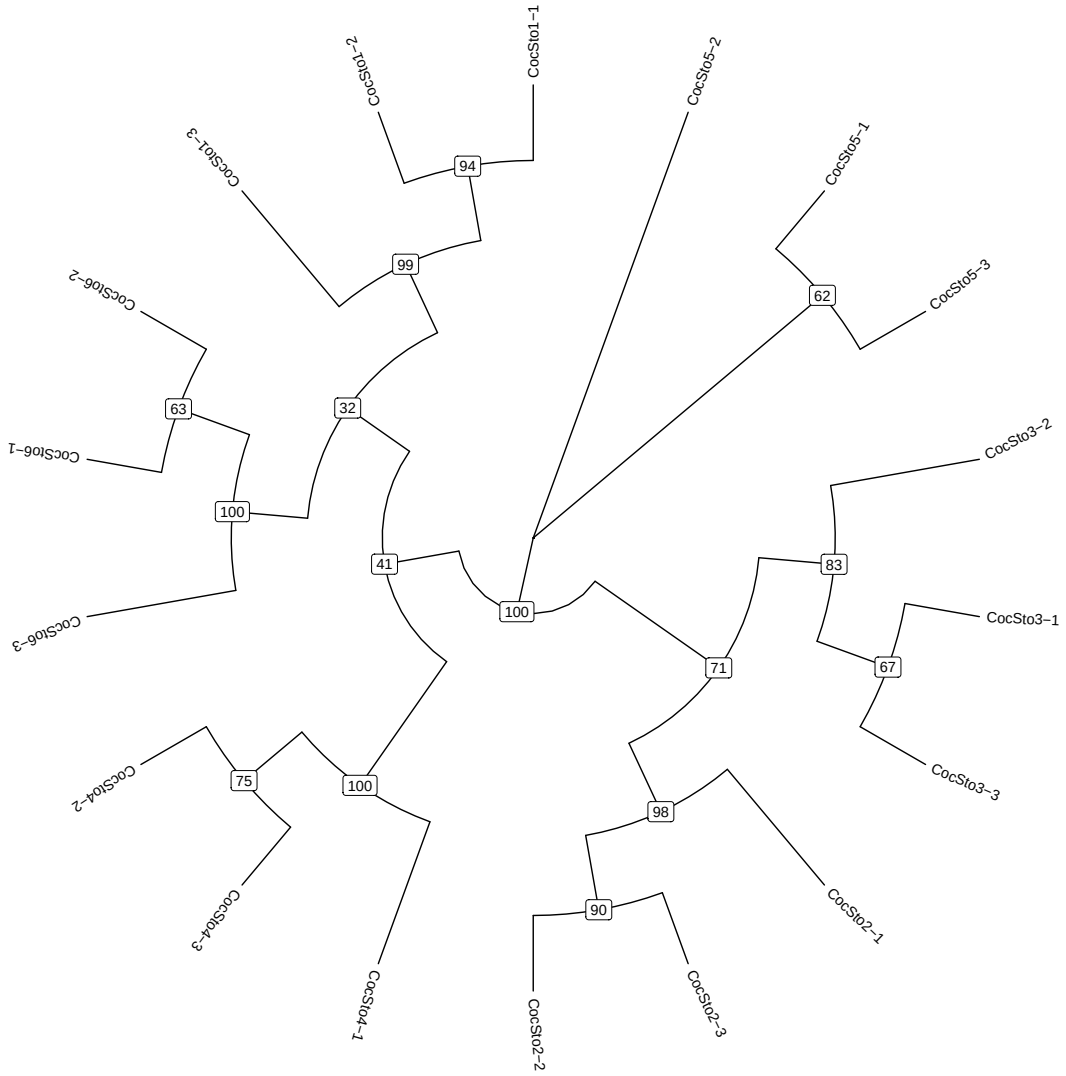
Fonte: Produzido pelo autor.

Figura 28 – Fenograma para membros da superfamília *Mutator*.



Fonte: Produzido pelo autor.

Figura 29 – Fenograma para membros da superfamília *Tc1-Mariner*.



Fonte: Produzido pelo autor.

13 Conclusão

O genoma de *C. canephora* abriga grande quantidade de MITEs frequentemente associados a regiões gênicas, e pelo menos 44 famílias identificadas são potenciais geradoras de pequenos RNAs. Entre os pequenos RNAs gerados, os de 24 nt foram os mais representados, levantando a hipótese de atuação de MITEs em mecanismos epigenéticos. Os pequenos RNAs de 21 nt, embora em menor número, também são importantes, uma vez que podem participar de mecanismos de regulação pós-transcricionais, agindo em mRNAs que carregam MITEs na região 3' não traduzida. Além disso, a identificação de um elemento da família *CocTo4* na porção 3' de um transcrito codificante reforça a necessidade de novos estudos das redes de regulação influenciadas por TEs nesse genoma à luz dos dados genômicos publicados.

Investigações aprofundadas da estrutura secundária das sequências de MITEs, e a validação experimental da maquinaria de biogênese de miRNAs e siRNAs, bem como dos seus produtos, podem contribuir para a elucidação da participação desses elementos nos processos regulatórios.

Referências

ALLEN, E. et al. Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nature genetics*, v. 36, n. 12, p. 1282–1290, 2004. ISSN 1061-4036. Citado na página 37.

ANDREWS, S. *FastQC: A quality control tool for high throughput sequence data.* 2010. http://www.bioinformatics.babraham.ac.uk/projects/ p. Citado na página 58.

AXTELL, M. J. Classification and comparison of small RNAs from plants. *Annual review of plant biology*, v. 64, n. January, 2013. ISSN 1545-2123. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23330790>. Citado 4 vezes nas páginas 24, 27, 37 e 53.

BAIDOURI, M. E.; PANAUD, O. Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biology and Evolution*, v. 5, n. 5, p. 954–965, 2013. ISSN 17596653. Citado na página 37.

BARRERA-FIGUEROA, B. E. et al. High throughput sequencing reveals novel and abiotic stress-regulated microRNAs in the inflorescences of rice. *BMC Plant Biology*, BMC Plant Biology, v. 12, n. 1, p. 132, 2012. ISSN 1471-2229. Citado na página 37.

BENNETZEN, J. L.; WANG, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual review of plant biology*, v. 65, n. February, 2014. ISSN 1545-2123. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24579996>. Citado 7 vezes nas páginas 13, 18, 19, 21, 29, 37 e 51.

BENSON, D. A. et al. GenBank. *Nucleic Acids Research*, Oxford University Press, v. 43, n. Database issue, p. D30, 2015. Citado na página 33.

BHATTACHARYYA, M. K. et al. The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. *Cell*, v. 60, n. 1, p. 115–22, jan 1990. ISSN 0092-8674. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/2153053>. Citado na página 19.

BIÉMONT, C. A brief history of the status of transposable elements: From junk DNA to major players in evolution. *Genetics*, v. 186, n. 4, p. 1085–1093, 2010. ISSN 00166731. Citado na página 13.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, v. 30, n. 15, p. 2114–2120, 2014. ISSN 1367-4803. Disponível em: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btu170>. Citado na página 58.

BOND, D. M.; BAULCOMBE, D. C. Small RNAs and heritable epigenetic variation in plants. *Trends in Cell Biology*, Elsevier Ltd, v. 24, n. 2, p. 100–107, 2014. ISSN 09628924. Disponível em: <http://dx.doi.org/10.1016/j.tcb.2013.08.001>. Citado na página 28.

BRAUER, E. K.; SINGH, D. K.; POPESCU, S. C. Next-generation plant science: putting big data to work. *Genome biology*, v. 15, n. 1, p. 301, 2014. ISSN 1465-6914. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24423368>. Citado na página 32.

BROWN, J. W. S. et al. Plant snoRNA database. *Nucleic Acids Research*, v. 31, n. 1, p. 432–435, 2003. ISSN 03051048. Citado na página 89.

BUDAK, H.; AKPINAR, B. A. Plant miRNAs: biogenesis, organization and origins. *Functional & Integrative Genomics*, v. 15, n. 5, p. 523–531, 2015. ISSN 1438-793X. Disponível em: <http://link.springer.com/10.1007/s10142-015-0451-2>. Citado na página 37.

CAMACHO, C. et al. BLAST+: architecture and applications. *BMC bioinformatics*, v. 10, p. 421, 2009. ISSN 1471-2105. Citado 3 vezes nas páginas 39, 57 e 60.

CANTU, D. et al. Small RNAs, DNA methylation and transposable elements in wheat. *BMC genomics*, v. 11, p. 408, 2010. ISSN 1471-2164. Citado 2 vezes nas páginas 53 e 62.

CHAN, P. P.; LOWE, T. M. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research*, v. 37, n. SUPPL. 1, p. 93–97, 2009. ISSN 03051048. Citado na página 89.

CHAVES, S. S. et al. New Insights on Coffea miRNAs: Features and Evolutionary Conservation. *Applied Biochemistry and Biotechnology*, v. 177, n. 4, p. 879–908, 2015. ISSN 0273-2289. Disponível em: <http://link.springer.com/10.1007/s12010-015-1785-x>. Citado na página 56.

CHEN, J. et al. P-MITE: A database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Research*, v. 42, n. D1, 2014. ISSN 03051048. Citado 4 vezes nas páginas 32, 53, 55 e 59.

COGNAT, V. et al. PlantRNA, a database for tRNAs of photosynthetic eukaryotes. *Nucleic Acids Research*, v. 41, n. D1, p. 1–7, 2013. ISSN 03051048. Citado na página 89.

CONNORS, J. et al. Circos : An information aesthetic for comparative genomics. n. 604, p. 1639–1645, 2009. Citado na página 57.

DAI, S. et al. Widespread and evolutionary analysis of a MITE family Monkey King in Brassicaceae. *BMC plant biology*, v. 15, n. 1, p. 149, jan 2015. ISSN 1471-2229. Disponível em: <http://www.biomedcentral.com/1471-2229/15/149>. Citado na página 61.

DEBAT, H. J.; DUCASSE, D. a. Plant microRNAs: Recent Advances and Future Challenges. *Plant Molecular Biology Reporter*, 2014. ISSN 0735-9640. Disponível em: <http://link.springer.com/10.1007/s11105-014-0727-z>. Citado 3 vezes nas páginas 24, 25 e 29.

DENOEUD, F. et al. The Coffee Genome Provides Insight into the Convergent Evolution of Caffeine Biosynthesis. *Science*, v. 345, n. 6201, p. In press, 2014. ISSN 10959203. Citado na página 56.

DEREEPER, A. et al. The coffee genome hub: a resource for coffee genomes. *Nucleic acids research*, v. 43, n. November 2014, p. 1–8, 2014. ISSN 1362-4962. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/25392413>. Citado 3 vezes nas páginas 56, 57 e 58.

DIAS, E. S. et al. Large distribution and high sequence identity of a Copia-type retrotransposon in angiosperm families. *Plant Molecular Biology*, Springer Netherlands, v. 89, n. 1-2, p. 83–97, 2015. ISSN 0167-4412. Disponível em: <http://link.springer.com/10.1007/s11103-015-0352-8>. Citado na página 56.

DONDOSHANSKY, I.; WOLF, Y. Blastclust (ncbi software development toolkit). *NCBI, Bethesda, Md*, 2002. Citado na página 57.

DUBREUIL-TRANCHANT, C. et al. Site-Specific Insertion Polymorphism of the MITE Alex-1 in the Genus Coffea Suggests Interspecific Gene Flow. *International Journal of Evolutionary Biology*, v. 2011, p. 1–9, 2011. ISSN 2090-052X. Disponível em: <http://www.hindawi.com/journals/ijeb/2011/358412/>. Citado 3 vezes nas páginas 53, 56 e 62.

EDGAR, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, v. 5, p. 113, 2004. ISSN 1471-2105. Citado 2 vezes nas páginas 57 e 59.

ERSON-BENSAN, A. E. Introduction to microRNAs in biological systems. In: *miRNomics: MicroRNA Biology and Computational Analysis*. [S.l.]: Springer, 2014. p. 1–14. Citado 2 vezes nas páginas 24 e 37.

FEDOROFF, N. V. Transposable Elements, Epigenetics, and Genome Evolution. *Science*, v. 338, n. November, 2012. Citado 2 vezes nas páginas 28 e 51.

FESCHOTTE, C. Transposable elements and the evolution of regulatory networks. *Nature reviews. Genetics*, v. 9, n. 5, p. 397–405, 2008. ISSN 1471-0064. Disponível em: <http://dx.doi.org/10.1038/nrg2337>. Citado na página 66.

FESCHOTTE, C.; JIANG, N.; WESSLER, S. R. Plant transposable elements: where genetics meets genomics. *Nature reviews. Genetics*, v. 3, n. 5, p. 329–341, 2002. ISSN 14710056. Citado 3 vezes nas páginas 18, 51 e 52.

FESCHOTTE, C.; PRITHAM, E. J. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics*, v. 41, n. 1, p. 331–368, 2007. ISSN 0066-4197. Disponível em: <http://www.annualreviews.org/doi/abs/10.1146/annurev.genet.40.110405.090448>. Citado 3 vezes nas páginas 18, 19 e 51.

FINN, R. D. et al. The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, v. 44, n. December 2015, p. gkv1344, 2015. ISSN 1362-4962. Disponível em: <http://nar.oxfordjournals.org/content/early/2015/12/15/nar.gkv1344.full>. Citado na página 66.

FINNEGAN, D. J. Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, v. 5, n. 4, p. 103–107, 1989. ISSN 01689525. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/2543105>. Citado na página 13.

FLUTRE, T. REPET : pipelines for the identification and annotation of transposable elements in genomic sequences. *Context*, 2009. Citado na página 23.

FLUTRE, T. et al. Considering transposable element diversification in de novo annotation approaches. *PLoS ONE*, v. 6, n. 1, 2011. ISSN 19326203. Citado na página 23.

FREITAS, A. A.; WIESER, D. C.; APWEILER, R. On the Importance of Comprehensible Classification Models for Protein Function Prediction. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, v. 7, n. 1, p. 172–182, 2010. ISSN 1557-9964. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20150679>. Citado na página 22.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, v. 5, n. 10, p. R80, 2004. ISSN 1465-6914. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=545600{&}tool=pmcentrez{&}rendertype=ab>. Citado na página 59.

GIM, J.-A. et al. Genome-Wide Identification and Classification of MicroRNAs Derived from Repetitive Elements. *Genomics and Informatics*, v. 12, n. 4, p. 261–267, 2014. Citado 2 vezes nas páginas 13 e 37.

GONZÁLEZ, J.; PETROV, D. Genetics. MITEs–the ultimate parasites. *Science (New York, N.Y.)*, v. 325, n. 5946, p. 1352–1353, sep 2009. ISSN 1095-9203. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/19745141>. Citado na página 51.

GOODSTEIN, D. M. et al. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, v. 40, n. D1, p. 1178–1186, 2012. ISSN 03051048. Citado na página 32.

GORDON, A.; HANNON, G. J. Fastx-toolkit. FASTQ/A short-reads pre-processing tools. *Unpublished http://hannonlab. cshl. edu/fastx_ toolkit*, 2010. Citado na página 58.

GUYOT, R. et al. Microcollinearity in an ethylene receptor coding gene region of the Coffea canephora genome is extensively conserved with Vitis vinifera and other distant dicotyledonous sequenced genomes. *BMC plant biology*, v. 9, p. 22–37, 2009. ISSN 1471-2229. Citado 3 vezes nas páginas 51, 60 e 62.

HADJIARGYROU, M.; DELIHAS, N. The Intertwining of Transposable Elements and Non-Coding RNAs. *International journal of molecular sciences*, v. 14, n. 7, p. 13307–13328, 2013. ISSN 1422-0067. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3742188/>. Citado 2 vezes nas páginas 13 e 37.

HAN, Y.; WESSLER, S. R. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, v. 38, n. 22, 2010. ISSN 03051048. Citado 2 vezes nas páginas 54 e 55.

JANICKI, M.; ROOKE, R.; YANG, G. Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. *Chromosome Research*, v. 19, n. 6, p. 787–808, aug 2011. ISSN 1573-6849. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21850457>. Citado na página 21.

JONES, R. N. McClintock's controlling elements: the full story. *Cytogenetic and genome research*, v. 109, n. 1-3, p. 90–103, jan 2005. ISSN 1424-859X. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15753564>. Citado na página 13.

JONES-RHOADES, M. W.; BARTEL, D. P.; BARTEL, B. MicroRNAS and their regulatory roles in plants. *Annual review of plant biology*, v. 57, p. 19–53, 2006. ISSN 1543-5008. Citado na página 26.

JURKA, J. et al. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, v. 110, n. 1-4, p. 462–467, 2005. ISSN 14248581. Citado 5 vezes nas páginas 21, 32, 39, 57 e 60.

KOZOMARA, A.; GRIFFITHS-JONES, S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, v. 42, n. D1, p. 1–6, 2014. ISSN 03051048. Citado 2 vezes nas páginas 32 e 39.

KREBS, J. E.; GOLDSTEIN, E. S.; KILPATRICK, S. T. *Lewin's genes XI*. [S.l.]: Jones & Bartlett Learning, 2014. Citado na página 24.

KUANG, H. et al. Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: New functional implications for MITEs. *Genome Research*, v. 19, n. 1, p. 42–56, 2009. ISSN 10889051. Citado 5 vezes nas páginas 19, 29, 44, 53 e 62.

KUMAR, A.; BENNETZEN, J. L. Plant retrotransposons. *Annual review of genetics*, v. 33, p. 479–532, 1999. ISSN 0066-4197. Disponível em: <http://www.annualreviews.org/doi/abs/10.1146/annurev.genet.33.1.479>. Citado na página 14.

KURTOGLU, K. Y.; KANTAR, M.; BUDAK, H. New wheat microRNA using whole-genome sequence. *Functional and Integrative Genomics*, v. 14, n. 2, p. 363–379, 2014. ISSN 14387948. Citado na página 37.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, v. 9, n. 4, p. 357–359, 2012. ISSN 1548-7105. Citado na página 58.

LEVY, A.; SELA, N.; AST, G. TranspoGene and microTranspoGene: Transposed elements influence on the transcriptome of seven vertebrates and invertebrates. *Nucleic Acids Research*, v. 36, n. SUPPL. 1, p. 47–52, 2008. ISSN 03051048. Citado 3 vezes nas páginas 30, 32 e 38.

LI, Y. et al. Domestication of transposable elements into microrna genes in plants. *PLoS ONE*, v. 6, n. 5, p. e19212, 2011. ISSN 19326203. Citado 4 vezes nas páginas 13, 29, 37 e 44.

LISCH, D. How important are transposons for plant evolution? *Nature reviews. Genetics*, Nature Publishing Group, v. 14, n. 1, p. 49–61, jan 2013. ISSN 1471-0064. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23247435>. Citado 9 vezes nas páginas 13, 15, 18, 19, 20, 28, 37, 51 e 54.

LIU, R.; ZHU, J.-K. Non-coding RNAs as potent tools for crop improvement. *National Science Review*, p. 1–4, 2014. ISSN 2095-5138. Disponível em: <http://nsr.oxfordjournals.org/cgi/doi/10.1093/nsr/nwu006>. Citado na página 26.

LOPES, F. R. et al. Transposable elements in Coffea (Gentianales: Rubiacea) transcripts and their role in the origin of protein diversity in flowering plants. *Molecular genetics and genomics : MGG*, v. 279, n. 4, p. 385–401, apr 2008. ISSN 1617-4615. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/18231813>. Citado 2 vezes nas páginas 56 e 66.

LOPES, F. R. et al. Transcriptional activity, chromosomal distribution and expression effects of transposable elements in Coffea genomes. *PLoS ONE*, v. 8, n. 11, 2013. ISSN 19326203. Citado na página 56.

LORENZ, R. et al. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, v. 6, n. 1, p. 26, 2011. ISSN 1748-7188. Citado na página 47.

LOSS-MORAIS, G. et al. Identification of novel and conserved microRNAs in Coffea canephora and Coffea arabica. *Genetics and molecular biology*, Sociedade Brasileira de Genética, v. 37, n. 4, p. 671–682, oct 2014. ISSN 1415-4757. Disponível em: <http://www.scielo.br/scielo.php?script=sci{_}arttext{&}pid=S1415-47572014000>. Citado 2 vezes nas páginas 56 e 58.

LU, C. et al. Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in Oryza sativa. *Molecular Biology and Evolution*, v. 29, n. 3, p. 1005–1017, 2012. ISSN 07374038. Citado 4 vezes nas páginas 53, 54, 55 e 62.

MARX, V. Biology: The big challenges of big data. *Nature*, v. 498, n. 7453, p. 255–260, 2013. ISSN 0028-0836. Citado na página 32.

MEYERS, B. C. et al. Criteria for annotation of plant MicroRNAs. *The Plant cell*, v. 20, n. 12, p. 3186–3190, 2008. ISSN 1040-4651. Citado 2 vezes nas páginas 30 e 44.

MIROUZE, M.; VITTE, C. Transposable elements, a treasure trove to decipher epigenetic variation: insights from Arabidopsis and crop epigenomes. *Journal of Experimental Botany*, v. 65, n. 10, p. 2801–2812, 2014. ISSN 14602431. Citado 2 vezes nas páginas 20 e 22.

MONDEGO, J. M. et al. An EST-based analysis identifies new genes and reveals distinctive gene expression features of Coffea arabica and Coffea canephora. *BMC plant biology*, BioMed Central Ltd, v. 11, n. 1, p. 30, jan 2011. ISSN 1471-2229. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3045888{&}tool=pmcentrez{&}rendertype=ab>. Citado 2 vezes nas páginas 60 e 66.

NOVAK, P. et al. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, v. 29, n. 6, p. 792–793, 2013. ISSN 1367-4803. Disponível em: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btt054>. Citado na página 23.

OU-YANG, F. et al. Transposable element-associated microRNA hairpins produce 21-nt sRNAs integrated into typical microRNA pathways in rice. *Functional and Integrative Genomics*, v. 13, n. 2, p. 207–216, 2013. ISSN 1438793X. Citado 2 vezes nas páginas 29 e 37.

PARISOD, C. et al. Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytologist*, v. 186, n. 1, p. 37–45, 2010. ISSN 0028646X. Citado 2 vezes nas páginas 13 e 20.

PASCHOAL, A. R. et al. Non-coding transcription characterization and annotation: A guide and web resource for non-coding RNA databases. *RNA Biology*, v. 9, n. 3, p. 274–282, 2012. ISSN 1547-6286. Citado na página 32.

PIERCE, B. A. *Genetics: A conceptual approach*. [S.l.]: Macmillan, 2010. Citado 2 vezes nas páginas 16 e 17.

PIRIYAPONGSA, J.; JORDAN, I. K. Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA (New York, N.Y.)*, v. 14, n. 5, p. 814–821, 2008. ISSN 1355-8382. Citado 6 vezes nas páginas 13, 29, 37, 44, 53 e 54.

PROOST, S. et al. PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research*, v. 43, n. D1, p. D974–D981, 2015. ISSN 0305-1048. Disponível em: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku986>. Citado na página 32.

QUAST, C. et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, v. 41, n. D1, p. 590–596, 2013. ISSN 03051048. Citado na página 89.

QUINLAN, A. R. BEDTools: The Swiss-Army tool for genome feature analysis. *Current Protocols in Bioinformatics*, v. 2014, p. 11.12.1–11.12.34, 2014. ISSN 1934340X. Citado na página 60.

QUINLAN, A. R.; HALL, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, v. 26, n. 6, p. 841–842, 2010. ISSN 13674803. Citado 3 vezes nas páginas 40, 57 e 58.

RAGUPATHY, R.; YOU, F. M.; CLOUTIER, S. Arguments for standardizing transposable element annotation in plant genomes. *Trends in Plant Science*, Elsevier Ltd, v. 18, n. 7, p. 367–376, 2013. ISSN 13601385. Disponível em: <http://dx.doi.org/10.1016/j.tplants.2013.03.005>. Citado 2 vezes nas páginas 14 e 37.

REBIJITH, K. B. et al. In silico mining of novel microRNAs from coffee (Coffea arabica) using expressed sequence tags. *Journal of Horticultural Science & Biotechnology*, v. 88, p. 325–337, 2013. Citado na página 56.

RICE, P.; LONGDEN, I.; BLEASBY, A. EMBOSS: the European molecular biology open software suite. *Trends in genetics*, v. 16, n. 6, p. 276–277, 2000. ISSN 13590294. Citado na página 40.

ROBERTS, J. T.; CARDIN, S. E.; BORCHERT, G. M. Burgeoning evidence indicates that microRNAs were initially formed from transposable element sequences. *Mobile genetic elements*, v. 4, n. December, p. e29255, 2014. ISSN 2159-2543. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4091103/>. Citado 3 vezes nas páginas 29, 31 e 37.

ROBERTS, J. T. et al. Continuing analysis of microRNA origins: Formation from transposable element insertions and noncoding RNA mutations. *Mobile genetic elements*, v. 3, n. 6, p. e27755, 2013. ISSN 2159-2543. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3891635/>. Citado 2 vezes nas páginas 29 e 37.

RUTHERFORD, K. et al. Artemis: sequence visualization and annotation. *Bioinformatics (Oxford, England)*, v. 16, n. 10, p. 944–945, 2000. ISSN 1367-4803. Citado na página 39.

SAMPATH, P. et al. Genome-wide comparative analysis of 20 miniature inverted-repeat transposable element families in Brassica rapa and B. oleracea. *PloS one*, Public Library of Science, v. 9, n. 4, p. e94499, jan 2014. ISSN 1932-6203. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0094499>. Citado na página 53.

SCHULMAN, A. H. Retrotransposon replication in plants. *Current Opinion in Virology*, Elsevier B.V., v. 3, n. 6, p. 604–614, 2013. ISSN 18796257. Disponível em: <http://dx.doi.org/10.1016/j.coviro.2013.08.009>. Citado 2 vezes nas páginas 13 e 15.

SMIT, A. F. A.; HUBLEY, R.; GREEN, P. *RepeatMasker Open-3.0*. 1996. Citado na página 21.

SONNHAMMER, E. L.; DURBIN, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, v. 167, n. 1-2, p. GC1–10, 1995. ISSN 0378-1119. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/8566757>. Citado 2 vezes nas páginas 57 e 59.

SOUZA, F. d. F. et al. *Características das principais variedades de café cultivadas em Rondônia*. Porto Velho: Embrapa Rondônia, 2004. ISSN 0103-9865. Citado na página 56.

STAMATAKIS, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, v. 30, n. 9, p. 1312–1313, 2014. ISSN 14602059. Citado na página 59.

SUN, J. et al. Characterization and evolution of microrna genes derived from repetitive elements and duplication events in plants. *PLoS ONE*, v. 7, n. 4, p. e34092, 2012. ISSN 19326203. Citado 4 vezes nas páginas 29, 37, 40 e 45.

SZCZEŚNIAK, M. W.; MAKAŁOWSKA, I. MiRNEST 2.0: A database of plant and animal microRNAs. *Nucleic Acids Research*, v. 42, n. D1, p. 74–77, 2014. ISSN 03051048. Citado na página 40.

TEAM, R. RStudio: Integrated Development for R. *RStudio, Inc., Boston, MA. URL http://www. RStudio. com/ide*, 2014. Citado 3 vezes nas páginas 58, 59 e 63.

TEMPEL, S.; POLLET, N.; TAHI, F. ncRNAclassifier: a tool for detection and classification of transposable element sequences in RNA hairpins. *BMC Bioinformatics*, v. 13, n. 1, p. 246, 2012. ISSN 1471-2105. Citado na página 38.

TEMPEL, S.; TAHI, F. A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Research*, v. 40, n. 11, p. 1–9, 2012. ISSN 03051048. Citado na página 30.

VOINNET, O. Origin, Biogenesis, and Activity of Plant MicroRNAs. *Cell*, Elsevier Inc., v. 136, n. 4, p. 669–687, 2009. ISSN 00928674. Disponível em: <http://dx.doi.org/10.1016/j.cell.2009.01.046>. Citado 4 vezes nas páginas 24, 25, 26 e 27.

WATERHOUSE, A. M. et al. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, v. 25, n. 9, p. 1189–1191, 2009. ISSN 1367-4803. Disponível em: <http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp033>. Citado 2 vezes nas páginas 57 e 59.

WEI, L. et al. Dicer-like 3 produces transposable element-associated 24-nt siRNAs that control agricultural traits in rice. *Proceedings of the National Academy of Sciences of the United States of America*, v. 111, n. 10, p. 3877–3882, 2014. ISSN 1091-6490. Disponível em: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3956178{&}tool=pmcentrez{&}rendertype=ab>. Citado na página 62.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics*, v. 8, n. 12, p. 973–982, 2007. ISSN 1471-0056. Citado 12 vezes nas páginas 13, 14, 15, 16, 17, 18, 39, 42, 52, 57, 59 e 60.

XIE, M.; YU, B. siRNA-directed DNA Methylation in Plants. *Current genomics*, v. 16, n. 1, p. 23–31, 2015. ISSN 1389-2029 (Print). Citado 2 vezes nas páginas 27 e 53.

YAN, Y. et al. Small RNAs from MITE-derived stem-loop precursors regulate abscisic acid signaling and abiotic stress responses in rice. *Plant Journal*, v. 65, n. 5, p. 820–828, 2011. ISSN 09607412. Citado na página 53.

YANG, G. MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. *BMC bioinformatics*, v. 14, n. 1, p. 186, jan 2013. ISSN 1471-2105. Disponível em: <http://www.biomedcentral.com/1471-2105/14/186>. Citado 3 vezes nas páginas 54, 55 e 57.

YANG, G. et al. Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science (New York, N.Y.)*, v. 325, n. 5946, p. 1391–1394, 2009. ISSN 0036-8075. Citado 2 vezes nas páginas 18 e 51.

YE, C.; JI, G.; LIANG, C. detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes. *Scientific Reports*, Nature Publishing Group, v. 6, n. January, 2016. ISSN 2045-2322. Disponível em: <http://www.nature.com/articles/srep19688>. Citado 2 vezes nas páginas 54 e 55.

YUYAMA, P. M. et al. FISH using a gag-like fragment probe reveals a common Ty3-gypsy-like retrotransposon in genome of Coffea species. *Genome / National Research Council Canada = Génome / Conseil national de recherches Canada*, v. 55, n. 12, p. 825–833, 2012. ISSN 1480-3321. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23231601>. Citado na página 56.

ZHANG, X. et al. Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell*, v. 126, n. 6, p. 1189–1201, 2006. ISSN 00928674. Citado na página 28.

ZHANG, Y.; JIANG, W. K.; GAO, L. Z. Evolution of microRNA genes in Oryza sativa and Arabidopsis thaliana: An update of the inverted duplication model. *PLoS ONE*, v. 6, n. 12, p. e28073, 2011. ISSN 19326203. Citado 3 vezes nas páginas 29, 37 e 44.

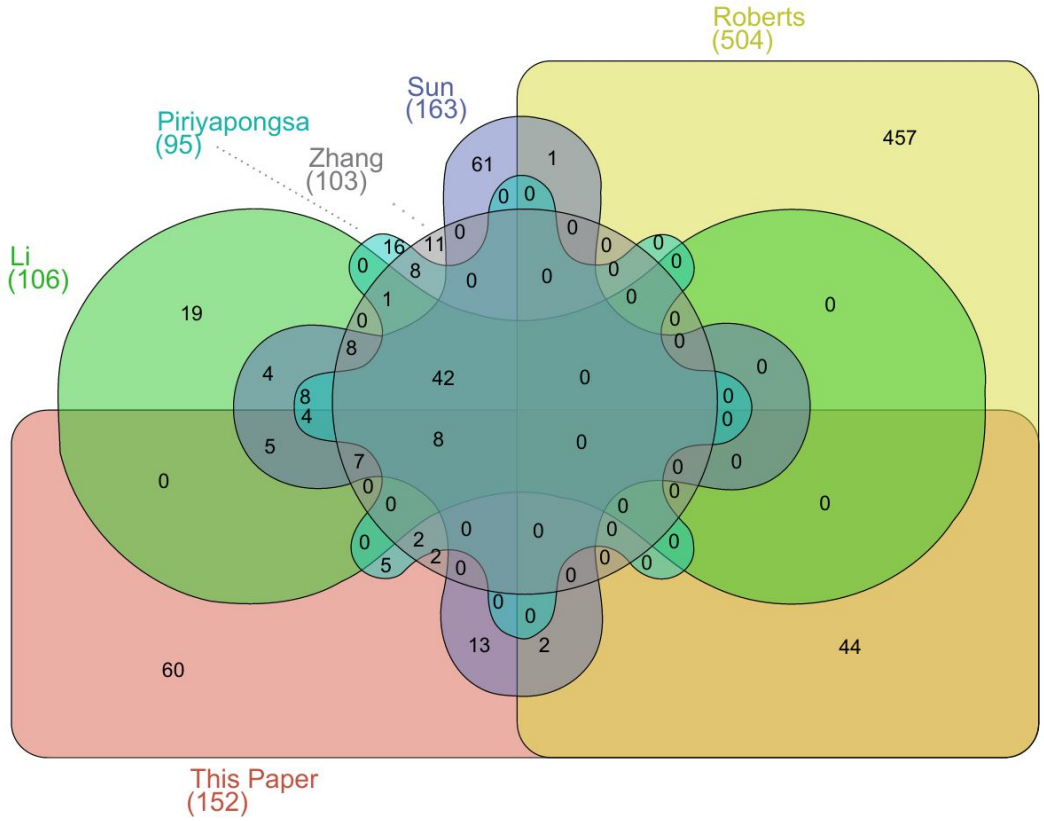
ZHAO, Y.; CHEN, X. Non-coding RNAs and DNA methylation in plants. *National Science Review*, v. 1, n. 2, p. 219–229, 2014. ISSN 2095-5138. Disponível em: <http://nsr.oxfordjournals.org/cgi/doi/10.1093/nsr/nwu003>. Citado 2 vezes nas páginas 27 e 28.

ZHOU, M. et al. Genome-wide analysis of clustering patterns and flanking characteristics for plant microRNA genes. *FEBS Journal*, v. 278, n. 6, p. 929–940, 2011. ISSN 1742464X. Citado na página 40.

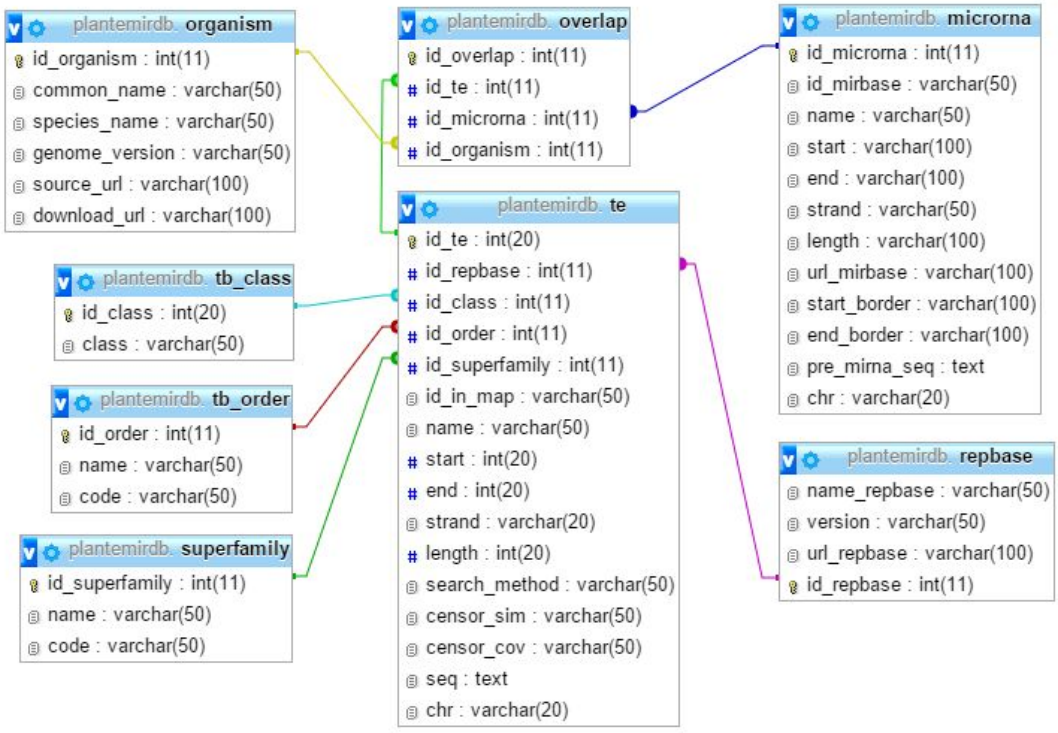
Apêndices

APÊNDICE A – *Electronic Supplementary Material*

ELECTRONIC SUPPLEMENTARY INFORMATION



Supplementary Fig. 1. Venn diagram showing the quantity of TE-related pre-miRNAs found by other authors compared to our study (Piriyapongsa; Jordan, 2008; Li et al., 2011; Zhang; Jiang; Gao, 2011; Sun et al., 2012; Roberts et al., 2013). Only miRBase reference data from Roberts et al. (2013) were used to compare results. Venn diagram was plotted using InteractiVenn (Heberle et al., 2015).



Supplementary Fig. 2. Entity-Relationship Diagram (ERD) of PlanTE-MIR DB.

Species Name	Common Name	Version	URL
Arabidopsis thaliana	thale cress	TAIR v10.0	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Athaliana/assembly/Athaliana_167.fa.gz
Arabidopsis lyrata	lyre-leaved rock cress	JGI v1.0	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000004255.1_v.1.0/GCA_000004255.1_v.1.0_genomic.fna.gz
Brachypodium distachyon	purple false brome	JGI v1.0	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Bdistachyon/assembly/Bdistachyon_192.fa.gz
Glycine max	soybean	Glyma v1.0	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Gmax/assembly/Gmax_189.fa.gz
Lotus japonicus	miyakogusa	Lj v2.5	ftp://ftp.kazusa.or.jp/pub/lotus/lotus_r2.5/pseudomolecule/Lj2.5_pseudomol.fna.gz
Malus domestica	apple	maldom pseudo v1.0	http://www.rosaceae.org/system/files/apple_genome/Malus_x_domestica.v1.0-primary.pseudo.fa.gz
Medicago truncatula	barrel medic	MedtrA17 v3.5	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Mtruncatula/assembly/Mtruncatula_198.fa.gz
Oryza sativa	rice	MSU v7.0	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Osativa/assembly/Osativa_204.fa.gz
Physcomitrella patens	moss	JGI v1.1	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Ppatens/assembly/Ppatens_152.fa.gz
Populus trichocarpa	poplar	JGI Poptr2.0	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000002775.2_Poptr2_0/GCA_000002775.2_Poptr2_0_genomic.fna.gz
Prunus persica	peach	JGI v1.0	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000346465.1_Prupe1_0/GCA_000346465.1_Prupe1_0_genomic.fna.gz
Solanum lycopersicum	tomato	SL v2.40	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Slycopersicum/assembly/Slycopersicum_225.fa.gz
Solanum tuberosum	irish potato	SolTub v3.0	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA_000226075.1_SolTub_3.0/GCA_000226075.1_SolTub_3.0_genomic.fna.gz
Sorghum bicolor	sorghum	JGI Sb v1.0	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Sbicolor/assembly/Sbicolor_79.fa.gz
Vitis vinifera	grape	genoscope march 2010	ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Vvinifera/assembly/Vvinifera_145.fa.gz

Supplementary Table 1. Fifteen species assessed by our analysis. Versions were retrieved from genome assembly source repositories.

Table format description

Alongside with detailed information showed on the website, data can be downloaded as a tab-separated table format. It was made this way for easy parsing of entries. Each one of twenty seven columns refers to following attributes:

1. Species Name: Species binomial nomenclature.
2. Common Name: Organism's popular name.
3. Assembly Version: Genome assembly version (see list in Supplementary Table 1).
4. Assembly URL: Download genome assemblies available at source website.
5. TE Name: Names were given to each annotated TE which had intersection with any pre-miRNA. Nomenclature rules and classification system were used according to Wicker and co-workers (2007).
6. TE Class: Higher level of most recent classification system for TEs. Based on Repbase.
7. TE Order: Another classification level (may be absent). Based on Repbase.
8. TE Superfamily: Lower level of classification system. Based on Repbase.
9. Repbase Name: Repbase's TE consensus name. Used to identify TE consensus in Repbase.
10. Repbase Consensus URL: Direct link to access TE consensus notes in Repbase.
11. Repbase Version: Version of Repbase's library used for CENSOR's annotation.
12. Search Method: Search method employed by CENSOR to find repetitions in genome assemblies.
13. Coverage (CENSOR): Fraction of TE consensus aligned to genome assemblies (values between 80% and 100%).
14. Similarity (CENSOR): Similarity between TE consensus and genome assemblies (values between 0.8 and 1.0). More information is available at CENSOR's help webpage: http://www.girinst.org/censor/help.html.
15. TE Chromosome or scaffold: Accession in genome assembly FASTA files (e.g. chr1).
16. TE Start Position: One-based start coordinate of annotated TE.
17. TE End Position: One-based end coordinate of annotated TE.
18. TE Strand: "+" (plus or sense) and "-" (minus or antisense).
19. Overlapping pre-miRNA: pre-miRNA name available at miRBase.
20. pre-miRNA ID: pre-miRNA accession available at miRBase.
21. pre-miRNA URL: pre-miRNA information URL at miRBase.
22. pre-miRNA Chromosome or scaffold: Accession in genome assembly FASTA files (e.g. chr1).
23. pre-miRNA Start position: One-based start coordinate of annotated TE.
24. pre-miRNA End position: One-based end coordinate of annotated TE.
25. pre-miRNA Strand: "+" (plus or sense) and "-" (minus or antisense).
26. pre-miRNA Sequence: pre-miRNA complete sequence (RNA sequence).
27. TE Sequence: annotated TE complete sequence (DNA sequence).

Generic Feature Format Version 3 (GFF3) description

GFF3 files are well known by scientific community and are described at http://www.sequenceontology.org/gff3.shtml. For better understanding, tab-separated column content are described below (bold text and quotes emphasize the attributes):

TE GFF3:

Column #1: TE Chromosome or scaffold
Column #2: Search Method
Column #3: Key (transposable_element in this case)
Column #4: TE Start Position
Column #5: TE End Position
Column #6: Score (in this case "." means not available)
Column #7: TE Strand
Column #8: Phase (in this case "." means not available)
Column #9: Features (comprises following additional information)
ID="**Repbase Name**";Name="**TE Name**";Alias="**Class**":"**Order**":"**Superfamily**";Note="**Similarity (CENSOR)**":"**Coverage (CENSOR)**":"**Repbase Version**"

E.g.:
ID=**ATMU3N1**;Name=**DTx_ATMU3N1_Chr1-1**;Alias=**Class II (DNA transposons) - Subclass I**:**TIR**:**Undefined**;Note=**0.9529**:**100.00**:**Repbase19.04**

pre-miRNA GFF3:

Column #1: pre-miRNA Chromosome or scaffold
Column #2: Source (miRBasev21 in this case)
Column #3: Key (ncRNA in this case)
Column #4: pre-miRNA Start Position
Column #5: pre-miRNA End Position
Column #6: Score (in this case "." means not available)
Column #7: pre-miRNA Strand
Column #8: Phase (in this case "." means not available)
Column #9: Features (comprises following additional information)
ID="**pre-miRNA ID**";Name="**pre-miRNA Name**"

E.g.:
ID=**MI0019229**;Name=**ath-MIR5635b**

FASTA header description

Respecting the presented titles, FASTA files for TEs contain the following header organization:
>"**TE Name**" "**Repbase Name**" intersects "**pre-miRNA Name**" ["**Species Name**"]

E.g.:
>**RLG_ATHILA4C_LTR_Chr2-1 ATHILA4C_LTR** intersects **ath-MIR8175** [**Arabidopsis thaliana**]

Using the same rules, the headers for pre-miRNAs follow this sample:
>"**pre-miRNA Name**" "**pre-miRNA ID**" intersects "**TE Name**" ["**Species Name**"]

E.g.:
>**ath-MIR8175 MI0026805** intersects **RLG_ATHILA4C_LTR_Chr2-1** [**Arabidopsis thaliana**]

Data visualization

We suggest using Artemis software for data visualization. Entire datasets and assemblies can be downloaded at Downloads section and loaded in Artemis. All files are ready to use, except for *Populus trichocarpa* assembly, which may need corrections in the FASTA file headers in order to match GFF3 accessions. The scaffold names must be "scaffold_" followed by their respective number (e. g. scaffold_1).

References

HEBERLE, H. et al. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. **BMC bioinformatics**, v. 16, n. 1, p. 169, 2015.

LI, Y. et al. Domestication of transposable elements into microRNA genes in plants. **Plos one**, v. 6, n. 5, p. e19212, 2011.

PIRIYAPONGSA, J.; JORDAN, I. K. Dual coding of siRNAs and miRNAs by plant transposable elements. **Rna**, v. 14, n. 5, p. 814-821, 2008.

ROBERTS, Justin T. et al. Continuing analysis of microRNA origins: Formation from transposable element insertions and noncoding RNA mutations. **Mobile genetic elements**, v. 3, n. 6, p. e27755, 2013.

SUN, J. et al. Characterization and evolution of microRNA genes derived from repetitive elements and duplication events in plants. **PloS one**, v. 7, n. 4, p. e34092, 2012.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973-982, 2007.

ZHANG, Y.; JIANG, W.; GAO, L. Evolution of microRNA genes in Oryza sativa and Arabidopsis thaliana: an update of the inverted duplication model. **PloS one**, v. 6, n. 12, p. e28073, 2011.

APÊNDICE B – Sequências utilizadas para limpeza dos *reads*

B.1 Adaptadores

```
>TruSeq2_PE_f
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
>TruSeq2_PE_r
AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
>3primeLinker
CTGTAGGCACCATCAAT
```

B.2 Sequências diversas

- Cloroplasto de *Coffea arabica* (NCBI *accession* NC_008535);
- Mitocôndria de *Erythranthe guttata* (NCBI *accession* NC_018041);
- Mitocôndria de *Nicotiana tabacum* (NCBI *accession* NC_006581);
- Mitocôndria de *Vitis vinifera* (NCBI *accession* NC_012119);
- rRNAs de *Coffea canephora*, subunidades maior e menor (SILVA rRNA *database*) (QUAST et al., 2013);
- tRNAs de *Solanum tuberosum* (PlantRNA *database*) (COGNAT et al., 2013);
- tRNAs de *Vitis vinifera* (GtRNAdb) (CHAN; LOWE, 2009);
- snoRNAs de diversas espécies (Plant snoRNA *database*) (BROWN et al., 2003).

APÊNDICE C – Artigo publicado no periódico *Functional & Integrative Genomics*

Funct Integr Genomics
DOI 10.1007/s10142-016-0480-5

ORIGINAL ARTICLE

PlanTE-MIR DB: a database for transposable element-related microRNAs in plant genomes

Alan P. R. Lorenzetti[1] · Gabriel Y. A. de Antonio[2] · Alexandre R. Paschoal[2] · Douglas S. Domingues[3]

Received: 9 November 2015 / Revised: 14 January 2016 / Accepted: 19 January 2016

Abstract Transposable elements (TEs) comprise a major fraction of many plant genomes and are known to drive their organization and evolution. Several studies show that these repetitive elements have a prominent role in shaping noncoding regions of the genome such as microRNA (miRNA) loci, which are components of post-transcriptional regulation mechanisms. Although some studies have reported initial formation of miRNA loci from TE sequences, especially in model plants, the approaches that were used did not employ systems that would allow results to be delivered by a user-friendly database. In this study, we identified 152 precursor miRNAs overlapping TEs in 10 plant species. PlanTE-MIR DB was designed to assemble this data and deliver it to the scientific community interested in miRNA origin, evolution, and regulation pathways. Users can browse the database through a web interface and search for entries using various parameters. This resource is cross-referenced with repetitive element (Repbase Update) and miRNA (miRBase) repositories, where sequences can be checked for further analysis. All data in PlanTE-MIR DB are publicly available for download in several file formats to facilitate their understanding and use. The database is hosted at http://bioinfo-tool.cp.utfpr.edu.br/plantemirdb/.

Keywords Transposable elements · MicroRNAs · Plant genomes · Database

Introduction

Transposable elements (TEs) are present in almost all living organisms and comprise a significant fraction of most plant genomes (Baidouri and Panaud 2013; Ragupathy et al. 2013). They are known to drive important modifications in host genomes, including the inactivation, creation, and mobilization of genes, chromosomal rearrangement, gene expression modulation, and epigenetic silencing (Lisch 2013; Bennetzen and Wang 2014). TEs are also known to have an important role in shaping long noncoding RNAs (lncRNAs) and small ncRNAs (e.g., piwi-interacting RNAs [piRNAs], small interfering RNAs [siRNAs], and microRNAs [miRNAs]) (Hadjiargyrou and Delihas 2013; Piriyapongsa and Jordan 2008; Li et al. 2011; Gim et al. 2014).

Mature plant miRNAs are usually 21-nucleotide-long hairpin RNA-derived (hpRNA) sequences that can bind to target mRNAs through Watson-Crick base pairing on the 3′ UTR. This pairing results in mRNA destabilization or translational repression, which are effective mechanisms for gene regulation. In plants, RNA polymerase II enzymatic complexes are generally committed to transcribe miRNA loci. Inside the cell nucleus, cleavage of foldback transcribed structures through DICER-LIKE 1 (DCL1) is executed.

Electronic supplementary material The online version of this article (doi:10.1007/s10142-016-0480-5) contains supplementary material, which is available to authorized users.

✉ Douglas S. Domingues
doug@rc.unesp.br

Alan P. R. Lorenzetti
alan.lorenzetti@uel.br

1 Graduation Program in Genetics and Molecular Biology, Universidade Estadual de Londrina, UEL, Londrina, Brazil

2 Bioinformatics Laboratory, Universidade Tecnológica Federal do Paraná, UTFPR, Cornélio Procópio, Brazil

3 Department of Botany, Instituto de Biociências, Universidade Estadual Paulista, UNESP, Rio Claro, Brazil

Published online: 18 February 2016

This step promotes conversion of primary miRNAs (pri-miRNAs) into precursor miRNAs (pre-miRNAs), which is followed by further processing to transform them into miRNA/miRNA* duplex. Finally, one of the dissociated miRNAs is loaded into ARGONAUTE (AGO) proteins to assemble the functional RNA-induced silencing complex (RISC) (Axtell 2013; Erson-Bensan 2014).

Hairpin structures are supposed to arise either by inverted duplication of the target gene locus (Allen et al. 2004) or juxtaposed TEs (Roberts et al. 2013). Piriyapongsa and Jordan (2008) also describe a model in which folded expressed miniature inverted-repeat transposable elements (MITEs) may be processed by the miRNA biogenesis pathway. In another report, Ou-Yang et al. (2013) used AGO1 immunoprecipitation and DCL mutants to find three MITE-associated bona fide miRNAs depending on those proteins, pointing to their functional activity.

Most miRNA loci originate from intergenic genomic sequences, but there is considerable evidence that they were initially formed from TE sequences (Hadjiargyrou and Delihas 2013; Roberts et al. 2014; Budak and Akpinar 2015). The "domestication" of TEs to form miRNA genes has been demonstrated by high throughput sequencing in rice (Li et al. 2011; Barrera-Figueroa et al. 2012), and other plant species were checked for TE-MIRs at the genomic level (Zhang et al. 2011; Sun et al. 2012; Kurtoglu et al. 2014). Similar findings have been reported for the human genome and other metazoan genomes (Levy et al. 2008; Tempel et al. 2012), for which there are publicly available resources showing matched overlaps between TE and miRNA loci. However, the comparable plant data has not yet been compiled into a user-friendly database enabling search and retrieval of this information.

In this study, we present PlanTE-MIR DB, available at http://bioinfo-tool.cp.utfpr.edu.br/plantemirdb/, which provides a user-friendly database for investigation of overlaps between TE and pre-miRNA loci in ten plant genomes.

Material and methods

Pre-miRNA annotation and curation

Our analysis relied on the annotated pre-miRNAs from miRBase (version 21) (Kozomara and Griffiths-Jones 2014) within 15 genome assemblies. Genome assemblies were retrieved from several repositories based on reference versions indicated by miRbase (Table S1). However, due to divergences between miRBase annotation file accession names and assembly headers, a checking step was executed.

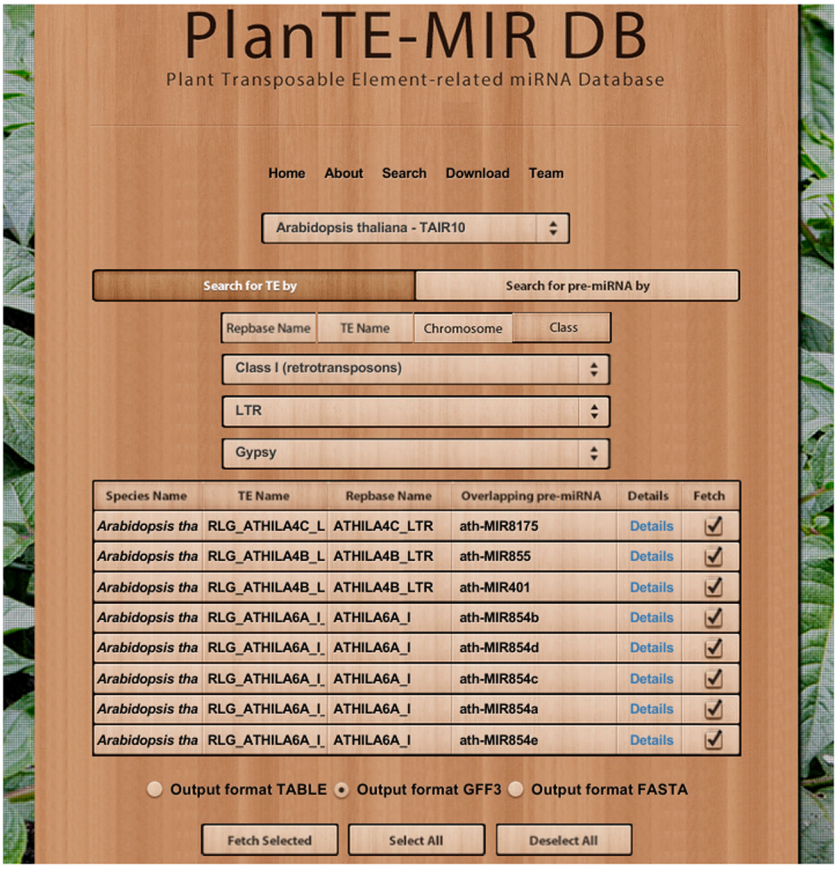


Fig. 1 Search section overview. Entries can be searched either by TE or pre-miRNA. Here, we present a search example for TE by Class I (retrotransposons), LTR order, and Gypsy superfamily

For that, miRBase pre-miRNA sequences in FASTA format were obtained for each of the studied plant species and BLAST (BLASTN, version 2.2.28+) (Camacho et al. 2009) searched against their respective genomes. An in-house bash script was made to run the program and perform the tasks. Only hits with 100 % query coverage and identity were maintained. In the case that some of the annotated pre-miRNAs from the same family showed indistinguishable sequences (e.g., mtr-MIR2669a, mtr-MIR2669b), they were aligned with more than one position. In these cases, we split single position hits from repeated ones and transformed them into GFF3 annotation files. Manual inspection was done using Artemis (version 15.1.1) (Rutherford et al. 2000) through the comparison of previously cited GFF3 files with an accession name corrected miRBase annotation file, all loaded on source genomes. New manually inspected GFF3 annotation files were created to match the accession names to source assembly headers.

Reference TEs

Plant TE libraries were obtained from Repbase Update (version 19.04 and 19.06 REPET edition) (Jurka et al. 2005). We used CENSOR (version 4.2.28 and 4.2.29) (Jurka et al. 2005) implemented with WU-BLAST (version 2.0 04-May-2006) as a search engine, using BLASTN and BLASTX algorithms according to well-established criteria (Wicker et al. 2007) to stringently remap reference TEs to genome assemblies. Initially, we used early versions of software and libraries with BLASTN, and then later versions with

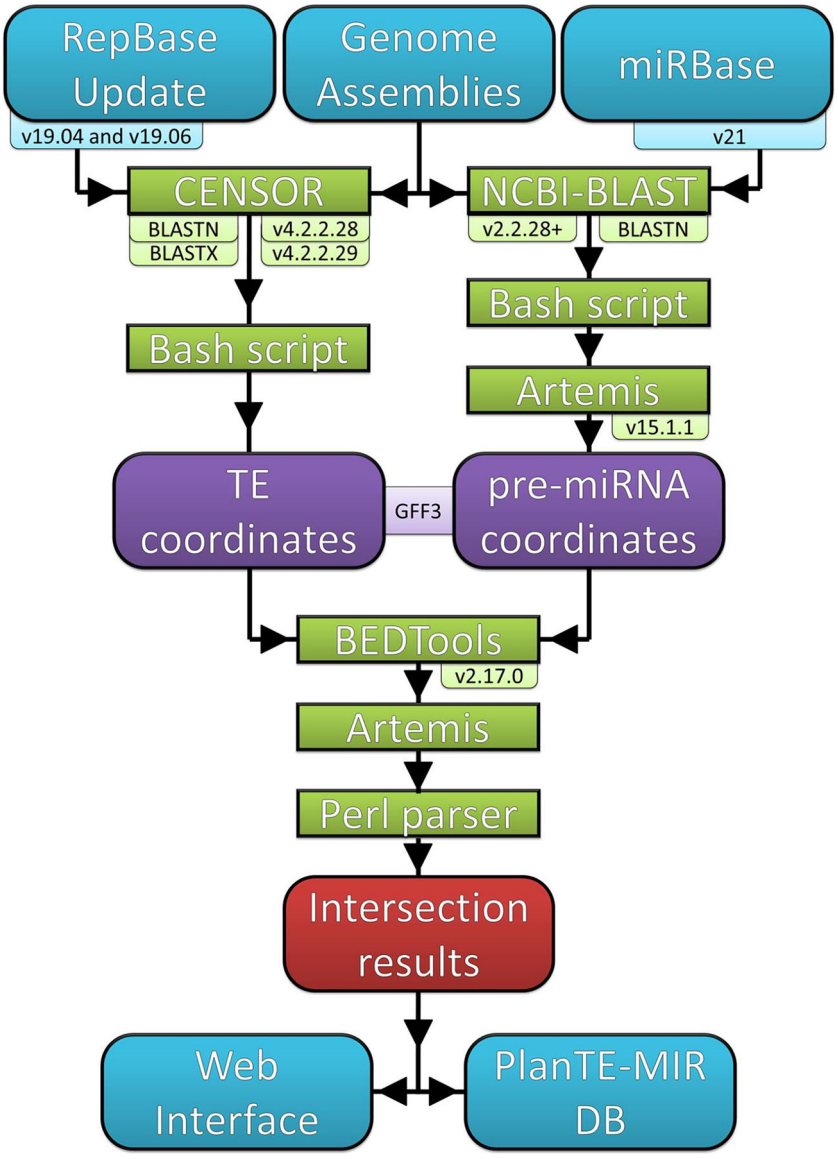


Fig. 2 Workflow diagram for the identification of TE-MIRs. CENSOR and BLAST programs were used to map TEs and pre-miRNAs. Bash script was used to filter and parse results to GFF3 file format. Using Artemis, pre-miRNA files were checked to confirm names and positions according to miRBase. Positional intersection analysis between TE-miRNA was run using BEDtools and manually checked with Artemis. These results were modelled to build PlanTE-MIR DB

BLASTX. A bash script was written to filter the results, according to the 80-80-80 rule proposed by Wicker and colleagues, and to parse TE coordinates to GFF3 annotation files.

TE-MIR relationship

We found positional overlaps between TEs and pre-miRNAs using the BEDTools (version 2.17.0) (Quinlan and Hall 2010) intersection function. Only pre-miRNAs having at least 36 % of their extension covered by a TE were maintained. Intersections were manually checked using Artemis. Whole sequences were captured from source assemblies through an in-house bash script running EMBOSS tools (version 6.6.0.0) (Rice et al. 2000).

Evolutionary conservation between TE-MIRs across taxa

We used a sequence-based similarity search method (Reciprocal Best BLAST Hit - RBH), following the rationale of Sun et al. (2012) to track evolutionary conservation. Thus, we BLAST (BLASTN, version 2.2.28+) searched our pre-miRNAs against all miRBase (version 21) hairpin sequences. Only hits with E values $\leq$ 1e-06 with at least 90 % query alignment and minimum of 80 % identity were maintained. We employed the same criteria used by Zhou et al. (2011) and Sun et al. (2012) to classify TE-MIRs at three levels: highly conserved (when TE-MIR homologs are present in both monocots and eudicots), low conserved (when TE-MIR homologs are present only in monocots or eudicots), and nonconserved (when a TE-MIR has no homologs outside a single species).

Transcriptional evidence for miRNAs

We checked for transcriptional evidence through browsing only high confidence miRNAs for each one of the plant species in the miRBase (version 21). Our pre-miRNAs were also used as queries and BLAST (BLASTN, version 2.2.28+) searched against the miRNEST (version 2.0) deep sequencing prediction file (Szcześniak and Makałowska 2014). Only hits presenting full identity and coverage of queries were maintained.

Database and web interface implementation

Annotation files were parsed to table using a Perl in-house script. Additional information was manually introduced using the Kingsoft Office Spreadsheet software. The data were then exported to a comma-separated values table and automatically inserted in MySQL Database Server (version 5.6) relational tables using a PHP script (version 5.3.10).

PlanTE-MIR DB was built on a 64-bit Windows (version 8.1) workstation. XAMPP (version 3.2.1) was executed to integrate Apache HTTP Server (version 2.2) with PHP and MySQL. The back-end was encoded in PHP and HTML5, using JavaScript jQuery library (version 1.11.2), with the plugins jQuery Vegas (version 1.3.5) and Ajax. For website design and structure customizing, we used Cascading Style Sheets (CSS3) as front-end. Except for the Windows operating system, only open source and cross-platform software

Table 1 Overall numbers of TEs, miRNAs, and TE-MIRs for the plant genomes analyzed in this study. Ten species presented at least one TE-MIR in their genome

Species	TEs[a]	TEs[b]	pre-miRNAs[c]	TE-MIRs[d]
Arabidopsis thaliana	6837	314	325	22
Brachypodium distachyon	3991	1402	317	2
Glycine max	18035	18290	573	4
Medicago truncatula	7481	1128	672	20
Oryza sativa	66794	2930	592	56
Physcomitrella patens	23744	3855	229	1
Populus trichocarpa	14587	2143	352	10
Sorghum bicolor	178106	29885	205	35
Solanum tuberosum	5530	14814	224	1
Vitis vinifera	16240	4787	163	1

[a]Annotated TEs using CENSOR with BLASTN

[b]Annotated TEs using CENSOR with BLASTX

[c]Pre-miRNAs retrieved from miRBase (version 21)

[d]MiRNAs intersecting at least one TE in each species

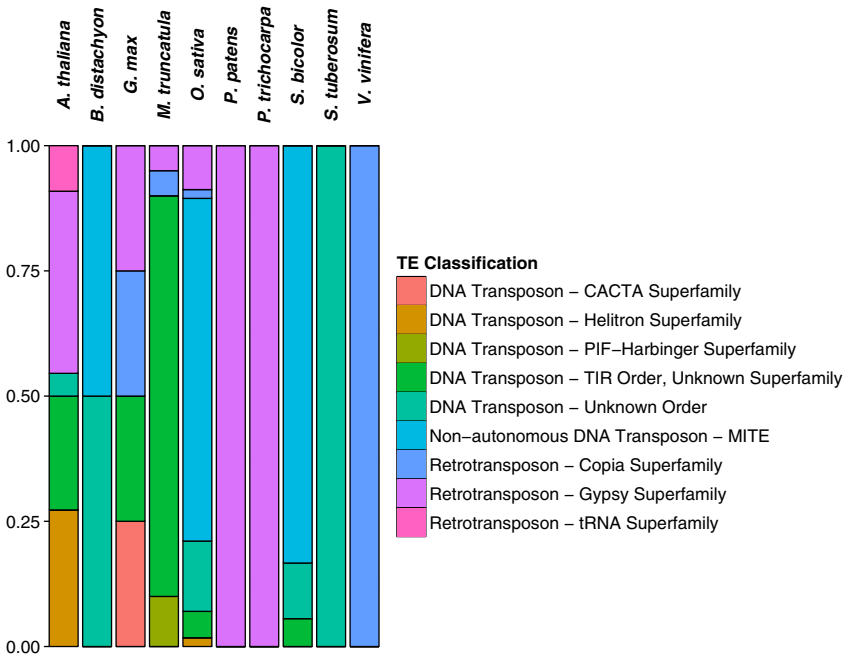


Fig. 3 Database composition by plant species and TE classification. MITEs are the most frequent type of repetition associated to miRNAs. DNA transposons were also highly related to this phenomenon. Classification data were collected from the Repbase Update

were used for database and web interface implementation. The complete system is hosted in the Information Office of the Federal University of Technology, Paraná, Brazil (UTFPR) and available at http://bioinfo-tool.cp.utfpr.edu.br/plantemirdb/.

Results and discussion

PlanTE-MIR DB: system and database overview

Plant Transposable Element-related miRNA Database (PlanTE-MIR DB, Fig. 1) was built as a resource for researchers interested in the evolution of TE and miRNA and their relationship. In this section, we detail the website and its functionalities.

The web interface was designed to be user-friendly, prioritizing easy ways of finding desired data through the use of filters, and providing alternative file formats when downloading entries. Accordingly, the website is divided in five sections: Home, About, Search, Download, and Team.

The Home and About sections concisely describe the purpose of the repository and its methods, and briefly instruct the user on how to interact with the search and download tools. They also contain information about assembly versions and reference libraries employed in the analysis.

In the Search section, users have a web interface for searching TE-MIR entries. The page was designed as an intuitive step-by-step form where the user can select options by name of the organism and TE or pre-miRNA attributes (Fig. 1). TEs can be found (1) by selecting the reference name (as supplied by the Repbase Update), (2) by TE name according to the nomenclature of Wicker and colleagues, (3) by TE position in the genome assembly, and (4) by TE class (Wicker et al. 2007). The last option is a hierarchical filter that allows the user to choose among TE classes, orders, and superfamilies. Similarly, pre-miRNAs may be found by miRBase ID, miRBase name, or position in the

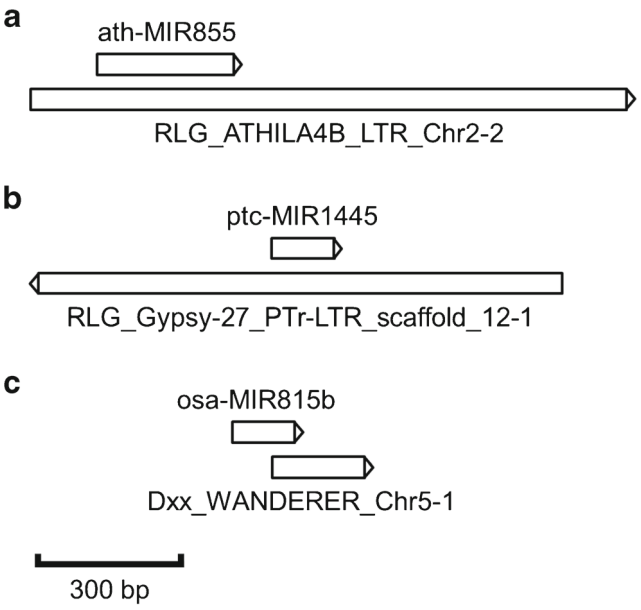


Fig. 4 Representative example of intersection patterns found by our analysis. Pre-miRNAs may intersect long terminal repeat (LTR) regions (**a**, **b**) or terminal regions of DNA transposons (**c**)

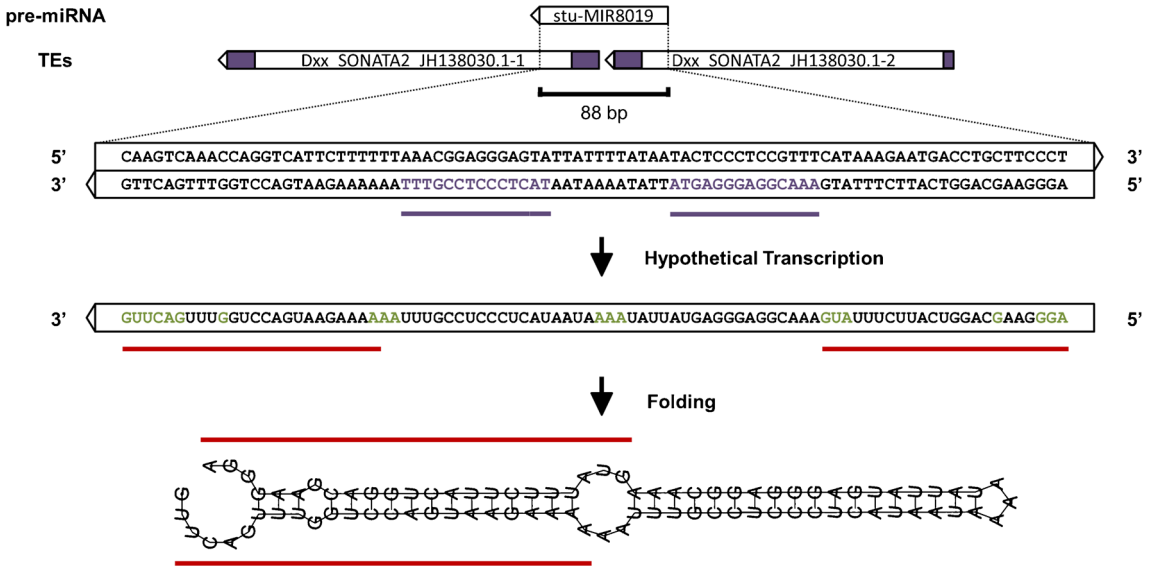


Fig. 5 Example of two juxtaposed non-autonomous DNA transposons overlapping a pre-miRNA in *Solanum tuberosum*. Terminal inverted repeats (TIRs) are highlighted in *purple*. Dxx_SONATA2_JH138030.1-1 is an intact element and Dxx_SONATA2_JH138030.1-2 lacks part of the 5′ TIR. A hairpin structure may emerge due to transcript complementarity. *Light green letters* show loop regions, and *red lines* emphasize mature miRNA regions. The secondary structure was plotted using RNAfold Webserver with the Minimum Free Energy (MFE) prediction method (Lorenz et al. 2011)

assembly. Next, a list of hits is shown to the user, allowing him or her to download search results by selecting Table file format, GFF3 file format, or FASTA file format. Furthermore, the user may access a detailed page containing information about the organism as well as annotations and cross-references obtained for each result. The description table shows species name, common name, assembly version, TE name, TE classification, Repbase Name, TE annotation details (such as Repbase version, CENSOR coverage, CENSOR similarity, start position, end position, and strand), overlapping pre-miRNA, pre-miRNA ID, and pre-miRNA annotation details. Further information relative to these items can be found in electronic supplementary material.

Bulk data for each species are provided through the Download section. All data are available in three formats: table, GFF3, and FASTA. GFF3 annotation files for TEs and pre-miRNAs can be directly loaded into publicly available assemblies using a genome browser tool (e.g., Artemis).

Identification of TE-MIRs

Since we found some inconsistencies between genome versions in miRBase and those in TE curated databases, we started our analyses by remapping reference TEs to standardized versions of plant genomes (Table S1). Figure 2 summarizes our approach to finding TE-MIR associations; in brief, we searched for positional intersections between pre-miRNAs and TEs. Among 10 species, we obtained a total of 152 pre-miRNAs that overlapped at least one TE (Table 1). Most pre-miRNAs were associated to MITEs (Fig. 3) and DNA transposons (Fig. 4). Our analysis provided 60 new cases of high confidence repetition-related miRNAs (Fig. 4, Fig. S1).

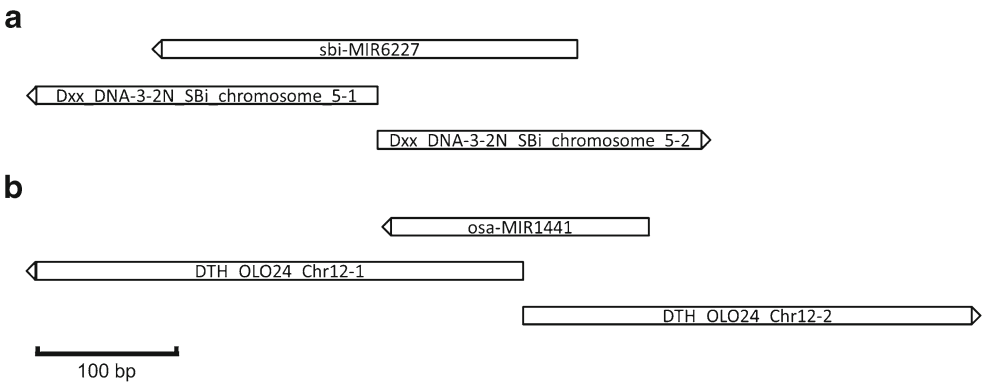


Fig. 6 Juxtaposed TEs possibly structuring pre-miRNAs in grasses. Both cases show an inverted insertion of the same TE. In *Sorghum bicolor*, two DNA-3-2N_SBi non-autonomous DNA transposons span sbi-MIR6227 locus (**a**). In *Oryza sativa*, two OLO24 non-autonomous DNA transposons intersect osa-MIR1441 (**b**)

Nine pre-miRNAs (ath-MIR401, ath-MIR854a, ath-MIR854b, ath-MIR854c, ath-MIR854d, ath-MIR855, osa-MIR812b, osa-MIR814a, and osa-MIR814b) indicated by Piriyapongsa and Jordan (2008) as the products of siRNA-miRNA dual coding TEs were confirmed by our analyses. Four members in a rice miRNA family (osa-MIR812f, osa-MIR812h, osa-MIR812i, and osa-MIR812j) and osa-MIR1850 were formerly classified as typical TE-MIRs (Li et al. 2011), since they are in conformity to a standardized protocol of miRNA annotation rules (Meyers et al. 2008). Ten miRNAs are indicated by miRBase to have transcriptional evidence, and ten were found in miRNEST deep-seq predictions. Only two of them are present in both repositories.

To our knowledge, few small RNA precursors have been reported to be formed by juxtaposed TE insertions in plant species (Kuang et al. 2009; Li et al. 2011; Zhang et al. 2011). One of these cases is osa-MIR1879, which was classified as a bona fide miRNA spanning two short non-autonomous retrotransposons. Other two pre-miRNAs (osa-MIR815b and osa-MIR815c) have similar structures, but were suggested as potential pre-evolved miRNAs (Li et al. 2011). Our analysis detected these pre-miRNAs, but they intersected only one TE. However, we found three cases of stem-loop structures formed by TE juxtaposition in potato, sorghum, and rice. In *Solanum tuberosum*, two SONATA2 non-autonomous DNA transposons on the same strand compose the stu-MIR8019 foldback structure (Fig. 5). In *Sorghum bicolor*, two DNA-3-2N_Sbi non-autonomous DNA transposons give rise to sbi-MIR6227 (Fig. 6a). Within *Oryza sativa*, the insertion of two OLO24 non-autonomous DNA transposons on opposite strands probably gave rise to osa-MIR1441 (Fig. 6b).

Using an adapted Reciprocal Best BLAST Hit method, we found that 92.11 % of the matches were species-specific. This result emphasizes that the repetitive element-related miRNAs tend to be species-specific (Sun et al. 2012).

Conclusions

To our knowledge, PlanTE-MIR DB is the first resource storing the putative relationship between TEs and miRNAs in plants. The database delivers, through a user-friendly web interface, several file formats to facilitate understanding and use of the available data. Future versions will update the database to support data provided by other studies. The discovery of new TE-MIRs strongly relies on comprehensive TE annotation, which is still a drawback for several species. Thus, de novo TE annotation for organisms for which there is available data in miRBase would be a valuable resource that would promote future discoveries. Also, new releases of Repbase Update, miRBase, and plant species genome assemblies should be considered in the next versions of PlanTE-MIR DB. In conclusion, we present a new resource, PlanTE-MIR DB, which allowed us to find new TE-MIR overlaps. We believe that PlanTE-MIR DB can supply insight into the evolution of TEs and miRNAs that will be of great value to the scientific community interested in this subject.

Acknowledgments We thank Romain Guyot (*Institute de recherche pour le développement* - IRD, Montpellier, France) for insightful comments on the TE annotation methods and for his web interface suggestions. APRL received a CAPES fellowship, and GYAdA received a *Fundação Araucária* fellowship. DSD studies on transposable elements are funded by a CAPES/CNPq "Science without boarders" grant (process 084/13). This manuscript was reviewed by a professional science editor and by a native English-speaking copy editor to improve readability.

References

Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW, Carrington JC (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. Nat Genet 36(12). doi:10.1038/ng1478

Axtell MJ (2013) Classification and comparison of small RNAs from plants. Annual review of plant biology 64(January), doi:10.1146/annurev-arplant-050312-120043

Baidouri ME, Panaud O (2013) Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. Genome Biol Evol 5(5). doi:10.1093/gbe/evt025

Barrera-Figueroa BE, Gao L, Wu Z, Zhou X, Zhu J, Jin H, Liu R, Zhu JK (2012) High throughput sequencing reveals novel and abiotic stress-regulated microRNAs in the inflorescences of rice. BMC Plant Biol 12(1). doi:10.1186/1471-2229-12-132

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annual review of plant biology 65(February), doi:10.1146/annurev-arplant-050213-035811

Budak H, Akpinar BA (2015) Plant miRNAs: biogenesis, organization and origins. Funct Integr Genom 15(5). doi:10.1007/s10142-015-0451-2

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinform 10. doi:10.1186/1471-2105-10-421

Erson-Bensan AE (2014) Introduction to microRNAs in biological systems. 26. In: miRNomics: MicroRNA Biology and Computational Analysis. Springer

Gim JA, Ha HS, Ahn K, Kim DS, Kim HS (2014) Genome-wide identification and classification of microRNAs derived from repetitive elements. Genom Inf 12(4)

Hadjiargyrou M, Delihas N (2013) The intertwining of transposable elements and non-coding RNAs. Int J Mol Sci 14(7). doi:10.3390/ijms140713307

Jurka J, Kapitonov VV, Pavlicek a KlonowskiP, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cy35 Togenet Genome Re 110(1-4). doi:10.1159/000084979

Kozomara A, Griffiths-Jones S (2014) MiRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42(D1). doi:10.1093/nar/gkt1181

Kuang H, Padmanabhan C, Li F, Kamei A, Bhaskar PB, Ouyang S, Jiang J, Robin Buell C, Baker B (2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: New functional implications for MITEs. Genome Res 19(1). doi:10.1101/gr.078196.108

Kurtoglu KY, Kantar M, Budak H (2014) New wheat microRNA using whole-genome sequence. Funct Integr Genom 14(2). doi:10.1007/s10142-013-0357-9

Levy A, Sela N, Ast G (2008) TranspoGene and microTranspoGene: Transposed elements influence on the transcriptome of seven vertebrates and invertebrates. Nucleic Acids Res 36(SUPPL. 1). doi:10.1093/nar/gkm949

Li Y, Li C, Xia J, Jin Y (2011) Domestication of transposable elements into microrna genes in plants. PLoS ONE 6(5). doi:10.1371/journal.pone.0019212

Lisch D (2013) How important are transposons for plant evolution? Nat Rev Genet 14(1). doi:10.1038/nrg3374

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) ViennaRNA Package 2.0. Algo Mol Biol 6(1). doi:10.1186/1748-7188-6-26

Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington JC, Chen X, Green PJ, Griffiths-Jones S, Jacobsen SE, Mallory AC, Ra M, Poethig RS, Qi Y, Vaucheret H, Voinnet O, Watanabe Y, Weigel D, Zhu JK (2008) Criteria for annotation of plant MicroRNAs. Plant cell 20(12). doi:10.1105/tpc.108.064311

Ou-Yang F, Luo QJ, Zhang Y, Richardson CR, Jiang Y, Rock CD (2013) Transposable element-associated microRNA hairpins produce 21-nt sRNAs integrated into typical microRNA pathways in rice. Funct Integr Genom 13(2). doi:10.1007/s10142-013-0313-8

Piriyapongsa J, Jordan IK (2008) Dual coding of siRNAs and miRNAs by plant transposable elements. RNA (New York, NY) 14(5). doi:10.1261/rna.916708

Quinlan AR, Hall IM (2010) BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26(6). doi:10.1093/bioinformatics/btq033

Ragupathy R, You FM, Cloutier S (2013) Arguments for standardizing transposable element annotation in plant genomes. Trends Plant Sci 18(7). doi:10.1016/j.tplants.2013.03.005

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16(6). doi:10.1016/j.cocis.2008.07.002

Roberts JT, Ea Cooper, Favreau CJ, Howell JS, Lane LG, Mills JE, Newman DC, Perry TJ, Russell ME, Wallace BM, Borchert GM (2013) Continuing analysis of microRNA origins: Formation from transposable element insertions and noncoding RNA mutations. Mob Genet Elem 3(6). doi:10.4161/mge.27755

Roberts JT, Cardin SE, Borchert GM (2014) Burgeoning evidence indicates that microRNAs were initially formed from transposable element sequences. Mobile Genet Elem 4. doi:10.4161/mge.29255

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Ma Rajandream, Barrell B (2000) Artemis: sequence visualization and annotation. Bioinformatics (Oxford England) 16(10). doi:10.1093/bioinformatics/16.10.944

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). doi:10.1371/journal.pone.0034092

Szcześniak MW, Makałowska I (2014) MiRNEST 2.0: A database of plant and animal microRNAs. Nucleic Acids Research 42(D1). doi:10.1093/nar/gkt1156

Tempel S, Pollet N, Tahi F (2012) ncRNAclassifier: a tool for detection and classification of transposable element sequences in RNA hairpins. BMC Bioinform 13(1). doi:10.1186/1471-2105-13-246

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8(12). doi:10.1038/nrg2165-c4

Zhang Y, Jiang WK, Gao LZ (2011) Evolution of microRNA genes in Oryza sativa and Arabidopsis thaliana: An update of the inverted duplication model. PLoS ONE 6(12). doi:10.1371/journal.pone.0028073

Zhou M, Sun J, Wang QH, Song LQ, Zhao G, Wang HZ, Yang HX, Li X (2011) Genome-wide analysis of clustering patterns and flanking characteristics for plant microRNA genes. FEBS J 278(6). doi:10.1111/j.1742-4658.2011.08008.x