UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CÂMPUS CORNÉLIO PROCÓPIO
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA [PPGBIOINFO]
MESTRADO ACADÊMICO EM BIOINFORMÁTICA

BRUNO HENRIQUE RIBEIRO DA FONSECA

MODELAGEM, INTEGRAÇÃO E ANÁLISE EXPLORATÓRIA DE DADOS PÚBLICOS DE MIRTRONS

CORNÉLIO PROCÓPIO, PR
2018

Bruno Henrique Ribeiro da Fonseca

Modelagem, integração e análise exploratória de dados públicos de mirtrons

Dissertação apresentada como requisito parcial a obtenção de grau de Mestre em Bioinformática pela Universidade Tecnológica Federal do Paraná - Campus Cornélio Procópio.

Orientador: Prof. Dr. Alexandre Rossi Paschoal
Coorientador: Prof. Dr. Douglas Silva Domingues

Cornélio Procópio, PR
2018

Dados Internacionais de Catalogação na Publicação

F676 Fonseca, Bruno Henrique Ribeiro da

Modelagem, integração e análise exploratória de dados públicos de mirtrons / Bruno Henrique Ribeiro da Fonseca. – 2018.
77 f. : il. color. ; 31 cm.

Orientador: Alexandre Rossi Paschoal.
Coorientador: Douglas Silva Domingues.
Dissertação (Mestrado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Bioinformática. Cornélio Procópio, 2018.
Bibliografia: p. 63-69.

1. RNA. 2. MicroRNA. 3. Banco de dados. 4. Bioinformática – Dissertações. I. Paschoal, Alexandre Rossi, orient. II. Domingues, Douglas Silva, coorient. III. Universidade Tecnológica Federal do Paraná. Programa de Pós-Graduação em Bioinformática. IV. Título.

CDD (22. ed.) 572.80285

Biblioteca da UTFPR - Câmpus Cornélio Procópio
Bibliotecário/Documentalista responsável:
Romeu Righetti de Araujo – CRB-9/1676

Título da Dissertação Nº 09:

"MODELAGEM, INTEGRAÇÃO E ANÁLISE EXPLORATÓRIA DE DADOS PÚBLICOS DE MIRTRONS".

por

Bruno Henrique Ribeiro da Fonseca

Orientador: Prof. Dr. Alexandre Rossi Paschoal
Coorientador: Prof. Dr. Douglas Silva Domingues

Esta dissertação foi apresentada como requisito parcial à obtenção do grau de MESTRE EM BIOINFORMÁTICA – Linha de Pesquisa: Biologia Computacional e Sistêmica, pelo Programa de Pós-Graduação em Bioinformática – PPGBIOINFO – da Universidade Tecnológica Federal do Paraná – UTFPR – Câmpus Cornélio Procópio, às 09h 00min do dia 14 de setembro de 2018. O trabalho foi _____ pela Banca Examinadora, composta pelos professores:

Prof. Dr. Alexandre Rossi Paschoal
(Presidente)

Prof. Dr. Fábio Fernandes da Rocha Vicente
(UTFPR-CP)

Prof. Dr. Francis de Morais Francos Nunes
(UFScar-SP)
Participação à distância via _____

Visto da coordenação:

André Yoshiaki Kashiwabara
Coordenador do Programa de Pós-Graduação em Bioinformática
UTFPR Câmpus Cornélio Procópio

A Folha de Aprovação assinada encontra-se na Coordenação do Programa.

AGRADECIMENTOS

De forma especial, agradeço a meus pais Edilaine e Célio, por todo amor, cuidado e apoio incondicional de sempre; a meu irmão Rodrigo, pela amizade, presença e companheirismo; aos irmãos Pietro e Lorena pela alegria que proporcionam; e aos demais familiares por todo o suporte oferecido.

Agradeço à Sthefani, namorada esposa e/ou esposa namorada, por toda paciência, dedicação, companheirismo e apoio durante esta caminhada. Sem tudo isso, o caminho certamente teria sido mais árduo.

Aos amigos, Vinicius, Juliano, Rômulo e tantos outros, pela amizade verdadeira, compreensão e suporte prestado.

Aos meus orientadores Alexandre Rossi Paschoal e Douglas Silva Domingues agradeço profundamente por todo profissionalismo, paciência, sabedoria, dedicação, e amadurecimento pessoal, científico e acadêmico proporcionado.

Aos queridos professores e pesquisadores André Kashiwabara, Franscismar Guimarães, Laurival Vilas Boas, Fabrício Lopes e Fábio Vicente agradeço por serem disseminadores do conhecimento e exemplos de profissionais.

À Sumara Philip e Humberto Rampazzo, por permitirem que meus compromissos profissionais não impedissem a realização do sonho acadêmico.

Aos colegas de turma do PPGBIOINFO-UTFPR-CP, Samara Mireza, Ricardo Medeiros, Isaque Katahira, Daniel Longhi e Tatianne Negri por todo o aprendizado e companheirismo.

RESUMO

FONSECA, B. H. R.. **Modelagem, integração e análise exploratória de dados públicos de mirtrons.** 2018. 76 p. Dissertação do curso de Mestrado em Bioinformática. Universidade Tecnológica Federal do Paraná, Cornélio Procópio, 2018.

MicroRNAs (miRNAs) são uma das classes de RNAs não-codificantes (ou do inglês *non-coding RNA* - ncRNAs) mais estudadas na literatura. Essa classe de pequenos ncRNAs atua no controle celular de diversos processos biológicos, por meio de seu papel regulatório pós-transcricional nos níveis de RNA mensageiro na célula. Em geral, para que miRNAs tornem-se maduros e aptos a executar seu papel regulatório, é necessário que duas clivagens ocorram em sua biogênese canônica. Estudos realizados em *Drosophila melanogaster* e *Caenorhabditis elegans* descreveram uma subclasse de miRNAs que utilizam um meio alternativo à primeira etapa de sua biogênese, os chamados mirtrons. Em suma, os mirtrons utilizam o processo de *splicing* como alternativa à primeira clivagem, e então prosseguem as demais etapas do processo biogênico canônico. Mirtrons localizam-se em pequenos introns e são associados a diversos processos regulatórios, como o de potencial silenciador de genes causadores de doenças em vertebrados e reguladores no processo de fotossíntese em plantas. Apesar de existirem diversos estudos sobre mirtrons, seus dados estão disponíveis de maneira dispersa, sem qualquer organização ou repositório para consulta. Diferenciar comparativamente miRNAs e mirtrons permite que sejam criadas abordagens em biologia computacional capazes de auxiliar estudos biológicos em ncRNAs. Deste modo, este trabalho apresenta duas principais contribuições: (i) desenvolver um repositório amigável sobre dados públicos de mirtrons; e (ii) realizar análise exploratória de modo a comparar e investigar características capazes de distingui-los de miRNAs. Tais contribuições permitem, portanto, a inclusão de uma nova camada na compreensão sobre mirtrons, bem como nas pesquisas sobre miRNAs.

Palavras-chave: RNA não-codificante, miRNA, mirtrons, banco de dados, análise exploratória, mineração de dados.

ABSTRACT

FONSECA, B. H. R .. **Modeling, integration and exploratory analysis of mirtrons public data**. 2018. 76 p. Master's thesis of Bioinformatics Course. Federal University of Technology - Paraná, Cornélio Procópio, 2018.

MicroRNAs (miRNAs) are the most studied non-coding RNA class in literature. This small ncRNA class acts in cellular control of several biological processes, through its post-transcriptional regulatory role in messenger RNA levels. Overall, to miRNAs become matures and able to perform their regulatory function, two cleavages must occur in their canonical biogenesis. Studies in *Drosophila melanogaster* and *Caenorhabditis elegans* have described a miRNA subclass that uses an alternative way to their biogenesis first stage, and they were called mirtrons. The mirtrons use the splicing process as an alternative to the first cleavage and then proceed in the canonical biogenic process. Mirtrons are located in small introns and associated with several regulatory processes, such as the potential diseases genes silencer in vertebrates and regulators in the photosynthesis process in plants. Although there are several studies about mirtrons, their data is available in a dispersed way, without any organization or repository to query. Differentiating comparatively miRNAs and mirtrons allows advances in computational biology that supporting biological studies in ncRNAs. Thus, this paper presents two main contributions: (i) to develop a friendly repository of public mirtrons data; and (ii) perform exploratory analysis to compare and investigate features capable of distinguishing mirtrons from miRNAs. These contributions allow a new layer to the understanding about mirtrons and miRNAs research.

Keywords: non-coding RNA, miRNA, mirtrons, database, exploratory analysis, data mining.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1. MOTIVAÇÃO

Dentre as diversas classes de RNAs não-codificantes (ncRNAs), a dos microRNAs (miRNAs) é uma das mais estudadas na literatura (GLASGOW; DE SANTI; GREENE, 2018). Estes pequenos ncRNAs apresentam processo biogênico e regulatório distintos em mamíferos e plantas, sendo sua principal função a de regulação dos níveis de RNA mensageiro (mRNA) nos organismos eucariotos. Para que o miRNA se torne maduro e apto a desempenhar seu papel regulatório, sua biogênese canônica é constituída de duas clivagens.

Os mirtrons são pequenos miRNAs de via biogênica alternativa, localizados em introns. Diferentemente dos miRNAs canônicos, os mirtrons tem a primeira etapa de clivagem de maturação via mecanismo de *splicing* (BEREZIKOV et al., 2007). Os mirtrons também possuem função regulatória, como o de potencial silenciador de genes causadores de doenças em vertebrados (SIBLEY et al., 2012) e de regulador no processo de fotossíntese em plantas (MENG; SHAO, 2012).

Sabe-se que há 76 artigos científicos sobre mirtrons, considerando NCBI PubMed até Novembro/2017. Deste total, 22 artigos possuem dados públicos passíveis de utilização para investigação em bioinformática, porém tais dados são encontrados de maneira dispersa e sem organização. Mesmo o miRBase (KOZOMARA; GRIFFITHS-JONES, 2013), o repositório estado-da-arte em miRNA, não possui seção para dados de mirtrons. Em outras palavras, não há repositório amigável e público para consulta e extração de informações sobre mirtrons.

Há de se considerar ainda que embora miRNAs e mirtrons possuam similaridades biogênicas, caracterizá-los e diferenciá-los comparativamente pode permitir uma nova camada no conhecimento de seus papéis regulatórios.

Nesse sentido, este trabalho foi realizado com o propósito de desenvolver e disponibilizar: (i) um repositório amigável sobre dados de mirtrons; e (ii) análise, de forma exploratória e comparativa, de características capazes de possibilitar a investigação de similaridades de miRNAs e mirtrons.

1.2. RNA NÃO-CODIFICANTE

RNA não-codificante ou não-codificadores (do inglês *non-coding RNA* - ncRNAs) são ácidos ribonucleicos (ARN e em inglês, RNA - *Ribonucleic Acid*) transcritos do DNA que não são traduzidos em proteínas, ainda que sua função não seja definida (ATIANAND; CAFFREY; FITZGERALD, 2017).

Os ncRNAs foram descritos relacionados a diversos processos biológicos, tais como no processo de tradução (snoRNA), processamento do mRNA (snRNA), repressão de transposons (piRNA), modificação e silenciamento da cromatina (lncRNA) e regulação de expressão gênica (miRNA) (PANIR et al., 2018).

A desregulação de alguns tipos de ncRNAs, tais como miRNAs e lncRNAs, é associada a origem de várias patologias fisiológicas e comportamentais, incluindo a tumorigênese e doenças de cunho neurodegenerativa (QU; ADELSON, 2012). Acredita-se, ainda, que os ncRNAs apresentem grande potencial e aplicabilidade para exploração terapêutica (CURTIS; SIBLEY; WOOD, 2012).

Os ncRNAs originam-se de várias regiões genômicas, que podem ser de regiões codificadoras (e.g. exônicas) ou não-codificadoras (e.g. intrônicas, intergênicas) (QU; ADELSON, 2012). A classificação dos ncRNAs é realizada considerando o tamanho, formato, ou função da molécula. De forma arbitrária, foi definido que RNAs pequenos são aqueles que possuem até 200 nucleotídeos (nt) de tamanho e os RNAs longos acima de 200 nt. A Tabela 1 apresenta as principais classes de RNA não-codificantes, classificados de acordo com seus tamanhos.

Dentre as classes, destaca-se os miRNAs, os quais são considerados a classe com maior interesse e dados em pesquisas científicas (PASCHOAL et al., 2012), os quais são inclusive objeto de estudo deste trabalho. Recomenda-se a leitura da revisão de Palazzo e Lee (2015) para detalhes sobre os ncRNAs, classes, quantidades e afins.

Tabela 1. Principais classes de RNA não-codificantes (ncRNA).

Nome	Acrônimo	Tamanho (nt	Notas
RNAs longos não-codificantes	lncRNA	>200	Transcritos de codificação não proteíca; classe heterogênea de RNAs
Região transcrita ultra conservada	T-UCR	≈50 - 570	Frequentemente localizado em locais frágeis e regiões genômicas associadas ao câncer; possivelmente regulado por miR
Circular RNA	circRNA	≈100 - 1600	Anéis de RNA fechados covalentemente; alguns possuem funções de codificação; potencial regulador genéticos e “armadilhas” de miR.
RNAs de pequena interferência	siRNA	20 - 25	RNAs de cadeia dupla semelhantes ao miR, operando através da via de RNA de interferência (RNAi); promove a degradação do mRNA
Y RNA	Y RNA	21 - 24	Necessário para a replicação do DNA através de interações com a cromatina e proteínas de iniciação; alvo de anticorpos auto-imunes
Micro-RNA	miRNA; miR	21-24	Função no silenciamento de RNA e regulação pós-transcricional da expressão gênica; pode ter uma localização extracelular
RNA de interaçãoes Piwi	piRNA	26-31	Silenciamento gênico epigenético e pós-transcricional de retrotransposons e outros elementos genéticos em células da linhagem germinativa
Pequenos RNAs nucleolares	snoRNAs	60-300	Guia modificações químicas de outros RNAs (rRNA, tRNA, snRNA)
Pequeno Ácido Ribonucléico Nuclear	snRNA; U-RNA	≈150	Função no processamento de RNA pré-mensageiro (hnRNA) no núcleo; Auxílio na regulação de fatores de transcrição; manutenção de telômeros

Fonte: Adaptado de Gulìa et al. (2017)

1.3. microRNA OU miRNA

MicroRNAs ou miRNAs são pequenos ncRNAs, cuja molécula madura apresenta tamanho entre 21 a 25 nt em animais (KATZ et al., 2016) e 21 a 22 nt em plantas (AXTELL; MEYERS, 2018). Esta classe de ncRNA está presente em genomas de eucariotos, tendo como função, em geral, a inibição pós-transcricional dos níveis de mRNAs na célula via o pareamento complementar (LAGOS-QUINTANA et al., 2001).

A biogênese do miRNA difere entre mamíferos e plantas (ZHANG et al., 2018) e o fato de os miRNAs estarem relacionados ao controle de diversos processos biológicos, inclusive a doenças como o câncer, torna-os de grande interesse de pesquisas na área (CURTIS; SIBLEY; WOOD, 2012). A Figura 1 apresenta a biogênese de miRNAs em mamíferos.

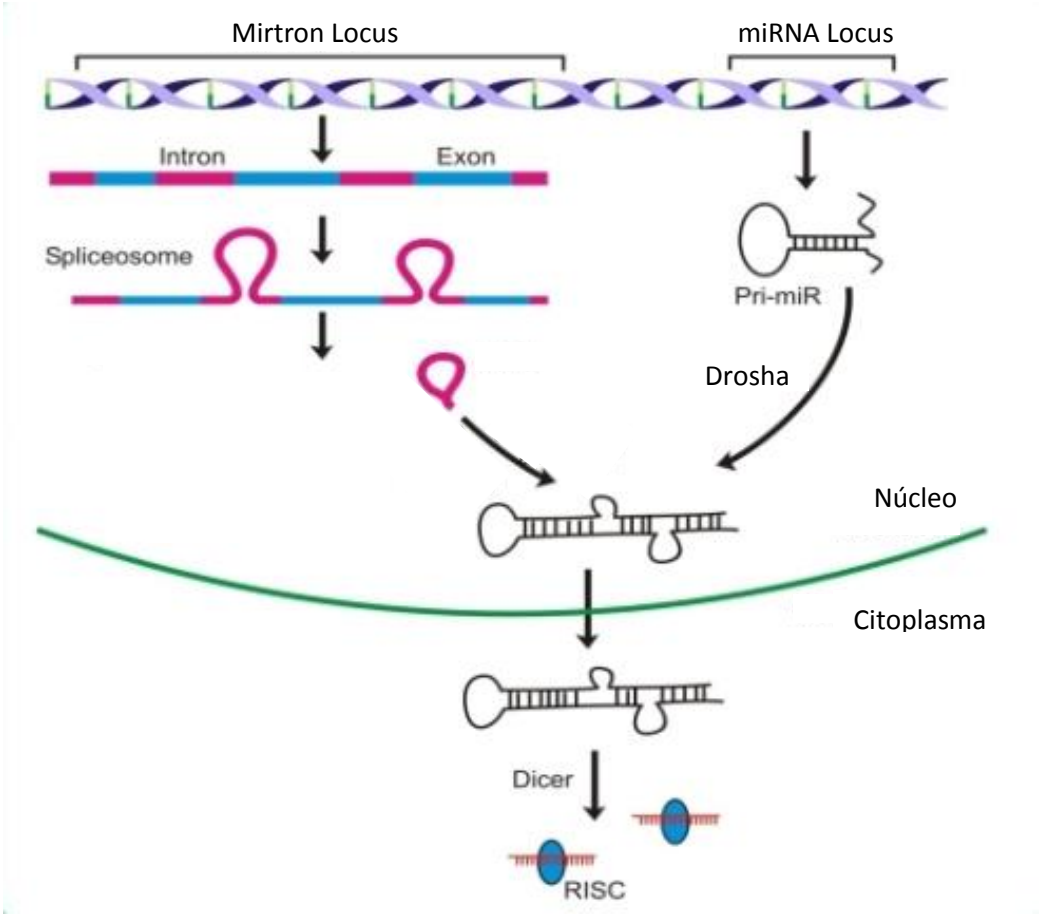


Figura 1. Biogênese de miRNAs canônicos e mirtrons, em mamíferos. Para que miRNAs tornem-se maduros e aptos a desempenhar seu papel é necessário que Drosha e Dicer clivem o transcrito, já para mirtrons apenas Dicer, sendo a primeira clivagem omitida dada sua conformação resultante de processo de splicing.
Fonte: Adaptado de Naqvi et al (2009)

Conforme apresentado na Figura 1, em mamíferos, após transcrição em miRNA primário (pri-miRNA), ocorrem duas etapas de clivagem. A primeira etapa é realizada pelo complexo Drosha-DGCR8 que gera o miRNA precursor (pre-miRNA), ou também conhecido *hairpin loop* ou *stem loop*. Em seguida, após a saída do núcleo celular, há a segunda clivagem, esta realizada pela enzima Dicer, quando então é formado o miRNA maduro, contendo cerca de 21~25 nucleotídeos. Posteriormente, o miRNA maduro é guiado para a realização de seu papel regulatório no mRNA (MENG; SHAO, 2012).

Diferentemente de animais, em plantas ambas as clivagens ocorrem por uma enzima homóloga a Dicer, chamada Dicer-like1 (DCL1). A DCL1 é expressa

somente no núcleo celular de plantas, indicando que ambas as reações acontecem dentro do núcleo. Posteriormente, o miRNA é transportado para o citoplasma para desempenhar seu papel regulatório (LELANDAIS-BRIÈRE et al., 2010).

O silenciamento gênico, por parte da atuação do miRNA, ocorre com o auxílio de um complexo ribonucleoprotéico, composto por um conjunto de proteínas associadas denominado de “Complexo de Indução do Silenciamento do RNA” (RISC, do inglês *RNA-induced silencing complex*). Deste modo, o miRNA pode se ligar de forma complementar ao RNA mensageiro alvo e impossibilitar que os ribossomos consigam acessar a informação genética necessária, acarretando a diminuição da síntese proteica específica do gene, seja através da degradação do mRNA ou inibindo sua tradução (BARTEL, 2004).

Para degradação do mRNA, os miRNAs maduros reconhecem sítios de ligação parcialmente complementares a suas sequências “*seeds*” (nucleotides da posição 2 a 8), e após se acoplarem, proteínas da família Argonaute, presentes no complexo de indução de silencionamento, realizam a clivagem e impedem que o mRNA realize sua função. Com relação a inibição da tradução, o pareamento entre a região *seed* do miRNA ao mRNA também pode atuar como impedidor da interação dos ribossomos com os mRNAs (BRAUN; HUNTZINGER; IZAURRALDE, 2012).

Recomenda-se a revisão de Moran et al (2017) para mais detalhes sobre miRNAs em animais e plantas.

1.4. MIRTRONS

Em 2007, Ruby, Jan, e Bartel (2007) e Okamura et al (2007) identificaram que em *D. melanogaster* e *C. elegans* haviam alguns pequenos íntrons que possuíam via biogênica alternativa, sem passar pela clivagem da Drosha, e que produziam um transcrito com as características estruturais dos pre-miRNAs. Estes elementos identificados foram denominados “mirtrons”. Para Ruby et al. (2007), os mirtrons são objeto de vários estudos para investigação de seu verdadeiro papel biológico.

Diferente do processo canônico de formação do miRNA, que se inicia com o reconhecimento e clivagem do miRNA primário (pri-miRNA) pela enzima RNase III Drosha, os mirtrons utilizam-se de sua maquinaria de *splicing* em íntrons curtos com potencial *hairpin* como alternativa. Posteriormente, miRNAs e mirtrons seguem a

mesma rota da biogênese via transporte pela Exportase-5, clivagem pela Dicer para a formação do miRNA maduro, e posterior execução de seus papéis regulatórios (OKAMURA et al., 2007). A Figura 1 apresenta as etapas do processo biogênico de mitrons.

Há indícios de que mirtrons desenvolveram-se de forma independente, em diferentes famílias de animais, tomando como base as estruturas já identificadas ao longo da história, as quais possuem *hairpins* curtos no citoplasma e são especificamente reconhecidos e processados pela enzima Dicer (CHUNG et al., 2011). Com a aplicação de técnicas "*deep sequencing*", as análises de bibliotecas de RNA tornaram-se um dos meios mais utilizados por pesquisadores para a descoberta de mirtrons (WESTHOLM; LAI, 2011).

Para Curtis, Sibley, e Wood (2012), uma importante aplicação de mirtrons é através de sua utilização em "*knockdown and replacement*" (nocaute e substituição) de um gene, em situações em que uma mutação provoca uma condição dominante e não saudável. Basicamente, mirtrons seriam utilizados para bloquear a expressão de um gene, e posteriormente, promover a substituição de seu alvo.

Do ponto de vista bioinformata, mirtrons possuem uma característica estrutural que difere do miRNA canônico, sendo essa uma explicação para a diferença na clivagem durante a biogênese (OKAMURA et al., 2007). Este fato é uma evidência que análises comparativas entre miRNA e mirtrons podem elucidar diferenças e similaridades entre ambos, o que é um problema em aberto.

1.5. RNA COMPETIDOR ENDÓGENO OU ceRNA

Em 2007, estudos apresentaram um novo conceito envolvendo regulação de miRNAs, o mimetismo alvo (*target mimicry*) (FRANCO-ZORRILLA et al., 2007). A hipótese é que transcritos que possuem região similar ou igual de ligação do miRNA, passam a "sequestra-lo", e como consequência, inibem o miRNA, permitindo a tradução do mRNA. Essa região de ligação do miRNA é denominada de elemento de reconhecimento do miRNA (MRE) (do inglês, *miRNA Recognition Element*) (SALMENA et al., 2011).

Os transcritos que competem pelo miRNA podem ser pseudogenes, *longos ncRNAs* (lncRNAs), RNA *circular* (circRNA), RNA viral ou mesmo mRNAs, sendo todos estes denominados ceRNAs (do inglês, *Competing endogenous RNAs*)

(THOMSON; DINGER, 2016). Existe, ainda, a hipótese sobre a ocorrência de RNAs competidores endógenos em dados de mirtrons, para a qual ainda não há registros de estudos publicados.

Destaca-se que lncRNAs são transcritos não-codificantes de proteínas com tamanho superior a 200 nt e que originam-se dos mais variados locais do genoma; os pseudogenes são genes que perderam sua capacidade de codificação de proteínas devido a várias incapacidades genéticas; enquanto os circRNAs são um tipo de RNA não-codificantes que formam um ciclo contínuo covalentemente fechado sem polaridade 5' a 3' ou cauda poliadenilada (LE et al., 2016).

Apesar de ceRNA ser um termo geral, em animais é conhecido como "*miRNA sponge*" ou "*miRNA decoy*" (EBERT; SHARP, 2010), enquanto que em planta denomina-se "*target mimic*" ou "*target mimicrys*" (RUBIO-SOMOZA et al., 2011). A Figura 2 ilustra a hipótese sobre como os ceRNAs atuam, embora ainda exista discussão sobre o processo (THOMSON; DINGER, 2016).

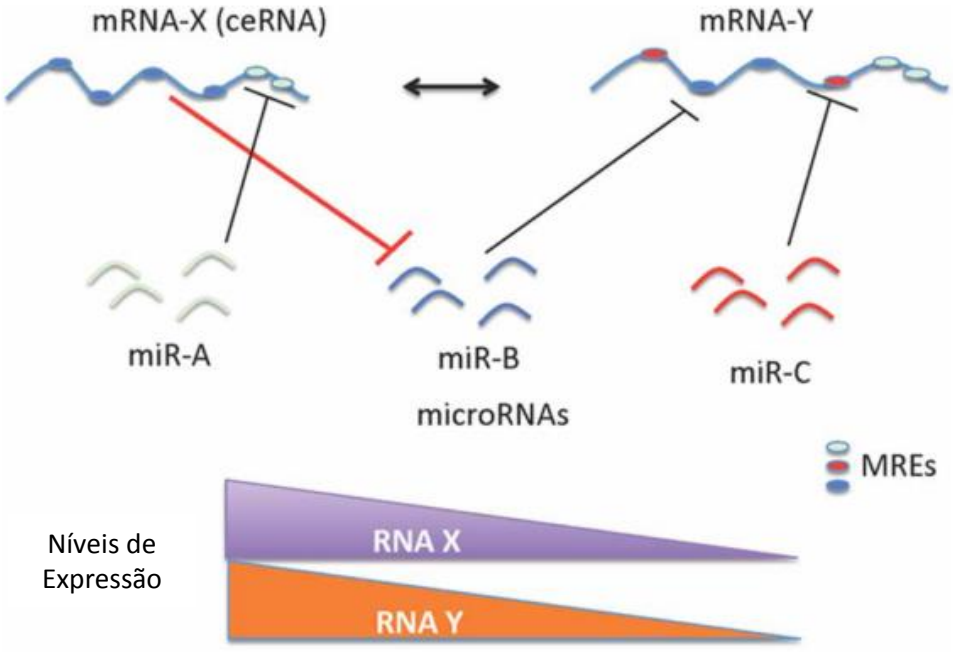


Figura 2. Esquema de regulação gênica por ceRNA; onde tem-se que o mRNA-Y é um alvo funcional de miR-B. Os níveis de expressão do mRNA-X, que contém MREs para o miR-B, podem atuar como esponjas e sequestrar o miR-B, impedindo sua atuação em mRNA-Y. Caso o nível de expressão de mRNA-X venha a diminuir, possivelmente os níveis de mRNA-Y decrescerão também, uma vez que os miR-B estarão livres a atuar e regular mRNA-Y. Adaptado de Kartha e Subramanian (2014)

1.6. BIOINFORMÁTICA NA PESQUISA DE MIRTRONS

Dos 140 bancos de dados de ncRNAs descritos na literatura, metade é dedicado a classe dos miRNAs (PASCHOAL et al., 2012; NRDR Site versão 2016). No que tange a fonte de dados, o banco de dados considerado estado-da-arte para miRNA é o miRBase (KOZOMARA; GRIFFITHS-JONES, 2013). Entretanto, apesar da quantidade de bancos, até o momento não há na literatura científica um repositório de mirtrons que se apresente de forma estruturada e com interface amigável para não bioinformatas. A hipótese é que isso ocorra em virtude da descoberta de mirtrons ser algo consideravelmente recente (descrito em 2007).

Por meio de pesquisa realizada neste trabalho, foi possível identificar que há 3.833 mirtrons dispersos em artigos científicos, principalmente para *Arabidopsis thaliana* e *Oryza sativa* (MENG; SHAO, 2012), *Homo sapiens* e *Mus musculus* (LADEWIG et al., 2012), e em *D. melanogaster* e *C. elegans* (CHUNG et al., 2011). Os dados estão dispersos e disponíveis em vários artigos científicos, com estruturas distintas, ou seja, sem um repositório integrado em ambiente web, tornando-se assim uma abordagem com potencial de exploração.

A aplicação da bioinformática para integrar dados de mirtrons possibilitará a condição mínima necessária para investigar as características dos mirtrons, inclusive de forma comparativa aos miRNAs. Por fim, tem-se o interesse na compreensão do evento de ceRNAs nos mirtrons, condição também ainda não explorada.

1.7. ANÁLISE EXPLORATÓRIA DE DADOS

1.7.1. Mineração de dados

O desenvolvimento de técnicas e tecnologias relacionadas a genômica e proteômica tem gerado expressivo aumento em relação a quantidade de dados biológicos disponibilizados nos últimos anos. Poder utiliza-los de maneira massiva a favor da geração de conhecimento requer, cada vez mais, sofisticadas análises computacionais. Para tanto, técnicas de *data mining* (mineração de dados) têm sido amplamente aplicadas (RAZA, 2010).

D*ata mining* (mineração de dados) é tido como o processo de exploração e análise de grandes quantidades de registros em busca da identificação de padrões e informações úteis. Em mineração de dados, dentre os modelos de identificação de padrões, há dois tipos principais: o preditivo e o descritivo. O modelo preditivo se baseia em técnicas de regressão e árvores de decisão para previsão de um determinado resultado, enquanto o modelo descritivo permite melhor compreensão dos dados por meio de análise, sem qualquer variável-alvo específica (LEVENTHAL, 2010).

Através da aplicação de modelos de identificação de padrões, pode-se: (i) classificar dados; (ii) estimar e predizer situações ou resultados com base em parâmetros e situações ocorridas previamente; (iii) associar ocorrências aos resultados; (iv) segmentar e agrupar dados e características, e (v) descrever e visualizar dados de forma analítica (RAZA, 2010). Para descrição e visualização de dados utiliza-se, principalmente, técnicas estatísticas de análise exploratória de conjunto de dados.

Para Kaur e Singh (2017), os principais problemas decorrentes em mineração de dados são: a baixa qualidade dos dados (dados ruidosos, sujos e de tamanho inadequado); a ocorrência de redundância de fontes; a aplicação de algoritmos de mineração de dados não efetivos; e a dificuldade no processamento de dados não estruturados.

Para que dados sejam minerados de maneira eficiente, é necessário que os dados sejam: (i) pré-processados, uma vez que, geralmente, os dados são coletados brutos, ruidosos, incompletos e inconsistentes; (ii) limpos, tratando valores ausentes, suavizando ruídos, excluindo valores discrepantes e resolvendo inconsistências; (iii) integrados, combinando dados de diferentes formatos e fontes; e (iv) tratados, para adequação à análise a ser realizada (LAN et al., 2018).

1.7.2. Análise exploratória

Análise exploratória de dados refere-se ao processo de sumarização dos dados utilizando informações numérico estatísticas, que podem ser apresentadas via tabelas e gráficos, visando a validação e qualidade dos dados (TEO, 2010). Para Lauretto (2001), a análise exploratória de dados é uma abordagem que busca examinar os dados anteriormente a aplicação de qualquer método estatístico.

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), para a obtenção de conhecimento por meio de análise massiva de dados, é necessário que estes sejam selecionados, preparados, tratados e padronizados, com conhecimento prévio do conjunto de dados e de técnicas específicas, para que então as análises sejam realizadas (Figura 3).

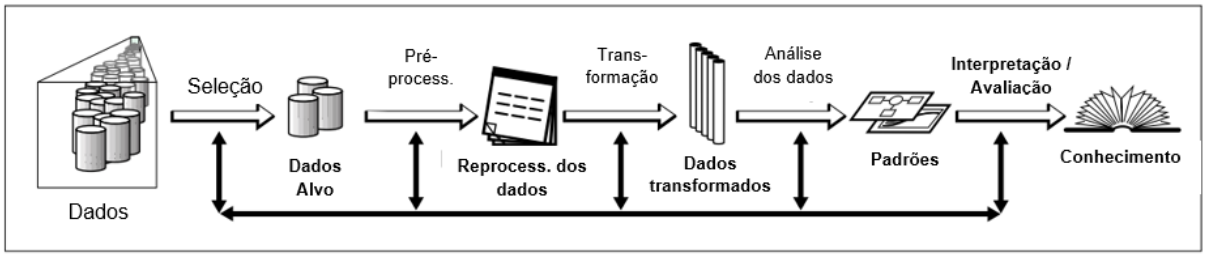


Figura 3. Processo de descoberta de conhecimento, desde o processo de coleta à análise e interpretação de dados para a geração do conhecimento.
Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996)

A análise exploratória de dados é uma abordagem que emprega uma variedade de técnicas que visam entre outras coisas a maximização do entendimento do conjunto de dados, o descobrimento de estruturas subjacentes, e a detecção de *outliers* e anomalias (NIST/SEMATECH, 2003). Fayyad, Piatetsky-Shapiro e Smyth (1996) afirmam ainda que, para que análises de dados sejam realizadas é necessário que haja conhecimento relevante e prévio sobre os objetivos da utilização e aplicação dos dados.

Dentre as técnicas mais empregadas para análise exploratória de dados, destacam-se a utilização de fluxogramas, histogramas, gráficos, técnicas estatísticas simples, tabelas e outras representações gráficas. Através da utilização de técnicas gráficas o entendimento é facilitado e promovido ao analista dos dados (NIST/SEMATECH, 2003).

1.8. OBJETIVOS

1.8.1. Geral

Modelar e disponibilizar dados curados de mirtrons, via repositório amigável, bem como ampliar o conhecimento sobre mirtrons, por meio de análise exploratória comparativa com miRNAs.

1.8.2. Específicos

i. Revisar a literatura sobre mirtrons.
ii. Coletar, modelar, e integrar dados públicos de mirtrons disponíveis na literatura em esquema de banco de dados.
iii. Criar e disponibilizar repositório web amigável para investigação de dados públicos de mirtrons.
iv. Analisar evento de ceRNAs em dados de mirtrons.
v. Realizar análise exploratória comparativa com dados de mirtrons e miRNAs, do ponto de vista de sequência e estrutura.

1.9. ORGANIZAÇÃO DA DISSERTAÇÃO

Este trabalho está estruturado em quatro capítulos principais. Neste Capítulo 1 são apresentados os principais conceitos e o estado-da-arte referente aos temas abordados, bem como os objetivos propostos. No Capítulo 2 encontra-se o artigo "*mirtronDB: a mirtron knowledge base*" (já submetido para publicação), o qual apresenta o repositório de dados de mirtrons resultante da primeira etapa desta pesquisa. O Capítulo 3 detalha a análise exploratória comparativa realizada entre dados de miRNAs e mirtrons. No Capítulo 4 são apresentadas as referências utilizadas para a elaboração deste trabalho.

2. mirtronDB: A MIRTRON KNOWLEDGE BASE

Na literatura atual não há registro de repositório que disponibilize dados de mirtrons de maneira organizada e centralizada. Organizar tal conteúdo, além da condição de exclusividade, permite que estudos para caracterização, atribuição de papéis biológicos, e criação ou aperfeiçoamento de preditores de mirtrons sejam realizados, bem como possibilita a identificação de interações de mirtrons em organismos.

Portanto, sabendo que 22 dos 76 artigos publicados sobre mirtrons na literatura atual possuem dados públicos disponíveis (considerando título e resumo no repositório NCBI Pubmed e período até Novembro de 2017) e que tais dados são disponibilizados de maneira desorganizada, descentralizada e despadronizada, neste capítulo apresentamos metodologia, resultados e discussões da primeira contribuição deste trabalho, a disponibilização do repositório de conhecimento de mirtrons, através do trabalho intitulado “mirtronDB: um repositório de conhecimento de mirtrons”.

Nesta etapa do trabalho os dados coletados foram filtrados, validados e importados para banco de dados modelado exclusivamente para mirtrons. No total, 1.407 mirtrons precursores e 2.426 mirtrons maduros foram identificados. Dentre os dados disponíveis apresentam-se mirtrons por: espécie, *host* gene, localização gênica, tipo (precursor ou maduro), sequência, e ano de publicação.

O repositório encontra-se disponível em <http://mirtrondb.cp.utfpr.edu.br/>, com interfaces para pesquisa, navegação, visualização e download dos dados. Com base nos dados disponíveis no repositório também foram realizadas análises de predição de *target gene*, predição de ceRNA e de caracterização de mirtrons.

Todo o conteúdo apresentado neste capítulo encontra-se na versão de submissão para publicação na revista acadêmica Bioinformatics | Oxford Academics (até a conclusão deste trabalho não havia nenhum retorno do editor responsável sobre a publicação do artigo).

2.1. ABSTRACT

Mirtrons are originated from short introns with atypical cleavage from the miRNA canonical pathway by using the splicing mechanism. Several studies

describe mirtrons in chordates, invertebrates and plants but in the current literature there is no repository that centralizes and organizes public and available data. To fill this gap, we created the first knowledge database dedicated to mirtron, called mirtronDB, available at http://mirtrondb.cp.utfpr.edu.br/. MirtronDB has a total of 1,407 mirtron precursors and 2,426 mature sequences in 18 species. Through a user-friendly interface, users can browse and search mirtrons by organism, organism group, type and name. In summary, mirtronDB is a specialized resource to explore mirtrons and their regulations, providing free, user-friendly access to knowledge on mirtron as a research topic.

2.2. INTRODUCTION

Studies in *D. melanogaster* and *C. elegans* identified that some short hairpin introns presented similar characteristics to microRNAs (miRNAs), however using the splicing mechanism as the first stage of the miRNA biogenesis cleavage (RUBY et al., 2007; OKAMURA et al., 2007). These non-canonical miRNAs described in small introns are collectively called "mirtrons". Mirtrons were later identified in 16 other organisms, including *H. sapiens* and *M. mulatta* (BEREZIKOV et al., 2007), *G. gallus* (GLAZOV et al., 2008) and *O. sativa* (ZHU et al., 2008).

In mammals, mirtrons can act in mRNA regulation processes via RISC complex (OKAMURA et al., 2007) and fat metabolism regulation (ROMAO et al., 2014). Its deregulation was identified as a potential source of several human pathologies (LAURYNAS et al., 2016), and its use in therapeutic treatments has been considered, through use of gene silencing techniques (CURTIS et al., 2017). In plants, an intronic miRNA, miR838, was located in the DCL1 primary transcript in *A. thaliana*, suggesting a feedback loop for the autoregulation of miRNA biogenesis (RAJAGOPALAN et al., 2006; BUDAK; AKPINAR, 2015).

In the current literature, there is no repository for accessing knowledge on mirtron data. Not even miRBase (KOZOMARA; GRIFFITHS-JONES, 2013), the miRNA state-of-the-art repository, has specific analysis for mirtrons. Until November 2017, 76 scientific papers were published using the term mirtrons/mirtron, according to title/abstract search in the PubMed website. From these, 22 articles have available public data. However, these datasets are dispersed, with no standardization or organization.

Organization of these data will enable studies on mirtron characteristics, roles, and interactions in organisms, among other potential scientific advances. In this context, to fill this gap, we provide mirtronDB (http://mirtrondb.cp.utfpr.edu.br/), a central mirtron knowledge data repository. For that, we modelled a total of 1,407 mirtron precursors and 2,426 mature mirtrons from 18 species (chordates, invertebrates and plants) based on published available literature.

MirtronDB has an online user-friendly interface for the user to search, browse, visualize, and download information about mirtrons. All datasets are publicly available in several formats. In particular, the user has access to: (i) precursor mirtron similarity analysis; (ii) target gene predictions; and (iii) ceRNA predictions in plants. We expect this resource can increase the amount studies on mirtrons research.

2.3. MATERIALS AND METHODS

As summarized in Figure 4, mirtronDB was built in four steps: (1) Data collection: literature investigation and data collection; (2) Data modelling: data organization and structured integration; (3) Data analysis: transforming raw data into usable information by using analytical and logical reasoning to examine each component of the data and specific tools to characterize the data collected; and (4) Website interface: a web portal for the scientific community.

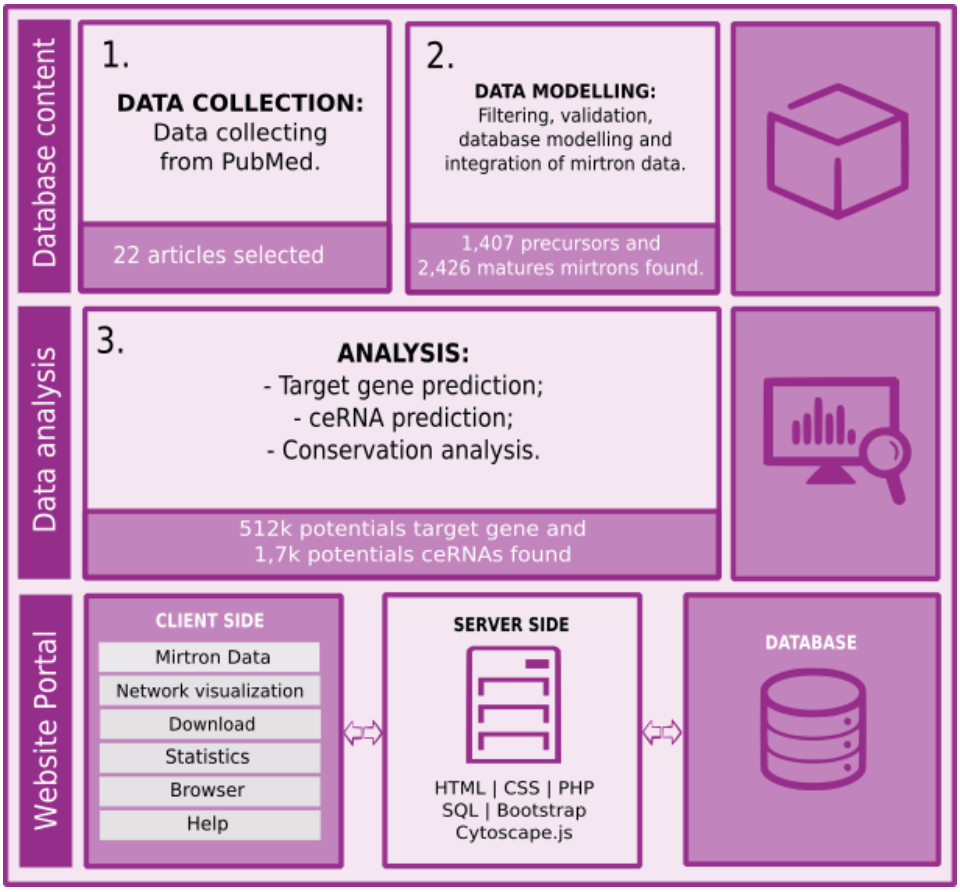


Figure 4. Schematic overview of steps for developing mirtronDB. The elaboration process was splitted in three major groups: database content, data analysis and website portal development.

2.3.1. Mirtron data collection and modelling

We collected the mirtron data available from June 2007 to November 2017 searching by the term "mirtron OR mirtrons" in the fields title/abstract in NCBI PubMed (Appendix 1) and in papers cited in them. The articles selected were manually analyzed and redundancies were removed. We created a standardized name: "organism name abbreviation + the word ‘mirtron’ + ID, and for mature we add the arm". We built a database and automatically imported the information.

We performed database modelling (Appendix 2), and the script implementation was written in the Python scripting language to automatically import the data from the articles to the database management system (PostgreSQL, 16). The importing script was implemented for future database automatic updates.

We performed analysis individually by organism and organism groups. We defined three major organism groups: chordates, invertebrates and plants.

2.3.2. Similarity analysis among organisms

We extracted genomic information from several sources. We used Ensembl 90 (ZERBINO et al., 2018) for chordates and *C. elegans*. For invertebrate genomes, we downloaded data from Flybase version FB2017_04 (GRAMATES et al., 2017). For plants, we used TAIR10 for *Arabidopsis thaliana* (BERARDINI et al., 2015), Phytozome v12.1 (GOODSTEIN et al., 2012) and Ensembl Plants 37 (BOLSER et al., 2007) for plants whose genome was not available in Phytozome. Further information is detailed in Appendix 3.

Based on genomic data, we performed a NCBI BLASTN version 2.6 (CAMACHO et al., 2009) alignment between all precursor mirtrons against all other species genomes (i.e., a mirtron precursor from *H. sapiens* was aligned against all organism genomes in which we found mirtrons, except in *H. sapiens*). We retained results above 95% query coverage and identity in the alignment.

2.3.3. Mirtrons and miRNAs similarity analysis

To identify sequence similarities with miRNAs, mature mirtrons were aligned to all miRNAs available in miRBase v22 (48,885 mature miRNAs) (KOZOMARA; GRIFFITHS-JONES, 2013) using the CD-HIT-EST-2D tool (LI; GODZIK, 2006) and considering the alignment of 9 nucleotides (nt) at 0.98 of cutoff identity.

2.3.4. Target gene prediction

We predicted the putative mirtron targets gene for *H. sapiens* and plants. For predicting human genome target, we used TargetScan (GRIMSON et al., 2007) with default parameters and mature human mirtrons. Data from human UTRs were obtained in the complementary material of the TargetScan tool.

The psRNATarget tool (version 2017 - DAI; ZHUANG; ZHAO, 2018) was used for target gene prediction in plants, with the seed region parameter from 2 to 8 nucleotides. As potential targets, data from plant cDNAs were obtained from the Ensembl Plants database (version 37).

2.3.5. ceRNA prediction in plants

We used TAPIR (version 1.2) (BONNET et al., 2010) to predict ceRNA in plants, with default parameters. All mature mirtrons from *A. thaliana, M. esculenta, M. truncatula, O. sativa* and *S. italica* were compared against all lncRNAs from 45 species obtained from GreeNC database version 1.12 (PAYTUV et al., 2016).

2.3.6. Website implementation

MirtronDB was developed using HTML 5, PHP 7.0, and CSS 4.0, and the front end was designed using the Bootstrap 3.3 framework. The interaction network visualization of targets and ceRNAs was done using the Cytoscape.js library (FRANZ et al., 2015). The relational database has been developed under the PostgreSQL database management system (version 9.6.6).

2.4. RESULTS

2.4.1. Database content

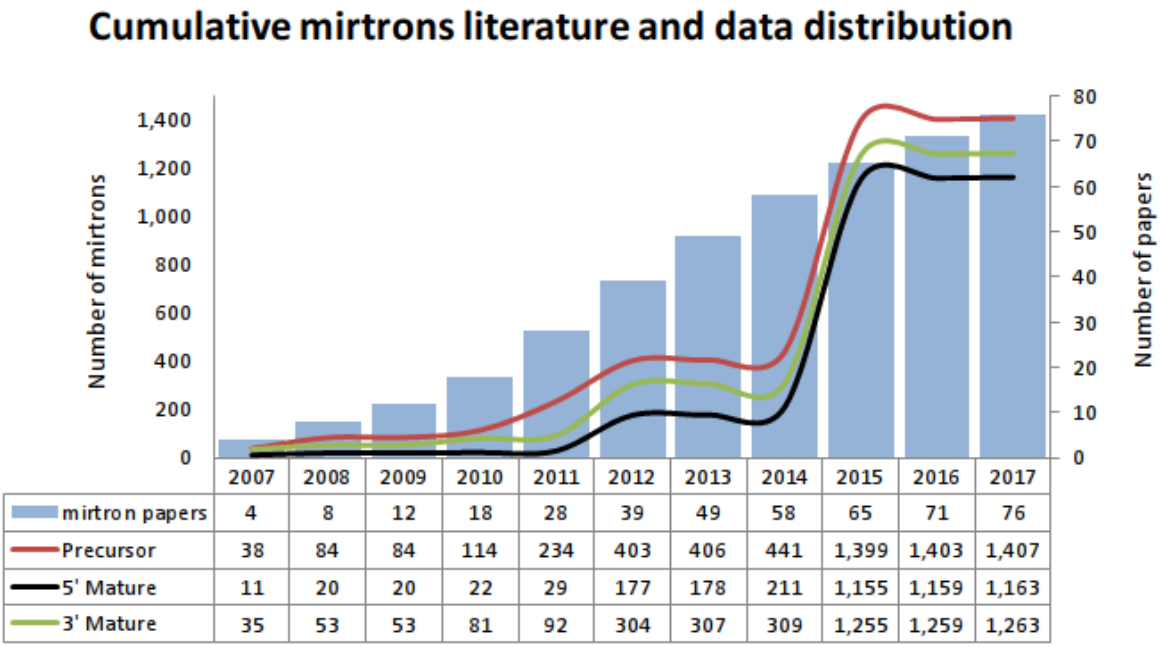
We found a total of 1,407 precursor mirtrons and 2,426 mature mirtrons in 18 species, based on 22 articles collected between July 2007 and November 2017. We extracted functional information, when available, such as species, host gene, genomic location, type (precursor or mature), sequence and year of publication. All mirtrons collected (precursor and mature) and respective target gene(s) and ceRNAs are detailed in Table 2. Mirtron precursor from chordates represents 85.4% of collected precursor data, invertebrates 9.4%, and plants 5.3%. Regarding mature mirtrons, 91.5% are from chordates, 5.7% from plants, and 2.8% from invertebrates.

Table 2. Overall mirtronDB data.

Caption:[1] target gene prediction tools: psRNATarget for plants and TargetScan for humans. [2] ceRNA prediction tool: TAPIR in plants.

Organism	Precursors	5' Mature	3' Mature	Target Gene	ceRNA
Arabidopsis thaliana	5	-	13	394 [1]	4
Bos taurus	1	1	-	-	-
Caenorhabditis elegans	32	8	11	-	-
Canis familiaris	4	4	4	-	-
Danio rerio	2	-	2	-	-
Drosophila melanogaster	75	3	23	-	-
Drosophila pseudoobscura	20	-	19	-	-
Drosophila simulans	5	-	4	-	-
Gallus gallus	73	2	17	-	-
Homo sapiens	585	568	568	512,298 [2]	-
Macaca mulatta	11	4	8	-	-
Manihot esculenta	1	-	1	34 [1]	-
Medicago truncatula	27	27	-	1,274 [1]	2
Mus musculus	517	516	516	-	-
Oryza sativa	39	23	72	2,121 [1]	8
Pan troglodytes	3	1	3	-	-
Setaria italica	2	1	2	61 [1]	0
Sus scrofa	5	5	-	-	-
Total	**1,407**	**1,163**	**1,263**	**516,182**	**14**

Since the first mirtron identification in *D. melanogaster* and *C. elegans* in 2007, this subject has been widely studied (Figure 5). More than half of publications (53.3%) occurred between 2011 and 2014. The highest number of publications per year were in 2012. The highest increase in available sequence data occurred in 2015, mainly due to study "*Analysis of Nearly One Thousand Mammalian Mirtrons Reveals Novel Features of Dicer Substrates*" (WEN et al., 2015).



	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
mirtron papers	4	8	12	18	28	39	49	58	65	71	76
Precursor	38	84	84	114	234	403	406	441	1,399	1,403	1,407
5' Mature	11	20	20	22	29	177	178	211	1,155	1,159	1,163
3' Mature	35	53	53	81	92	304	307	309	1,255	1,259	1,263

Figure 5. Cumulative distribution of mirtron papers, precursor and mature mirtron sequences per year. The mirtrons studies and sequence numbers presents growth during the period, mainly in 2015.

2.4.2. Precursor mirtron similarity analysis

We obtained 944 aligned precursor mitrons in the similarity analysis, where 896 were aligned in chordates (94.9%), 46 aligned in invertebrates (4.9%) and 2 in plants (0.2%) (Appendix 4). Only four species had more than 3 mirtrons aligned in another genome: *H. sapiens, M. mulatta, P. troglodytes* and *D. melanogaster*.

2.4.3. Mature mirtron characterization

We analyzed mature mirtron size distribution among species, which is detailed in Appendix 5 and 6. Almost three quarters (74.1%) of mature mirtrons are between 21 and 23 nt. In chordates and invertebrates, most mature mirtrons have 22 nucleotides (32.1%) and, in plants, most mature mirtrons (28%) have 21 nt sequences.

Regarding mature mirtron arms, 3' mature mirtrons have in most cases 21, 22 and 24 nucleotides in size, for chordates, invertebrates and plants respectively. For 5' mature mirtrons average size, chordates and invertebrates have 22 nucleotides and plants have 21 nucleotides.

We obtained logo sequences for mirtron arms (CROOKS et al., 2004) (Appendix 7). Chordates present more GC bases than invertebrates and plants.

2.4.4. Mirtrons availability in miRBase

We investigated if mature mirtron sequences were represented in the gold standard miRNA database, miRBase release 22 (Table 3). We observed that 966 mirtrons (39.8%) are available in miRBase, reinforcing the novelty provided by mirtronDB. Particularly, only 2% of plant mirtrons are available in miRBase. However, most invertebrate mirtrons (94.1%) are in miRBase.

Table 3. Mature mirtron availability in miRBase and data exclusively presented in mirtronDB

Organism Group	miRBase v22		mirtronDB	Exclusively available in mirtronDB	
	#	%	#	#	%
Chordate (*H. sapiens / M. musculus*)	874	40.3%	2,168	1,294	59.7%
Other chordates	25	49.0%	51	26	51.0%
Invertebrate	64	94.1%	68	4	5.9%
Plant	3	2.2%	139	136	97.8%
Total	**966**	**39.8%**	**2,426**	**1,460**	**60.2%**

2.4.5. Target gene analysis

We used TargetScan and psRNATarget to perform the target gene prediction in human and plant genomes, respectively. We identified a total of 512,298 and 3,884 potential targets in humans and plants, respectively (Table 2).

The largest potential interaction in human transcripts occurs with the transcript ENST00000609686.1, in which we identified 495 potential mature mirtron interactions. The mirtron that has more interactions with this transcript is "hsa-mirtron-1339-5p", with 6 potential interactions.

The human transcript ENST00000609686.1 is originated from gene GRIN2B (glutamate ionotropic receptor NMDA type subunit 2B). This gene is located in chromosome 12 (MISHRA et al., 2015) and it has 13 exons (LEMKE et al., 2013). The gene GRIN2B is normally associated with epilepsy and its mutations are

associated with neurodevelopmental disorders, such as intellectual disabilities (HU et al., 2016), schizophrenia (MISHRA et al., 2015) , and autism (PAN et al., 2015).

The mirtron "hsa-mirtron-1339-5p", which presents more potential interactions with the transcript ENST00000609686.1, is originated from gene RAPGEF1. Gene RAPGEF1 (Uniprot accession Q13905) plays an important role in regulating neural cell polarity in rats (SHAH et al., 2016) and cell adhesion, proliferation, apoptosis and actin reorganization in humans (MITRA et al., 2011). These interactions present an important potential mirtron role in chordate specialized cell development.

In plants, we identified 3,884 potential target interactions: 2,121 in *O. sativa*, 1,274 in *M. truncatula*, 394 in *A. thaliana*, 61 in *S. italica* and 34 in *M. truncatula* (Table 2). The *M. truncatula* mirtron, namely "mtr-mirtron-1838-5p", is the one with the highest amount of interactions in plants (78). This mirtron is originated from gene Medtr5g023640.1, an unknown protein.

The plant transcript with more mirtron target interactions is OS03T0348800-01 (14 interactions), in *O. sativa.* The *O. sativa* transcript OS03T0348800-01 is originated from gene Os03g0348800, which codes a triacylglycerol lipase (KIKUCHI et al., 2003).

2.4.6. ceRNA and mirtrons in plants

In plants, we verified if mirtrons could act as ceRNA candidates (Appendix 8). A total of 1,738 potential ceRNA interactions were identified in plant mirtrons using TAPIR. *O. sativa* contains 82.8% of the total interactions (1,439 results), followed by *A.thaliana* (210), *M.truncatula* (84), *S. italica* (4) *and M. esculenta* (1). However, most of these interactions did not represent a ceRNA interaction in the same species. We identified 4 mirtrons from *A. thaliana*, 2 from *M. truncatula* and 8 from *O. sativa* that can act as ceRNAs in targets within the same species.

2.4.7. mirtronDB: the repository

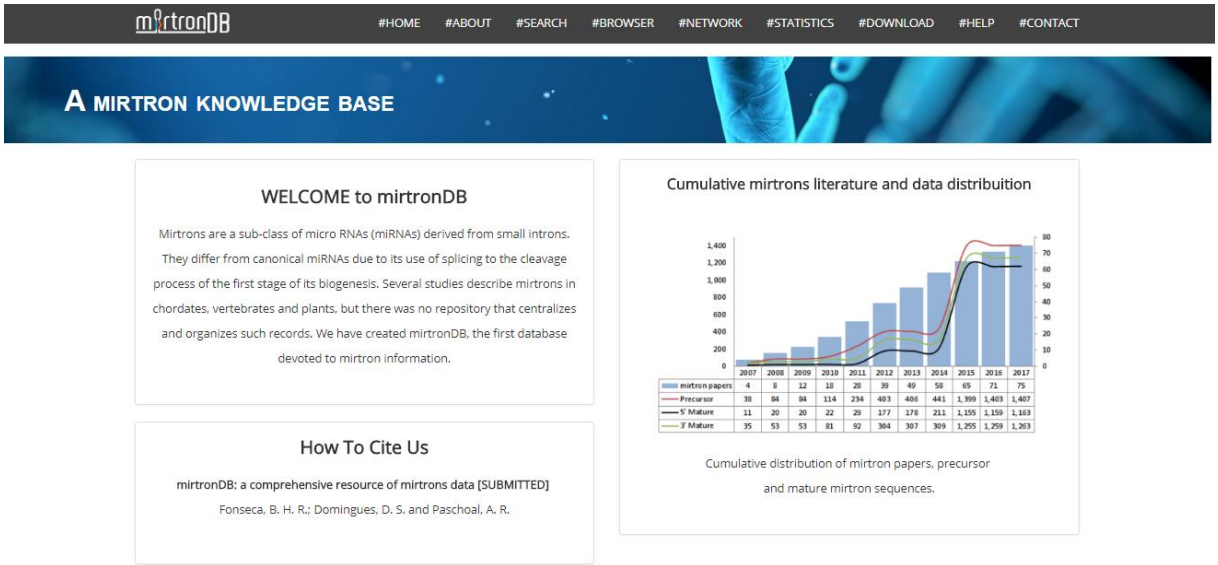
The mirtronDB portal provides a user-friendly web interface to search, browse, network visualization, statistics, and download all curated information on mirtrons (Figure 6).

Users can search mirtrons by organism, organism group, mirtron type and mirtron name in the search page. A table with summarized results is presented with two options: download result table and details page. In the first one, the user can download the results information in a table format. In the second (“Details” page) users have access to all the information on that mirtron, such as identification code, mirtron name, origin organism, mirtron sequence, host gene, and publication data.

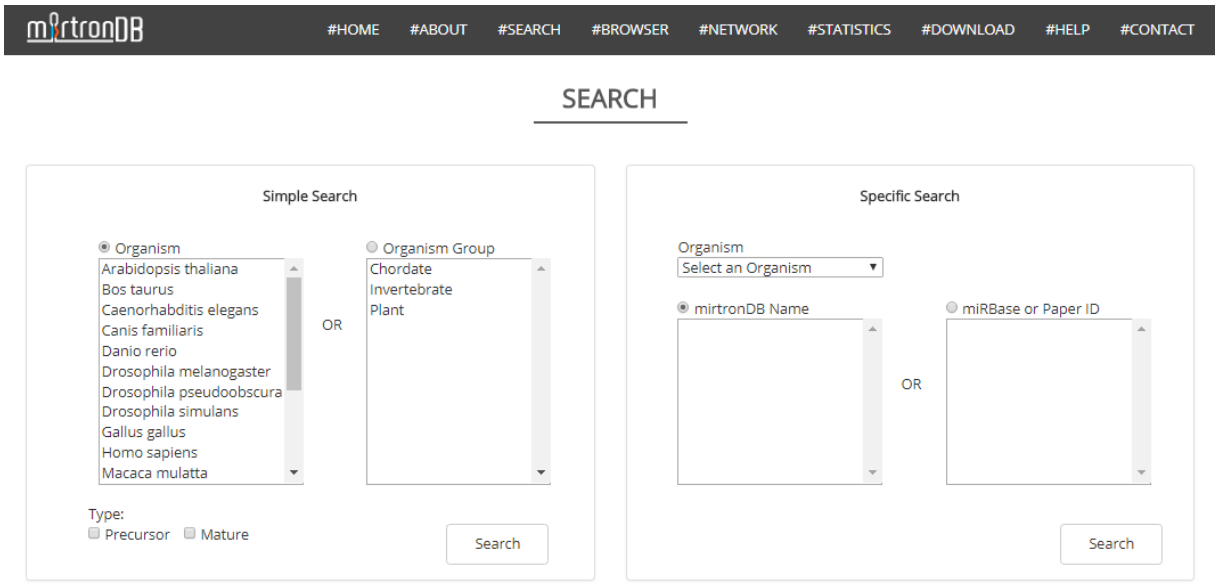
In the "Browse" section, users can find mirtron data by organism. The “Network Visualization” page is dedicated to mirtron regulatory interaction in a graphical view. Interactions are predicted between target gene and ceRNA prediction. In "Statistics", we present the total number of available organisms, collected mirtrons, mirtrons by organism, and papers used for data analysis. All mirtron sequence data in FASTA and GFF format is available for the scientific community in the “Download” page. Finally, “Help” shows users how to use mirtronDB.

Figure 6. Search methods and results from mirtronDB

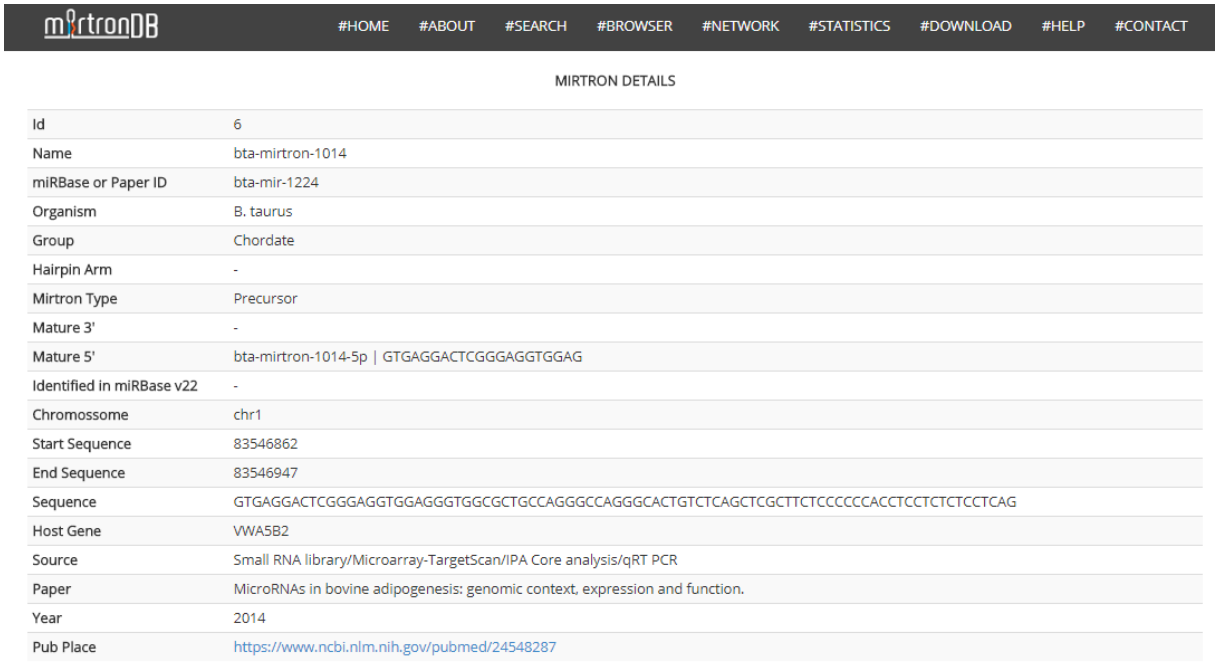
(A) General webpages



(B) Simple search and available filters

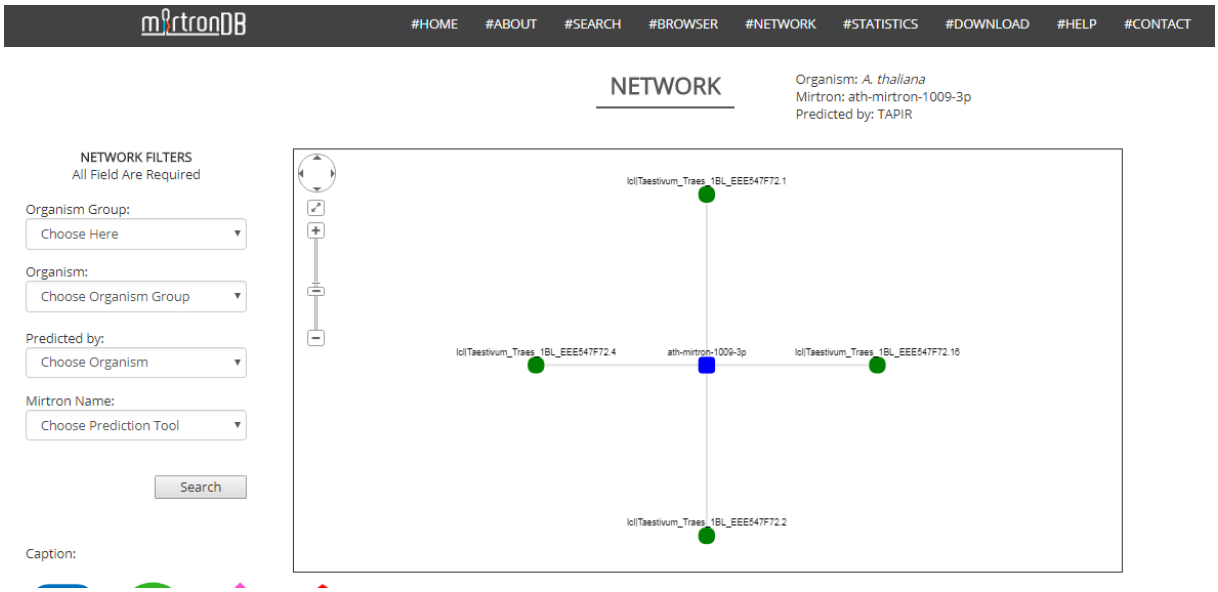


(C) Detailed search results for a specific mirtron



Id	6
Name	bta-mirtron-1014
miRBase or Paper ID	bta-mir-1224
Organism	B. taurus
Group	Chordate
Hairpin Arm	-
Mirtron Type	Precursor
Mature 3'	-
Mature 5'	bta-mirtron-1014-5p | GTGAGGACTCGGGAGGTGGAG
Identified in miRBase v22	-
Chromossome	chr1
Start Sequence	83546862
End Sequence	83546947
Sequence	GTGAGGACTCGGGAGGTGGAGGGTGGCGCTGCCAGGGCCAGGGCACTGTCTCAGCTCGCTTCTCCCCCCACCTCCTCTCTCCTCAG
Host Gene	VWA5B2
Source	Small RNA library/Microarray-TargetScan/IPA Core analysis/qRT PCR
Paper	MicroRNAs in bovine adipogenesis: genomic context, expression and function.
Year	2014
Pub Place	https://www.ncbi.nlm.nih.gov/pubmed/24548287

(D) Results of a mirtron interaction network



2.5. DISCUSSION

Several mirtrons studies have been published, but all data generated by them is available in distinct and non-standardized structures, without an integrated web repository. Integrating mirtron data provides an initial assessment for large-scale mirtron characterization and analysis, even comparatively.

MirtronDB is a web database that centralizes, standardizes, integrates and provides mirtron data available in literature. We highlight that (i) all data collected is available in several formats (e.g., FASTA, GFF, CSV); (ii) curated data make this repository a reference to explore mirtron information; (iii) mirtron sequence, structure and conservation analysis are provided; and (iv) potential targets and occurrence of ceRNA in mirtrons are also investigated.

In our mirtron data analysis, mammals were the most represented organisms in available data, reflecting that most studies on mirtrons were focused on *H. sapiens* and *M. musculus*. Consequently, we identified more similarity results among chordates than in the other groups. We also highlight that chordates have mature sequences with the highest GC content and, consequently, greater molecule stability. In plants, they presented few available precursor mirtrons, and in this group we can notice that average mature mirtron size and other characteristics are dissimilar from mammals.

Using mature mirtron size characterization by organism group, we could identify that, in most cases, mature mirtrons in chordates and invertebrates present 22 nucleotides in size, and in plants, 21 nucleotides. This is an example of information that was not easily available, as mirtron data distribution was not centralized in a repository. When characterizing mirtrons acting as targets, we could identify mirtrons presenting potential to play regulatory roles in mRNA and act as ceRNAs in plants such as *A. thaliana*, *M. truncatula* and *O. sativa*.

Mirtron data availability facilitates the development of new studies in biology. For example, information resulting from mirtron similarity analysis and target gene prediction could then be tested in wet lab analyses.

MirtronDB data could also provide novel standardized approaches to predict and identify mirtrons in organisms that do not have them described yet.

2.6. CONCLUSION

MirtronDB is a comprehensive and unique database about mirtrons. Its resources allow users to query mirtron data and download them in several formats. The analyses presented in this paper provide initial mirtron characterization as well, and can be used as a guide about mirtrons potential as ceRNAs and gene expression regulators. This repository also has the potential to promote advances in computational biology, since consolidated data available can now be used to improve or model mirtron approaches and biological experiments.

3. ANÁLISE EXPLORATÓRIA COMPARATIVA: miRNAs E MIRTRONS

3.1. INTRODUÇÃO

MicroRNAs (miRNAs) são pequenos RNAs não-codificantes com tamanho de 18 a 25 nucleotídeos (LIAO et al., 2018) e potencial de atuação regulatória pós-transcricional nos níveis de RNA mensageiro na célula (BUDAK; AKPINAR, 2015). Sua desregulação em humanos é associada a doenças como o câncer (LAURYNAS et al., 2016), e em plantas, relacionados ao tempo de floração e desenvolvimento de raízes (FERDOUS; HUSSAIN; SHI, 2015).

Os mirtrons são uma subclasse de miRNAs de via biogênica alternativa (WESTHOLM; LAI, 2011). Eles são pequenos introns que utilizam-se do processo de *splicing* para não passar pela primeira etapa da clivagem do transcrito primário, realizada pela *Drosha*. Posteriormente estes introns seguem o processo biogênico canônico (RUBY; JAN; BARTEL, 2007). Mirtrons também desempenham importantes papéis biológicos, tais como atuar como potencial silenciador do gene origem da doença de Parkinson (SIBLEY et al., 2012) e regulatoriamente no processo de fotossíntese em plantas (MENG; SHAO, 2012).

Dado os importantes papéis biológicos que miRNAs e mirtrons desempenham e sua distinção biogênica, pesquisas têm sido realizadas buscando identificar características capazes de diferenciá-los e predize-los. Estudos como "*Distinguishing mirtrons from canonical miRNAs with data exploration and machine learning methods*" (RORBACH; UNOLD; KONOPKA, 2018) e "*Comparing miRNA structure of mirtrons and non-mirtrons*" (TITOV; VOROZHEYKIN, 2018) buscam compará-los e distingui-los, porém para tal utilizam pequeno e restrito conjunto de dados, limitando suas análises ao conjunto geral de características, sem a distinção de miRNAs e mirtrons entre grupos de organismos.

Este capítulo tem a finalidade de identificar características capazes de distinguir miRNAs e mirtrons, sendo utilizado dados de sequências para tal análise exploratória comparativa descritiva. As análises levam em consideração o tipo da estrutura (precursor e maduro) e a distinção por grupo de organismo (cordados, invertebrados e plantas).

A abordagem comparativa apresentada neste trabalho é inédita, leva em consideração seis características (de sequência e estrutura), e dada a quantidade, organização e estruturação dos dados utilizados, é possível afirmar que apresenta ganhos em assertividade estatística nos resultados das análises, em relação a outros comparativos já realizados, e proporciona uma nova concepção sobre as relações e caracterizações de miRNAs e mirtrons.

3.2. MATERIAIS E MÉTODOS

3.2.1. Coleta de dados

As sequências de miRNAs e mirtrons foram coletadas, em estrutura precursora e madura, dos repositórios miRBase v22 (KOZOMARA; GRIFFITHS-JONES, 2013) e mirtronDB, respectivamente.

O miRBase, considerado estado-da-arte com relação a banco de dados de miRNAs, possui em sua versão 22 o total 38.589 sequências de miRNAs precursores e 48.885 sequências de miRNAS maduros. O mirtronDB possui 1.407 mirtrons precursores e 2.426 sequências de mirtrons maduros.

Os dados foram divididos em 3 grupos: cordados, invertebrados e plantas. Para miRNAs os organismos definidos para representar os grupos foram *H. sapiens*, *D. melanogaster* e *A. thaliana*, respectivamente. Para mirtrons, devido à quantidade distinta de dados, os grupos de organismos foram compostos conforme apresentado a seguir: (i) Cordados: *B. taurus, C. familiaris, D.rerio, G. gallus, H. sapiens, M. mulatta, M. musculus, P. troglodytes* e *S. scrofa;* (ii) Invertebrados: *C. elegans, D. melanogaster, D. pseudoobscura* e *D. simulans;* (iii) Plantas: *A. thaliana, M. esculenta, M. truncatula, O. sativa e S. italica.*

3.2.2. Análise de conjunto de dados

Visando a identificação e remoção de duplicidade nos dados, uma vez que existem mirtrons disponíveis no miRBAse, utilizou-se a ferramenta de alinhamento local CD-HIT-2D (LI; GODZIK, 2006), com cut-off de 0,98 de identidade. Do total de 6.074 sequências de miRNAs coletados, 860 foram identificadas como mirtrons. A

Figura 7 apresenta o Diagrama de Venn com a interseção de miRNAs e mirtrons, precursores e maduros.

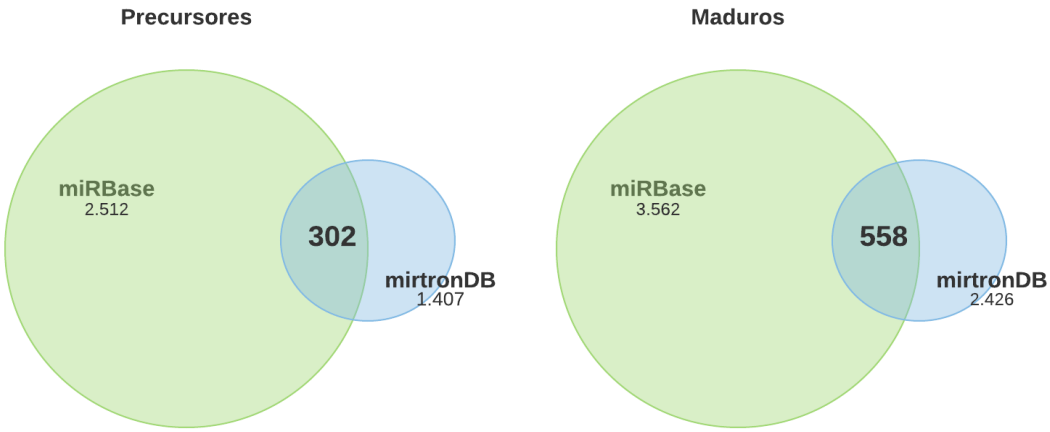


Figura 7. Remoção de duplicidade de dados de miRNAs e mirtrons

Fonte: Autoria Própria

A Figura 8 apresenta a quantidade de registros utilizados nas análises deste trabalho, após remoção de duplicidade de dados, por grupo de organismo e tipo de sequência (precursor e maduro). No total foram utilizados 5.214 registros de miRNAs e 3.833 de mirtrons.

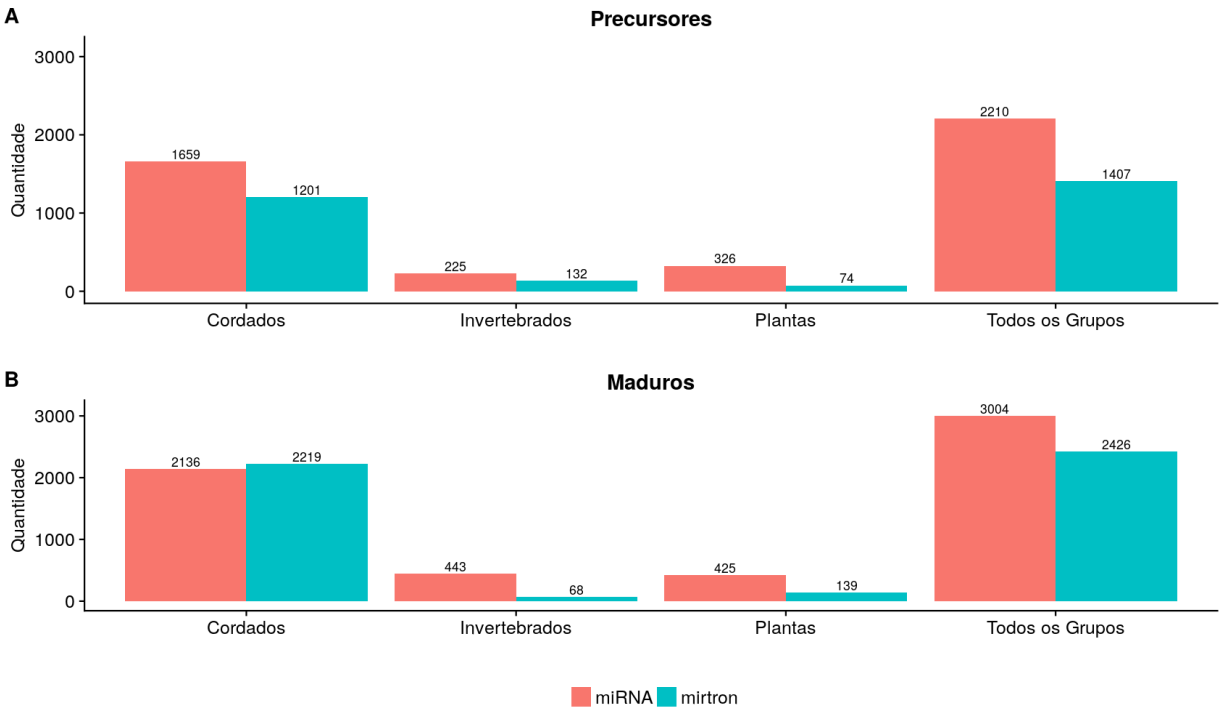


Figura 8. Quantidade de registros analisados, por grupo de organismo e tipo de sequência
Fonte: Autoria Própria

3.2.3. Seleção de características analisadas

As características analisadas neste trabalho são relacionadas a sequência e estrutura de miRNAs e mirtrons, e os resultados apresentam distinção por grupo de organismo e tipo de estrutura (maduro e precursor), sendo:

- Para maduros e precursores:
  - (i) Distribuição de tamanho de sequência e (ii) Relação de conteúdo GC, conforme utilizado em "*Distinguishing mirtrons from canonical miRNAs with data exploration and machine learning methods*" (RORBACH; UNOLD; KONOPKA, 2018);
  - (iii) Relação de conteúdo de G em função de C (*GC ratio*), conforme utilizado em "*MiRNAfe: A comprehensive tool for feature extraction in microRNA prediction*" (YONES et al., 2015);
  - (iv) Distribuição de nucleotídeos (k-mer de 1 a 3), conforme utilizado em *"MicroRNA categorization using sequence motifs and k-mers"* (YOUSEF et al., 2017) ;

- Para precursores:

(i) Frequência de energia mínima livre (MFE), conforme utilizado em "*MiRNAfe: A comprehensive tool for feature extraction in microRNA prediction*" (YONES et al., 2015);

- Para maduros:

(i) Distribuição de frequência de bases, conforme utilizado em "*MiRNAfe: A comprehensive tool for feature extraction in microRNA prediction*" (YONES et al., 2015);

3.2.4. Distribuição de tamanho de sequências

Os tamanhos das sequências também foram analisados. Para tanto, um script em linguagem R foi implementado para o cálculo de tamanho das sequências de miRNAs e mirtrons, tanto precursores, quanto maduros.

3.2.5. Frequência de bases para miRNAs e mirtrons maduros

A representação gráfica da frequência de nucleotídeos (nt) para miRNAs e mirtrons maduros foi realizada através da ferramenta WebLogo (CROOKS et al., 2004). A representação foi realizada apenas para as sequências que apresentaram maior frequência de ocorrência, com relação a tamanho, por grupo de organismos.

3.2.6. Relação de conteúdo GC

Com base no script elaborado em R para a contagem de nucleotídeos por sequência, a relação proporcional de conteúdo das bases Guanina (G) e Citosina (C) também foi calculada. O cálculo foi realizada com base na representatividade da soma da quantidade das bases G e C em função do tamanho total da sequência, soma de Adeninas (A), Citosinas (C), Timinas (T) e Guaninas (G):

$$Conteúdo\ GC\ = \frac{G\ +\ C}{A\ +\ C\ +\ T\ +\ G}$$

3.2.7. Relação de Guaninas em função de Citosinas (*GC Ratio*)

O cálculo da relação entre Guanina (G) e Citosina (C) nas sequências precursoras e maduras também foi realizada através da linguagem R, uma vez que a quantidade de bases já haviam sido calculadas. Este indicador de análise estrutural de sequências também é conhecido como *GC ratio*.

$$GC\ ratio\ = \frac{G}{C}$$

O indicador GC ratio é utilizado para representar o potencial de estabilidade térmica da molécula. Quanto maior seu valor, mais estável a molécula tende a ser.

3.2.8. Distribuição de nucleotídeos

Com utilização da linguagem R e apoio da biblioteca “seqinr” (CHARIF; LOBRY, 2007), a qual é desenvolvida especificamente para análise e visualização de dados biológicos, um script para cálculo de frequência de mononucleotídeos, dinucleotídeos e trinucleotídeos foi realizada. As frequências foram calculadas para maduros e precursores, de miRNAs e mirtrons e os resultados são graficamente apresentados com base na ordem alfabética das da primeira letra do nome das bases e suas combinações.

3.2.9. Frequência de energia mínima livre (MFE)

A estabilidade da estrutura secundária de miRNA e mirtrons precursores pode ser quantificada com base no potencial de energia livre liberada em sua formação. A energia mínima livre demonstra a capacidade de dobramento da sequência precursora e seu potencial de formação de grampos, sendo que quanto menor for seu valor, mais estável é a estrutura secundária (SOEMEDI et al., 2017).

Para cálculo da frequência de Energia Mínima Livre (*Minimum Free Energy ou* MFE), de miRNAs e mirtrons precursores, foi utilizada a aplicação RNAfold, a qual é disponibilizada através do pacote ViennaRNA 2.0 (LORENZ et al., 2011).

3.3. RESULTADOS

A análise exploratória descritiva foi realizada de forma comparativa para precursores e maduros de miRNAs e mirtrons. Considerou-se para análise os grupos: cordados, invertebrados, plantas e acumulado dos grupos.

3.3.1. Distribuição de tamanho de sequências

Para precursores, comparativamente entre os grupos, nota-se que há variação de tamanho, principalmente em plantas e cordados. Por exemplo, em cordados a amplitude da distribuição e o tamanho das sequências foram maiores em mirtrons do que em miRNAs, fato oposto ao observado no grupo de plantas. Invertebrados possuem comportamento de distribuição de tamanho relativamente similar entre miRNAs e mirtrons (Figura 9).

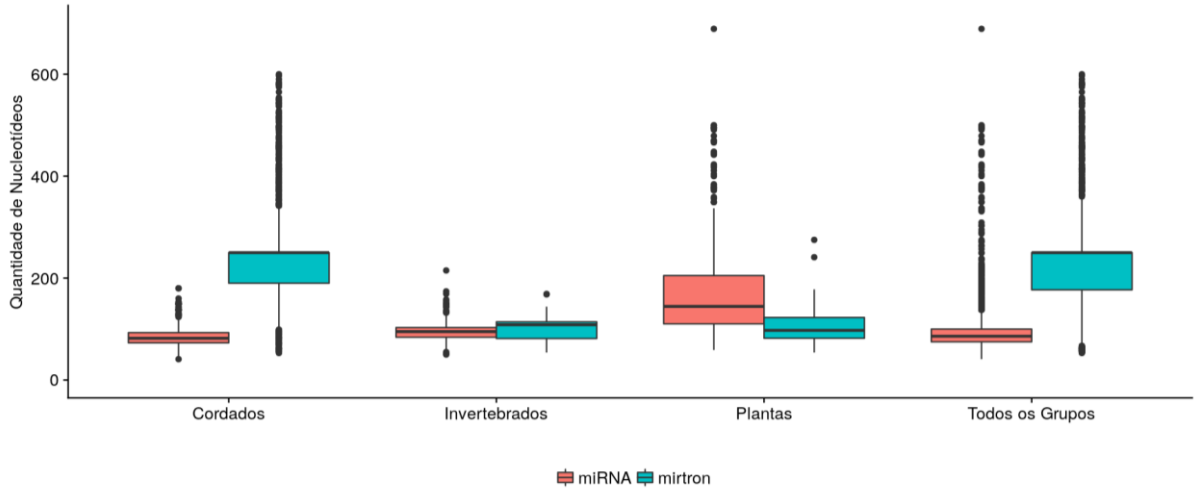


Figura 9. Distribuição de tamanho de sequências precursoras
Fonte: Autoria Própria

Em geral, as sequências precursoras de mirtrons apresentam maior amplitude de tamanho entre si, fato não observado em miRNAs. Enquanto mirtrons possuem tamanho de precursor médio de 220 nucleotídeos, os miRNAs apresentam estrutura de precursor com 100 nucleotídeos de tamanho médio.

Os mirtrons precursores de cordados são os que apresentam o maior tamanho médio, seguido por invertebrados e plantas. Para miRNAs precursores, plantas apresentam maior tamanho médio, seguidos de invertebrados e cordados.

Quanto a distribuição de tamanho das sequências de miRNAs e mirtrons maduros destaca-se que embora no geral existam similaridades, por grupo de organismo há distinção. Observa-se que acumulado dos grupos, tanto miRNAs quanto mirtrons, apresentam maior quantidade de maduros com tamanho médio de 22 nucleotídeos (Figura 10).

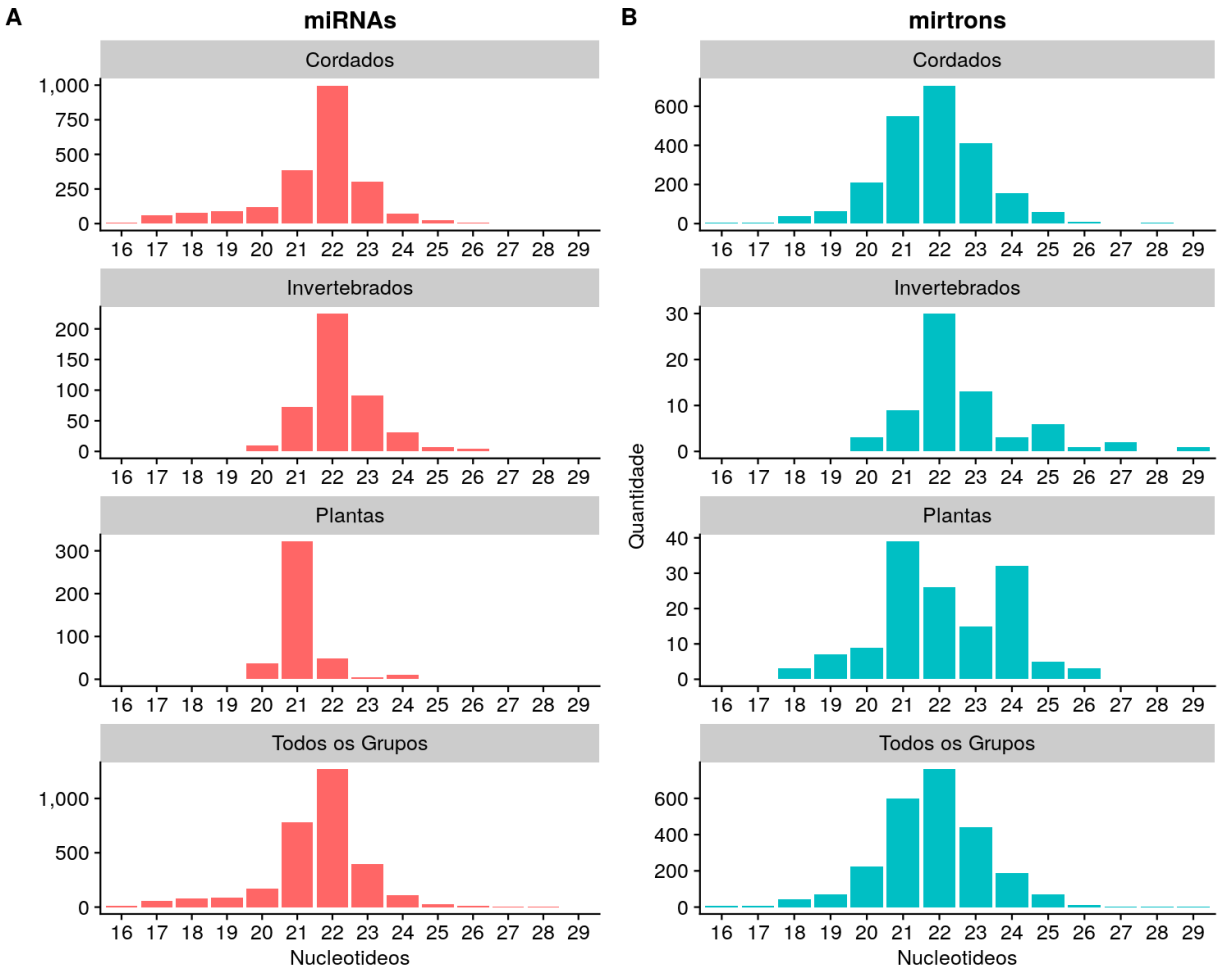


Figura 10. Distribuição de tamanho de maduros

Fonte: Autoria Própria

Especificamente para maduros por grupos de organismo, cordados e plantas apresentam distribuições distintas de tamanho. Embora miRNAs e mirtrons de cordados apresentem maioria de maduros com 22 nucleotídeos, em mirtrons há maior representatividade de maduros com tamanho de 21 e 23 nucleotídeos, do que em miRNAs. Para maduros de plantas, enquanto miRNAs apresentam sequências com 21 nucleotídeos de tamanho, em mirtrons além de sequências com 21

nucleotídeos, destacam-se expressivamente sequências com 24 e 22 nucleotídeos de tamanho.

3.3.2. Frequência de bases em maduros

A Figura 11 apresenta a frequência de distribuição de bases por posição para as sequências de maduros de maior representatividade de tamanho, por grupo de organismo.

Mirtrons	miRNAs
Cordados 	

Invertebrados 	

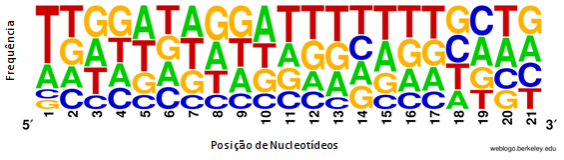
Plantas 	

Figura 11. Frequência de bases por posição

Fonte: Autoria Própria

Nota-se que embora por tamanho de sequência os miRNAs e mirtrons maduros possuam similaridade de características, visualmente sua composição

apresenta considerável distinção. Por exemplo, para cordados de 22 nucleotídeos nota-se a predominância de Guaninas em mirtrons e Timinas em miRNAs. Para invertebrados de 22 nucleotídeos pode-se destacar que enquanto miRNAs possuem quantidade expressiva de Timinas, para mirtrons apenas entre as posições 15 a 19 o mesmo ocorre. Para plantas de 21 nucleotídeos, a variação da distribuição das bases é expressiva, levando em consideração a região *seed* como exemplo (nucleotídeos 2 a 8).

3.3.3. Relação de conteúdo GC

Para precursores, com exceção dos invertebrados, os grupos de organismo apresentam conteúdo GC superior em mirtrons do que em miRNAs (Figura 12.A).

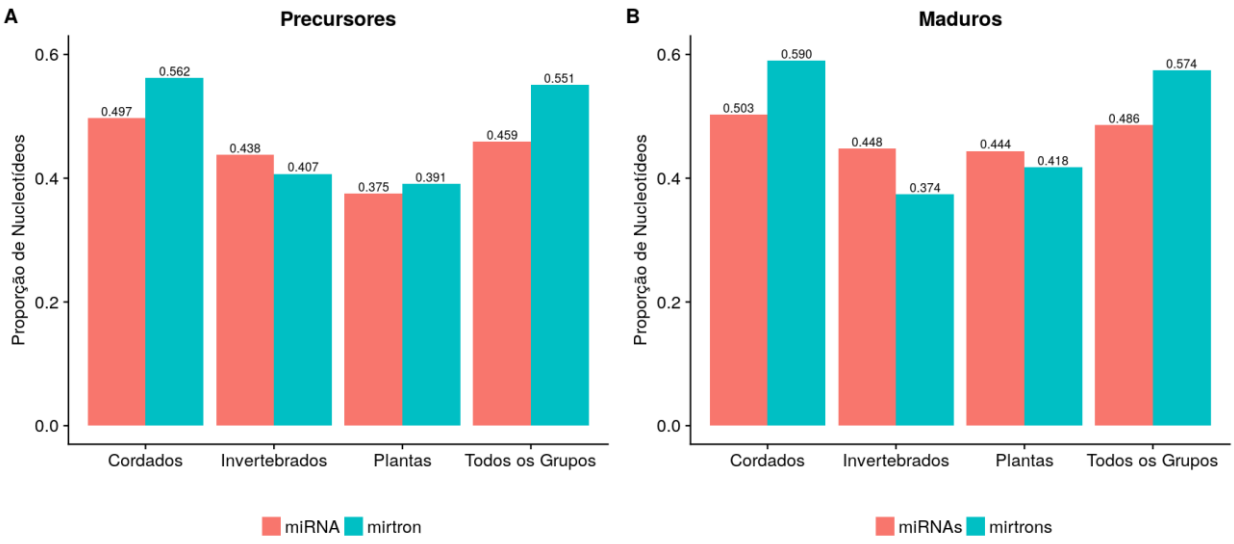


Figura 12. Proporção de conteúdo GC

Fonte: Autoria Própria

Com relação aos maduros, de acordo com a Figura 12.B, é possível identificar que no geral mirtrons possuem maior concentração de conteúdo GC, se comparados a miRNAs, embora nos grupos de invertebrados e plantas a relação não seja evidenciada.

3.3.4. Relação de *GC ratio*

Para os precursores, a relação entre a ocorrência de guaninas (G) em função de citocinas (C) apresenta no geral caracterização similar, embora para plantas exista maior diferenciação entre mirtrons e miRNAs (Figura 13. A).

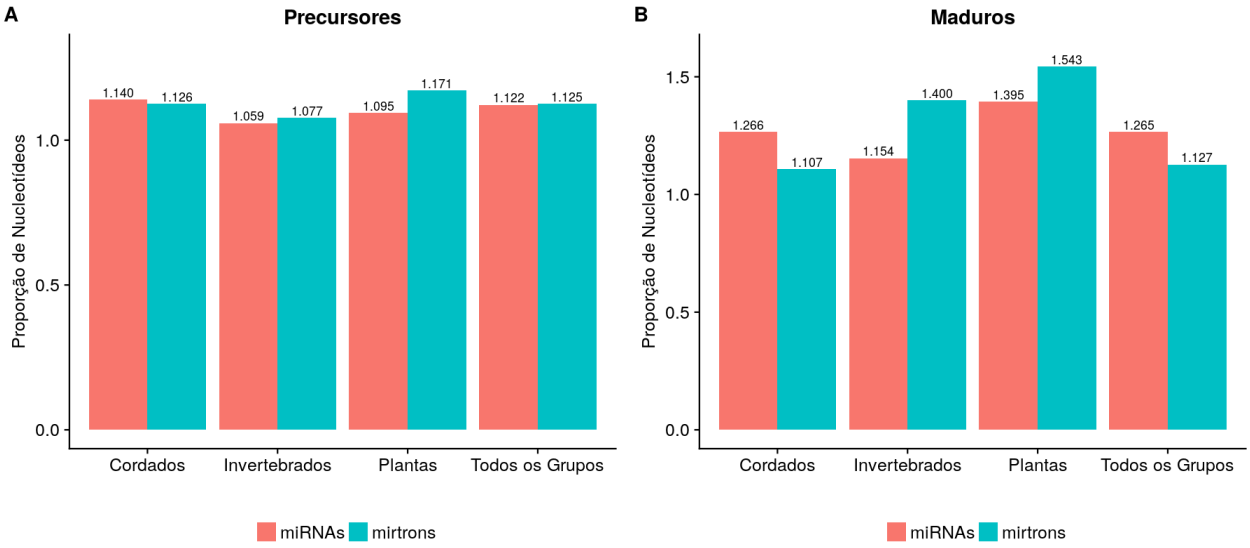


Figura 13. Proporção de *GC ratio*

Fonte: Autoria Própria

Considerando maduros, o GC *ratio* varia entre os grupos de organismos. Em plantas, tanto para miRNAs, quanto para mirtrons, há a maior proporção de *GC ratio* entre os grupos de organismos. Considerando o mesmo grupo, miRNAs maduros de cordados possuem índice superior de *GC ratio* do que mirtrons, enquanto que para invertebrados os mirtrons são os que possuem maior representatividade (Figura 13.B).

3.3.5. Distribuição de nucleotídeos

3.3.5.1. Distribuição de mononucleotídeos

Com relação a grupo de organismos, a distribuição da proporção de mononucleotídeos entre miRNAs e mirtrons precursores apresentam maior similaridade entre invertebrados e plantas, do que em cordados. Neste último grupo, destaca-se a maior representatividade das bases G e C em mirtrons, do que em

miRNA (Figura 14). Nota-se também no total dos grupos a inversão de distribuição proporcional das bases, uma vez que enquanto miRNAs precursores apresentam maior conteúdo de T e A, em mirtrons tem-se G e C.

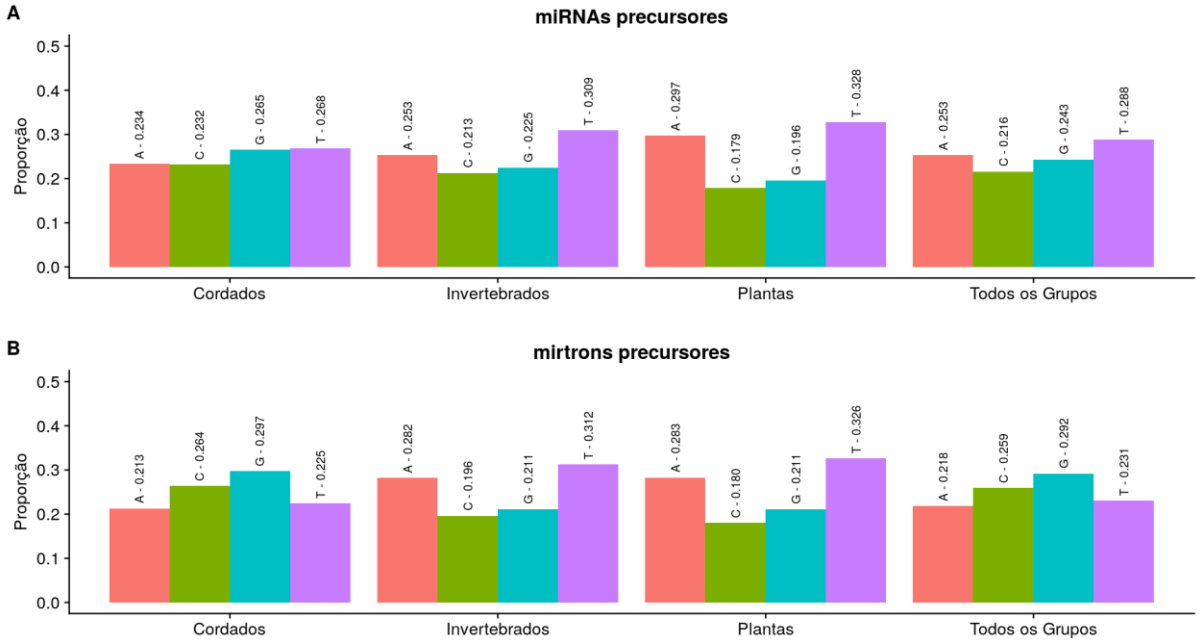


Figura 14. Mononucleotídeos por grupo de organismo, para precursores
Fonte: Autoria Própria

Através da Figura 15 é possível identificar que plantas e invertebrados são os grupos de organismos que apresentam maior proporção das bases mais representativas de miRNAs precursores (T e A). Já para mirtrons precursores, G e C apresentam expressiva maioria proporcionalmente em cordados, do que nos demais grupos.

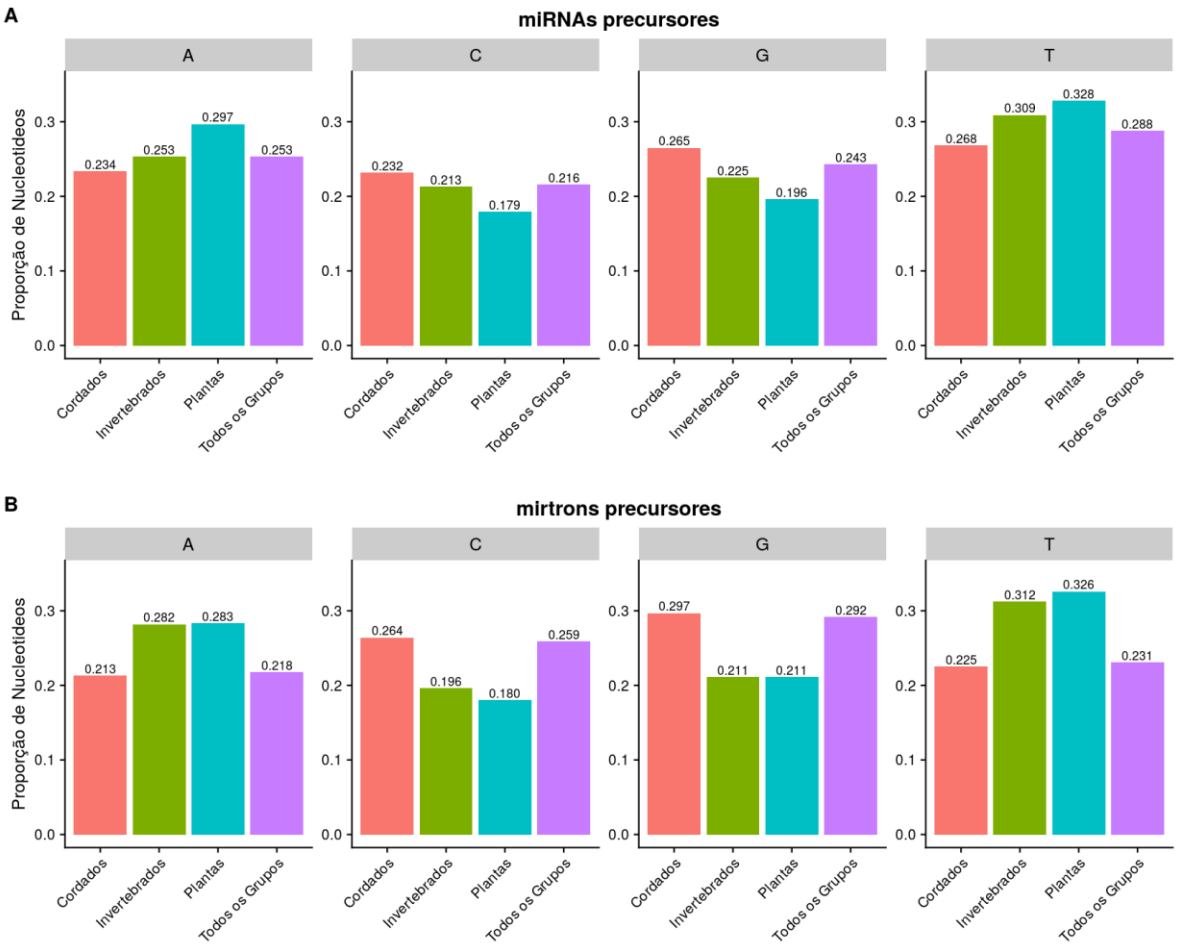


Figura 15. Proporção de mononucleotídeos por grupo de organismo, para precursores

Fonte: Autoria Própria

A distribuição de mononucleotídeos entre miRNAs e mirtrons maduros é apresentada através da Figura 16. Tanto mirtrons, quanto miRNAs em cordados, apresentam G como maioria de sua composição. Para invertebrados e plantas, a base de maior frequência é a T. Com relação à base de menor frequência, tem-se a C para todos os grupos de miRNAS e para a maioria dos mirtrons, sendo exceção os mirtrons em cordados. No total dos grupos a base de maior representatividade é a T para miRNAs, e a G para mirtrons.

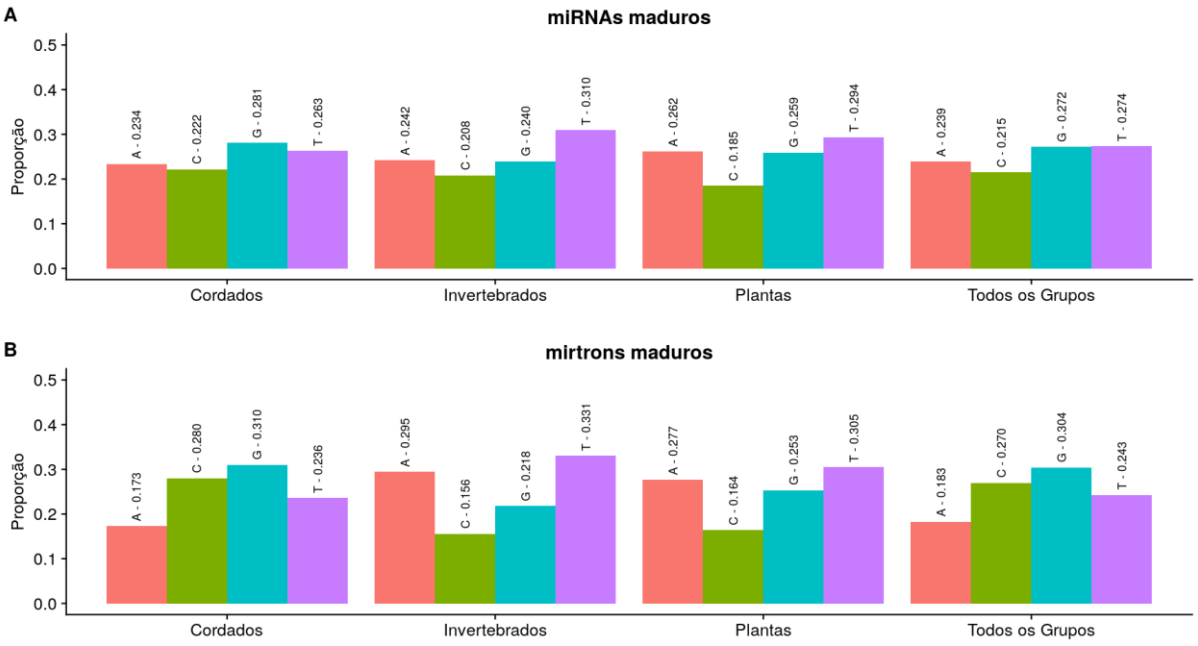


Figura 16. Mononucleotídeos por grupo de organismo, para maduros

Fonte: Autoria Própria

Há que se destacar ainda que invertebrados apresentam maior representatividade de T em miRNAs, e cordados são o grupo de organismo que apresentam maior representatividade de G em mirtrons (Figura 17).

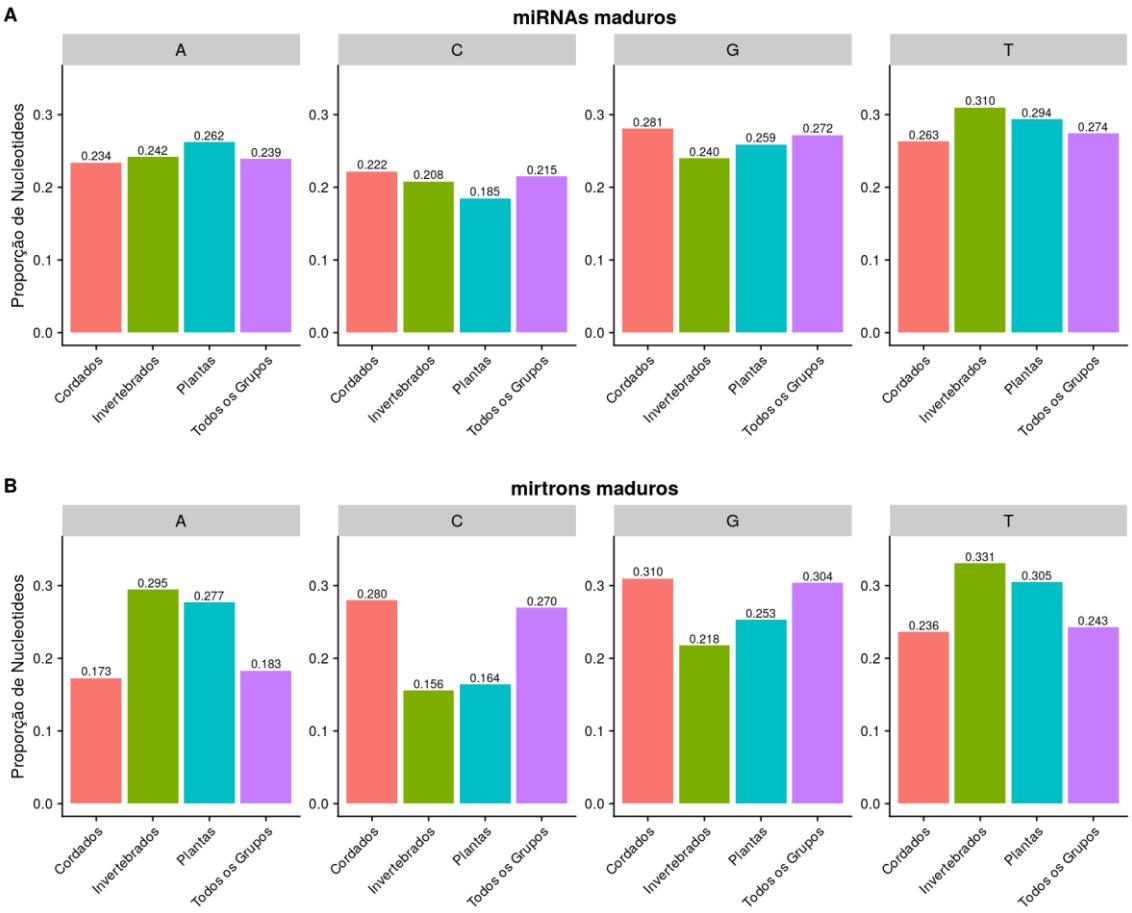


Figura 17. Proporção de bases nitrogenadas

Fonte: Autoria Própria

3.3.5.2. Distribuição de dinucleotídeos

A Figura 18 apresenta a distribuição de dinucleotídeos para miRNAs e mirtrons precursores:

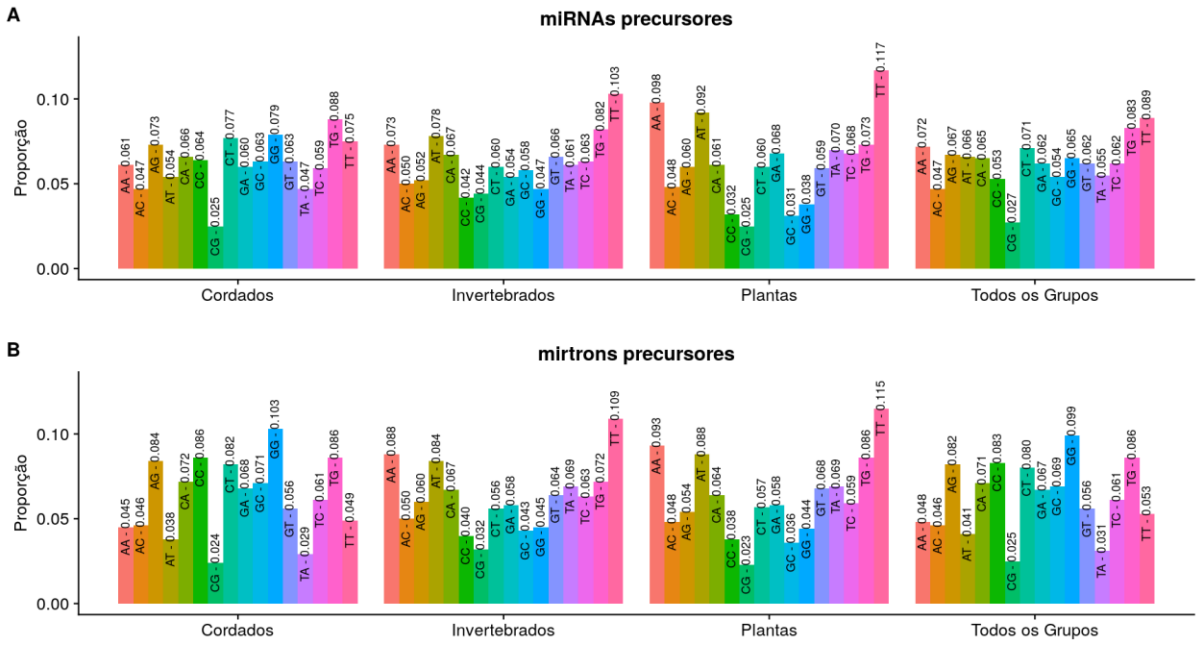


Figura 18. Distribuição de dinucleotídeos de miRNAs e mirtrons precursores, por grupo de organismos

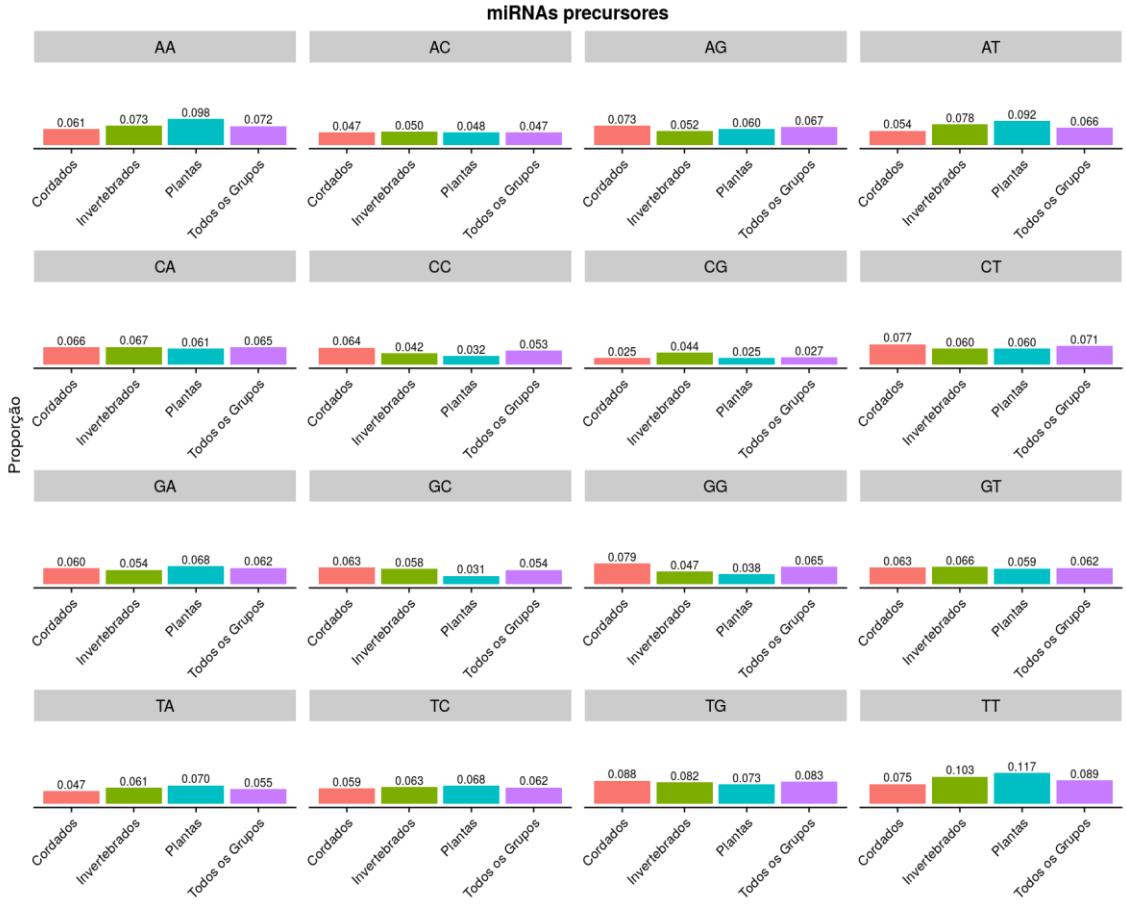
Fonte: Autoria Própria

Por grupo de organismo nota-se que mirtrons precursores de cordados apresentam maioria de dinucleotídeo GG, enquanto invertebrados e plantas possuem TT. Com relação a miRNAs, cordados apresentam maior frequência de TG, e invertebrados e plantas de TT. O dinucleotídeo CG apresenta menor expressividade para a maioria dos grupos de organismo, sendo exceção em miRNAs de invertebrados.

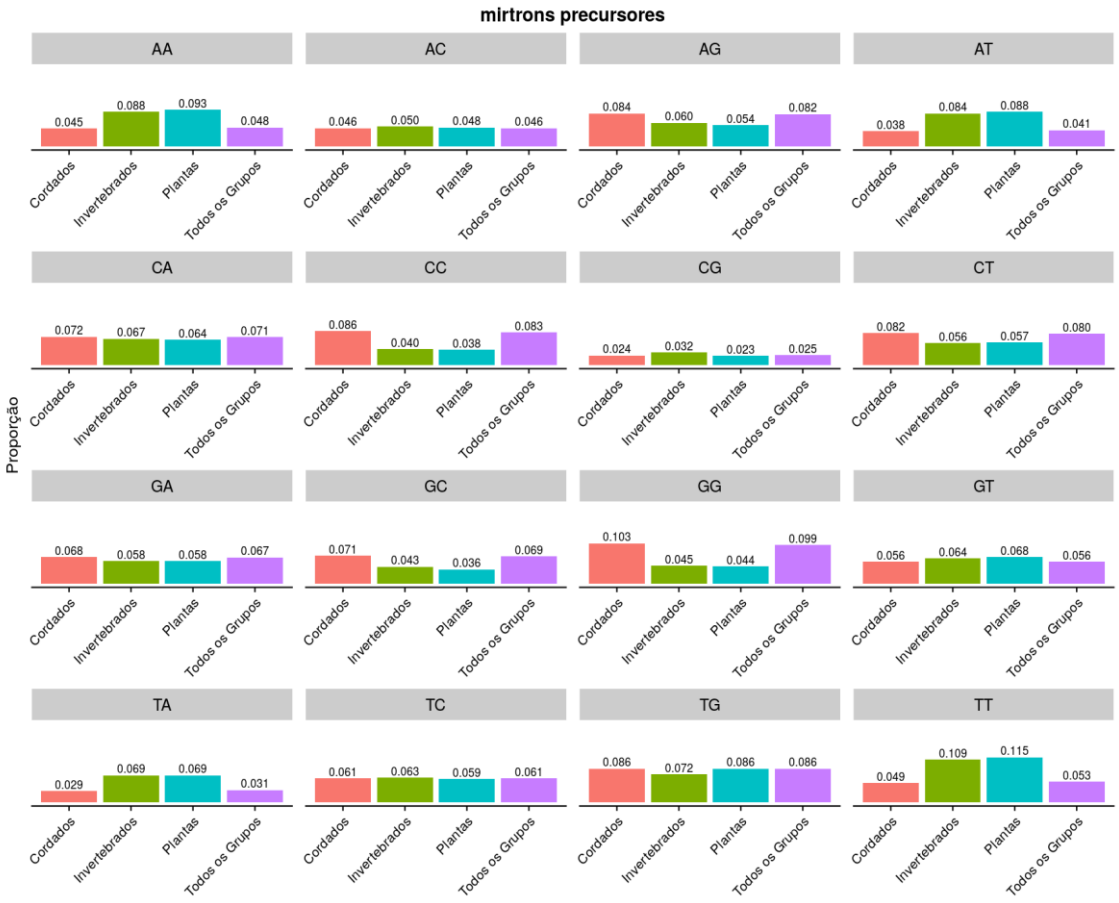
No geral, pode-se destacar distinções de miRNAs e mirtrons precursores com relação a distribuição dos dinucleotídeos AA, AT, GG, TA e TT. Para estes dinucleotídeos, plantas não apresentam a maior representatividade apenas para GG, o qual é apresentado em maioria em cordados (Figura 19).

Figura 19. Distribuição de dinucleotídeos de mirtrons e miRNAs precursores

A.



B.



Fonte: Autoria Própria

Em linhas gerais, com relação a distribuição de dinucleotídeos de miRNAs e mirtrons maduros tem-se a destacar que em cordados há variação principalmente da distribuição dos dinucleotídeos AA, AT, CC, GG e TT; em invertebrados as variações ocorrem principalmente em AA, CG, GC e GG; e para plantas, enquanto miRNAs apresentam maioria de TG, em mirtrons a maior ocorrência é de AT (Figura 20). No geral, a distribuição de dinucleotídeos apresenta grande variação em AA, AT, CC, GG e TT.

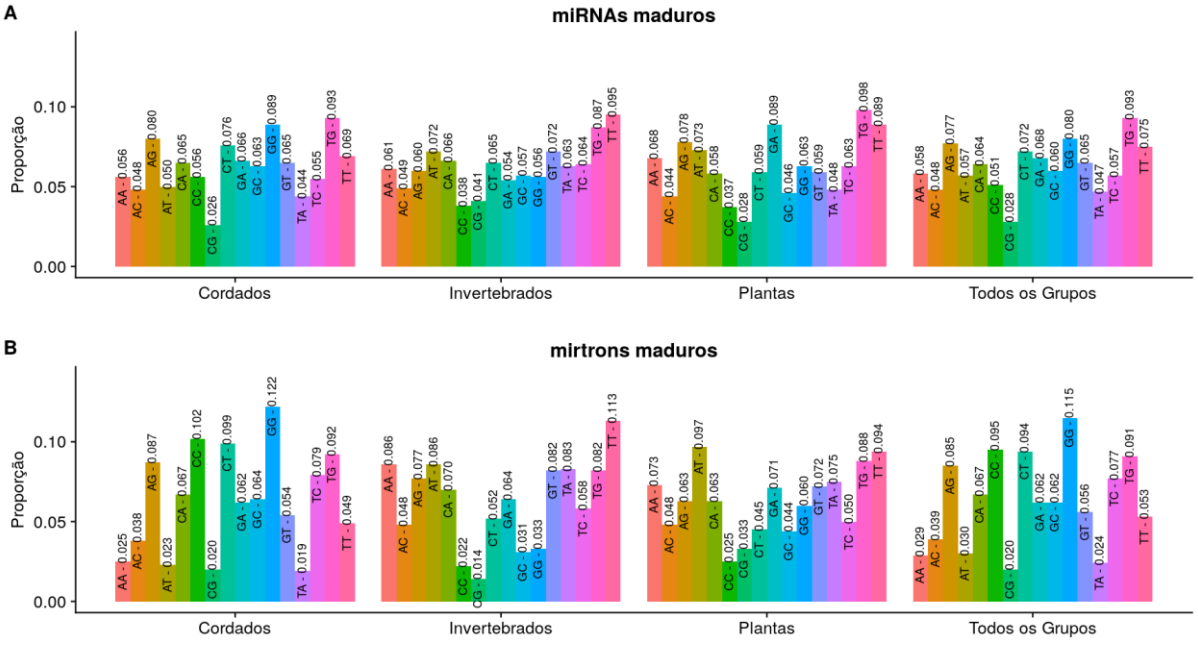


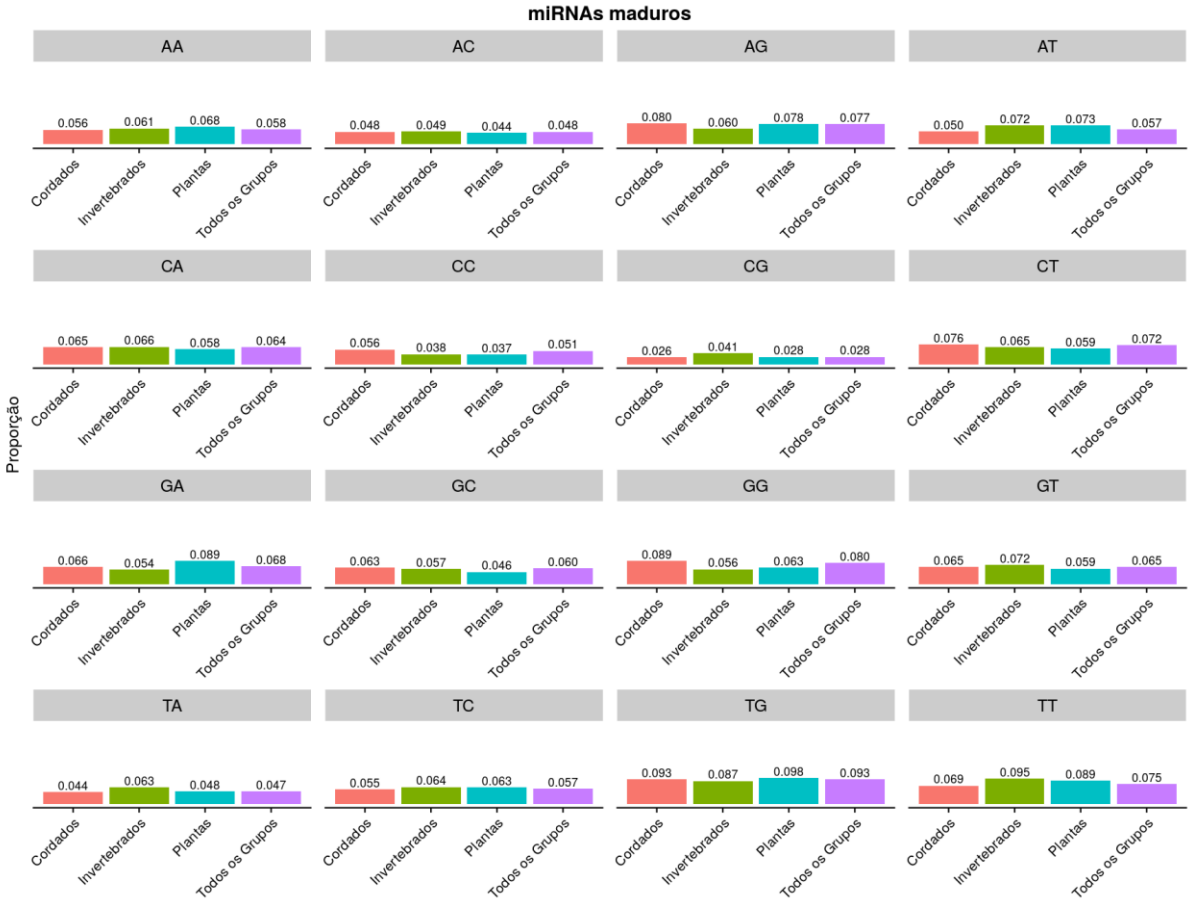
Figura 20. Distribuição de dinucleotídeos de miRNAs e mirtrons maduros, por grupo de organismo

Analisando os dinucleotídeos de maior variação entre miRNAs e mirtrons maduros (Figura 21) é possível identificar que enquanto AT apresenta maioria representativa em plantas, CC e GG destacam-se em cordados, e TT em invertebrados, o dinucleotídeo AA apresenta maior representatividade em miRNAs de plantas e em mirtrons de invertebrados.

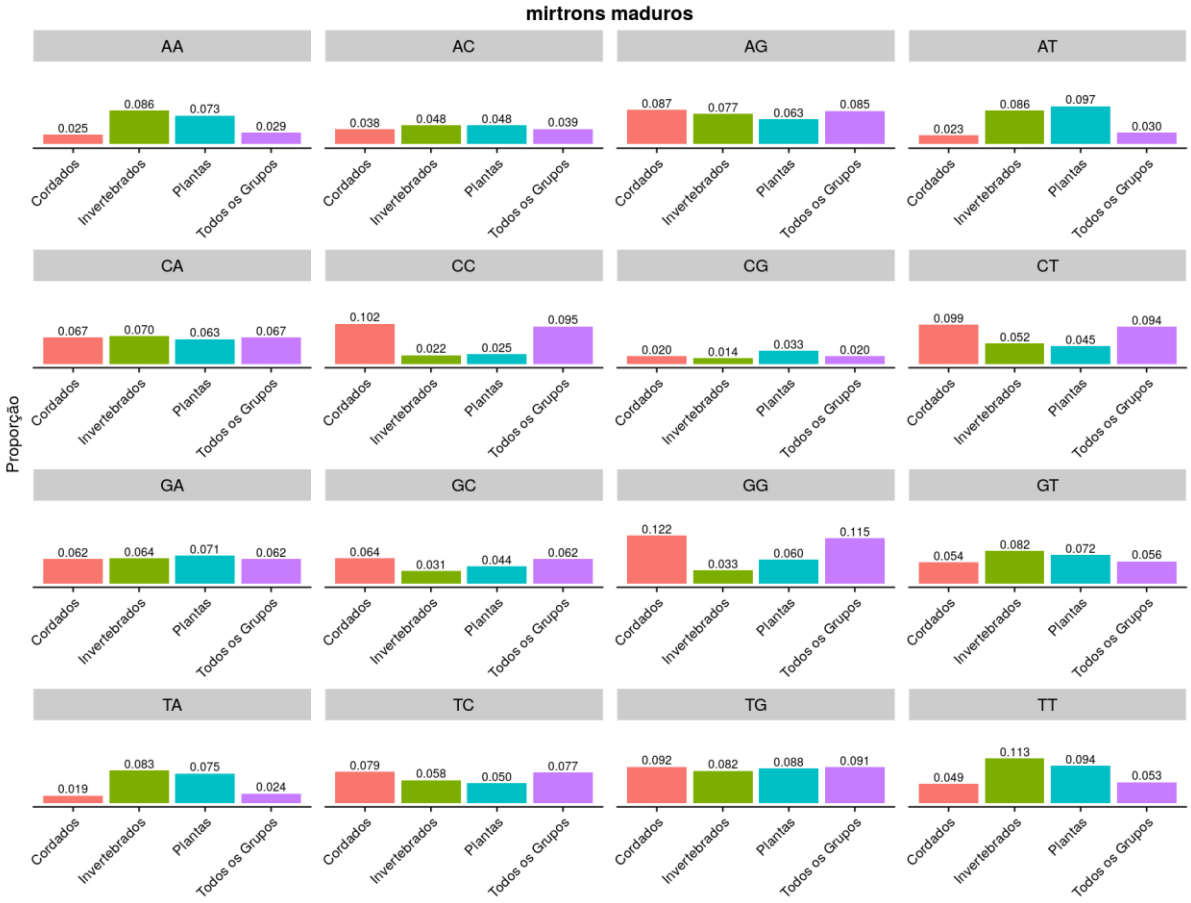
Importante destacar ainda que os dinucleotídeos AA, AT, GG e TT foram identificados como de consideráveis variações tanto para miRNAs e mirtrons precursores, quanto maduros.

Figura 21. Distribuição de dinucleotídeos de mirtrons e miRNAs maduros

A.



B.

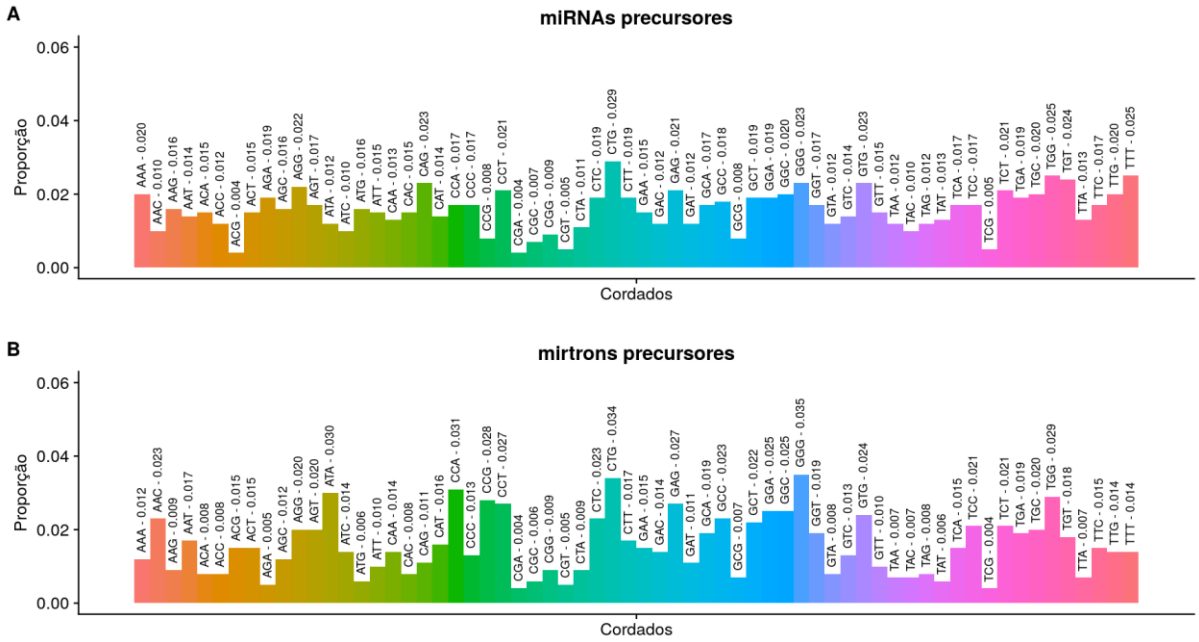


3.3.5.3. Distribuição de trinucleotídeos

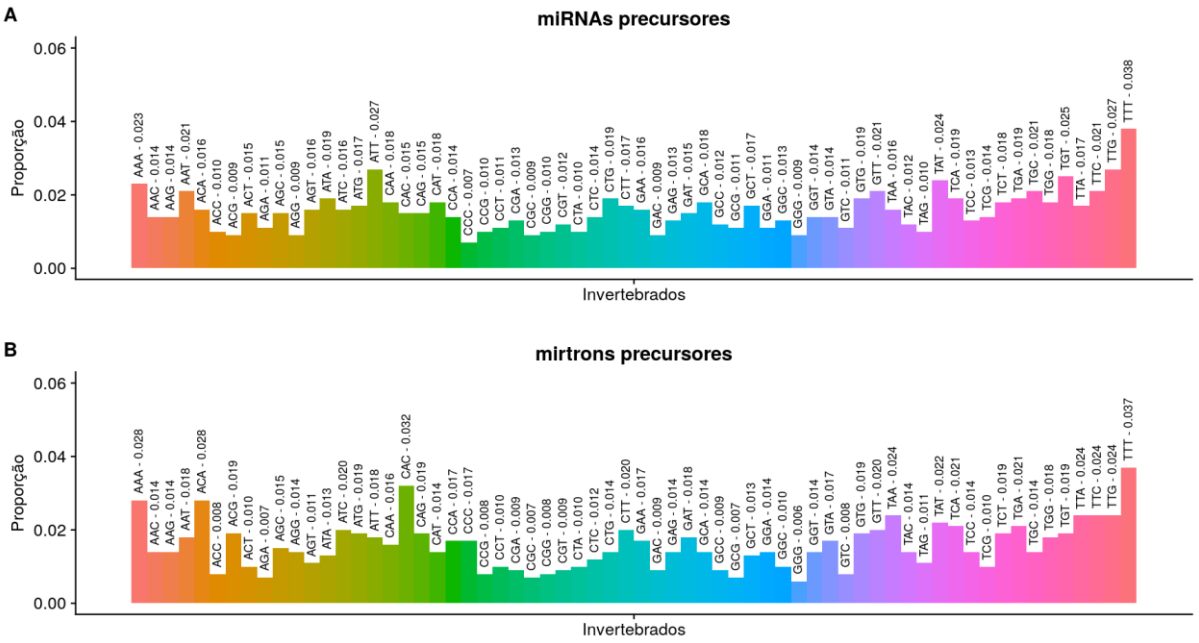
Considerando o apresentado na Figura 22, pode-se destacar como principais variações quanto a distribuição de trinucleotídeos entre miRNAs e mirtrons precursores: (i) para cordados: AAA, AAC, ATA, ACG, CCA, CCG e TTT; (ii) invertebrados: ATT e CAC; (iii); e em plantas: ACA, ACG, AGA, AGA e ATA e (iv) no geral dos grupos, pode-se notar grandes variações, dentre eles destacam-se AGA e TTT em maiores proporções em miRNAs, e CTG e GGG em mirtrons.

Figura 22. Distribuição de trinucleotídeos de mirtrons e miRNAs precursores, por grupo de organismos

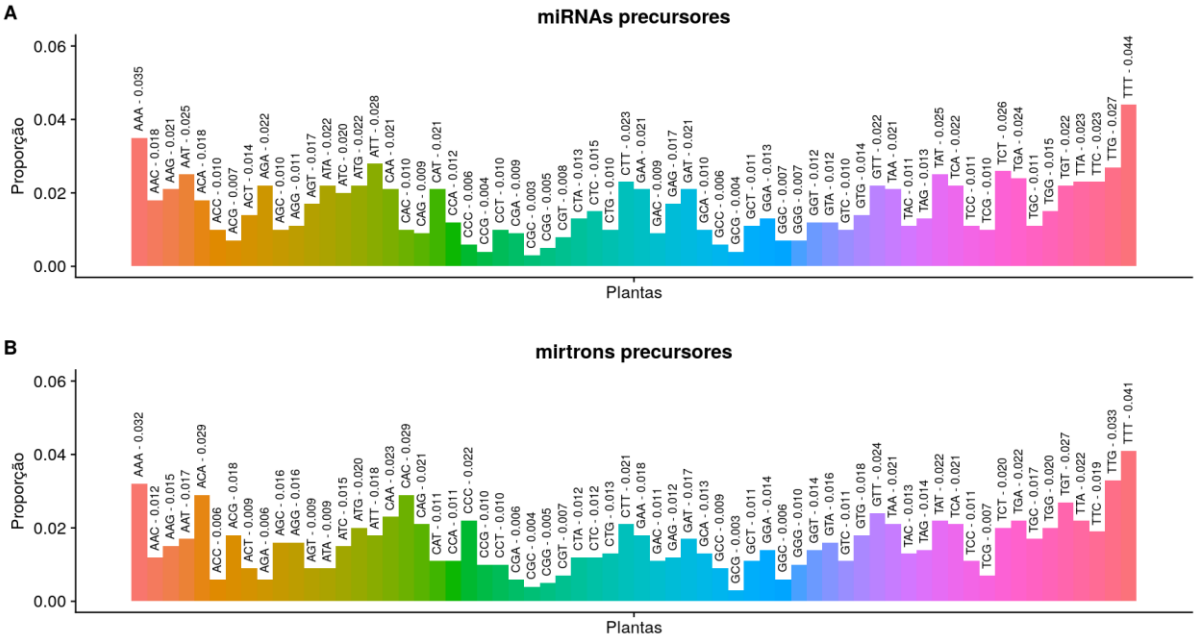
(i) Cordados



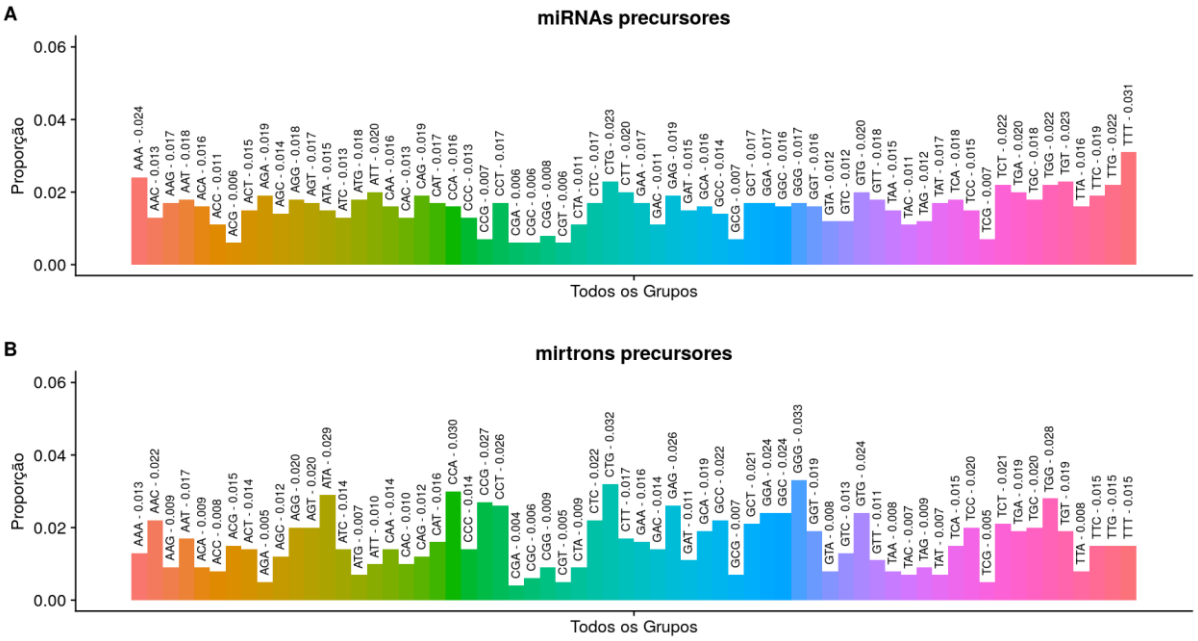
(ii) Invertebrados



(iii) Plantas



(iv) Todos os Grupos

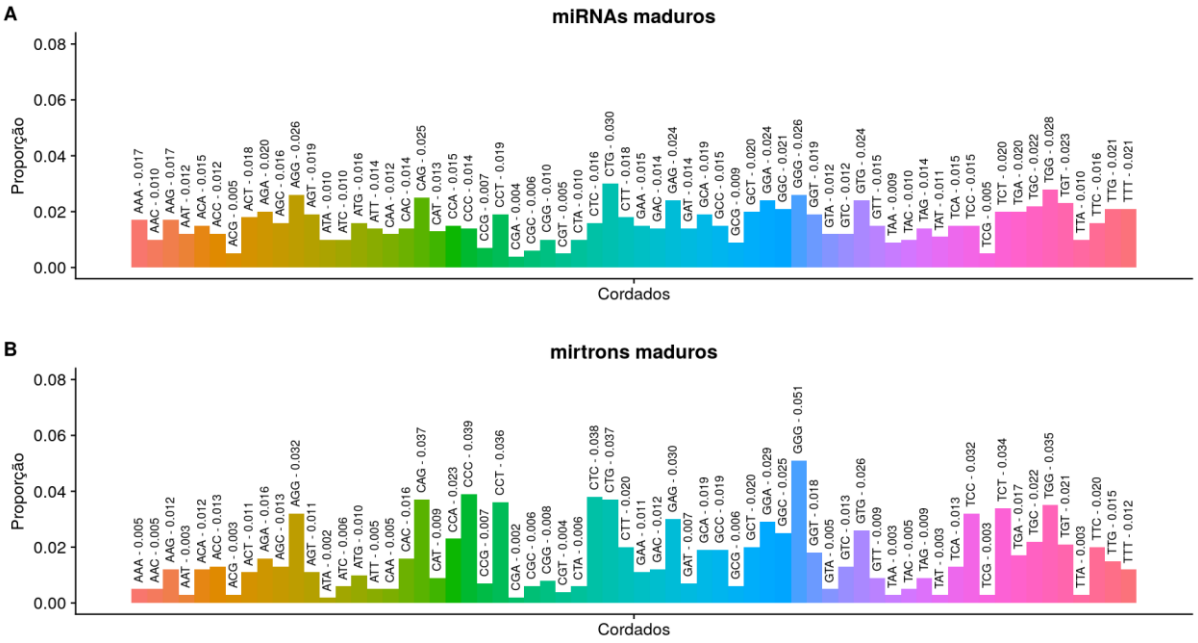


Para distribuição de trinucleotídeos em maduros (Figura 23), dentre as variações identificadas pode-se destacar comparativamente que: (i) em cordados mirtrons possuem maior proporção de CCC, GGG e TCC, e miRNAs de AAA; (ii) para invertebrados, enquanto miRNAs possuem maior proporção de ACG e CGT,

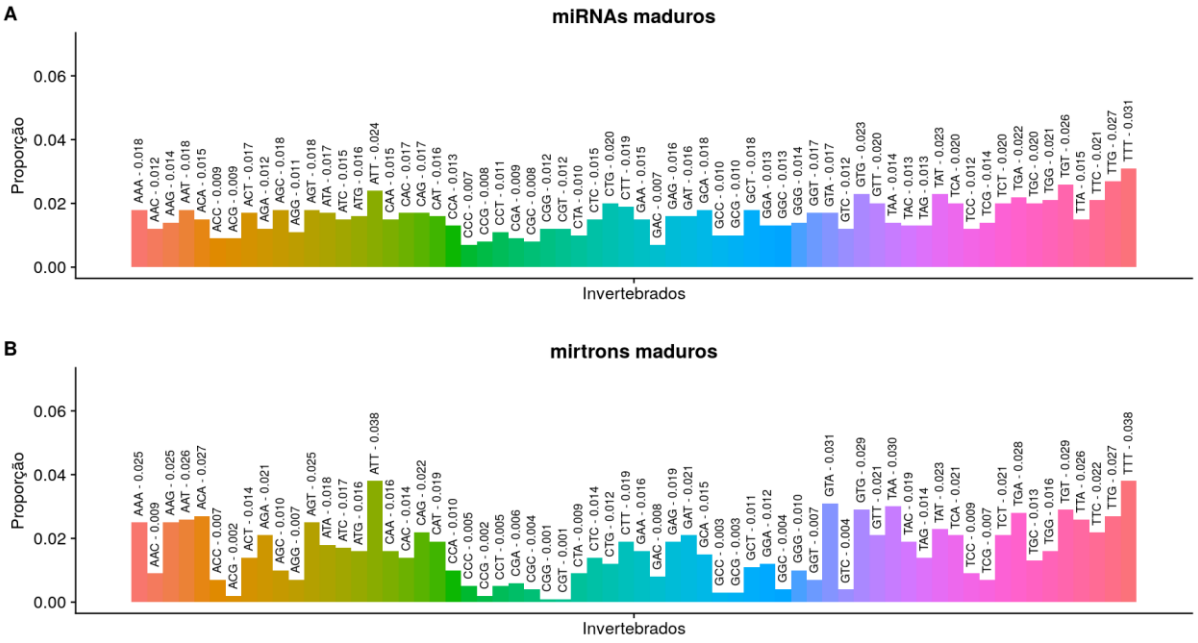
em mirtrons destacam-se GTA e TAA; e (iii) para plantas, miRNAs possuem maior proporção de GAG e mirtrons de GTA.

Figura 23. Distribuição de trinucleotídeos de mirtrons e miRNAs maduros, por grupo de organismos

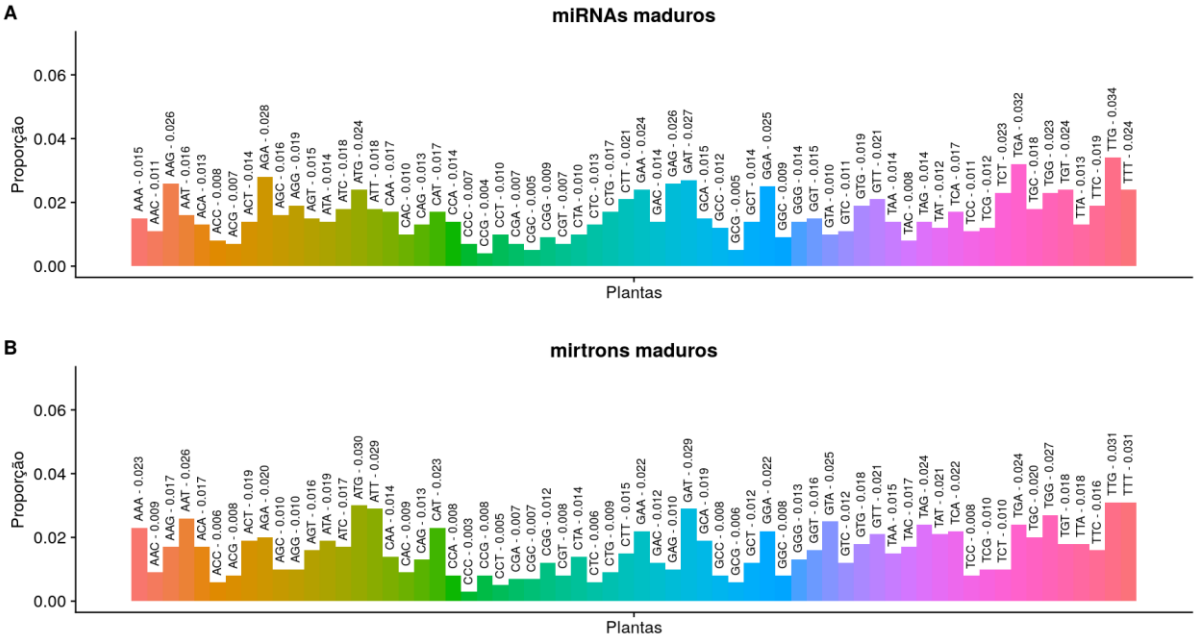
(i) Cordados



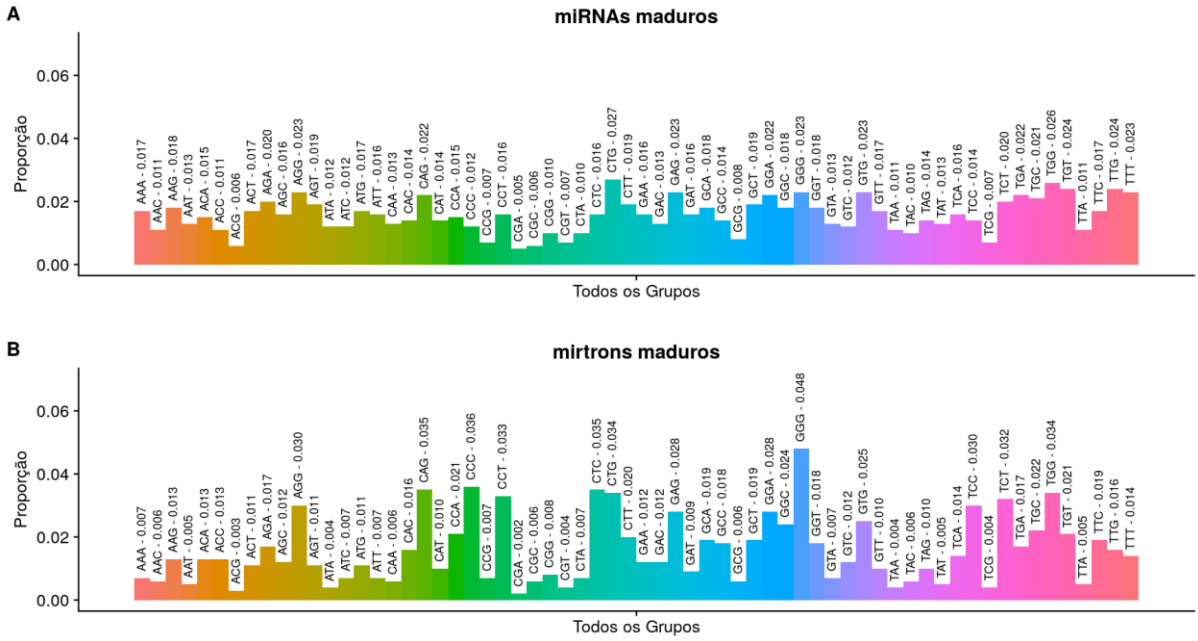
(ii) Invertebrados



(iii) Plantas



(iv) Todos os Grupos



No acumulado dos grupos, pode-se destacar que mirtrons maduros se diferenciam de miRNAS maduros principalmente em relação a proporção dos trinucleotídeos CCC, CTC, GGG e TCC.

3.3.6. Distribuição de mínimo de energia livre (MFE)

A Figura 24 apresenta a distribuição de energia mínima livre para mirtrons e miRNAs, em Kcal/mol.

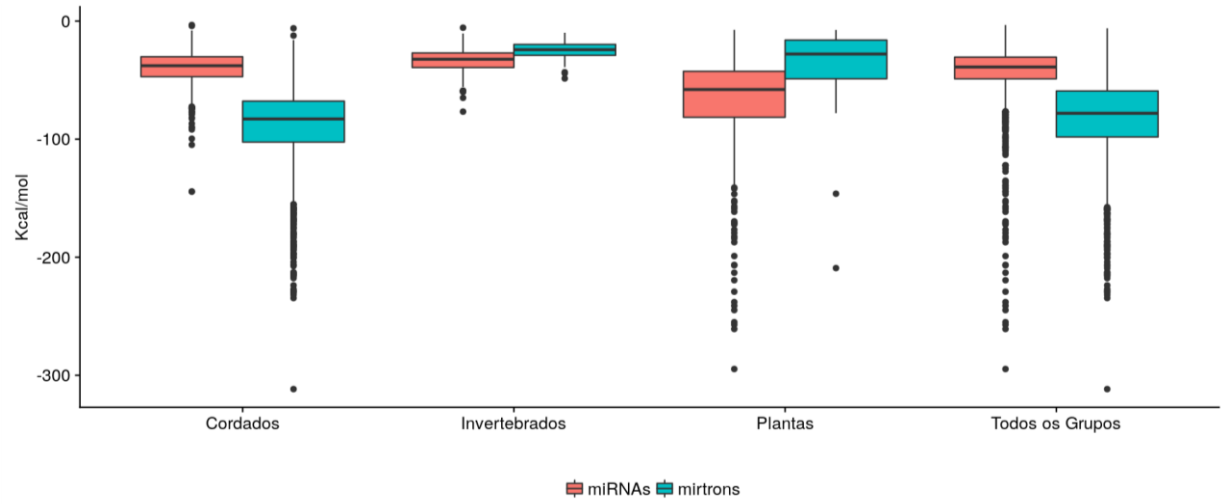


Figura 24. Distribuição de Energia Mínima Livre (MFE)
Fonte: Autoria Própria

Observa-se que enquanto miRNAs precursores de cordados e invertebrados apresentam constância e homogeneidade de energia entre os grupos de organismos, para mirtrons há considerável variação, principalmente entre mirtrons de cordados e plantas. Entre miRNAs e mirtrons, é possível evidenciar que mirtrons apresentam maior estabilidade termodinâmica, se comparada a miRNAs, dado seu menor valor médio de energia mínima livre.

3.4. DISCUSSÃO

Mirtrons são considerados uma sub-classe de miRNAs e ambos desempenham importantes papéis de regulação gênica (OKAMURA et al., 2007). Embora entre eles exista similaridade em relação a biogênese, estudos apresentam características capazes de diferenciá-los.

Neste capítulo, foi realizada análise comparativa de características, através de dados originados dos diretórios mirtronDB e miRBase. No total sete características foram analisadas comparativamente, dentre elas distribuição de tamanho de sequências, distribuição de nucleotídeos (tamanho de 1 a 3 nt), e frequência de energia mínima livre (MFE). Tais análises foram realizadas para maduros e precursores, inclusive com distinção por grupo de organismo. Até o momento não é de nosso conhecimento qualquer abordagem comparativa como a aqui apresentada, principalmente pela quantidade e estruturação dos dados utilizados e classificação de resultados por grupos de organismos.

A partir dos resultados obtidos foi possível observar indicativos de similaridades e divergências entre miRNAs e mirtrons, principalmente entre grupos de organismo. Por exemplo, para precursores ficou evidenciado que mirtrons tendem a possuir maior estabilidade termodinâmica, sendo os mirtrons precursores de cordados os que possuem menor valor de MFE. Para maduros, observou-se que tanto miRNAs quanto mirtrons possuem em sua maioria sequências de 22 nucleotídeos de tamanho, porém para plantas pôde-se observar que tanto miRNAs quanto mirtrons apresentam maioria com 21 nucleotídeos de tamanho.

É possível afirmar que este estudo apresenta características de sequência e estrutura capazes de distinguir miRNAs e mirtrons, sendo que a utilização em conjunto apresenta potencial de ganhos em assertividade. Dentre as principais características potenciais, para precursores e maduros, destacam-se conteúdo GC, frequência de bases, MFE e distribuição de nucleotídeos.

Há de se destacar que enquanto “*Distinguishing mirtrons from canonical miRNAs with data exploration and machine learning methods*” utilizou em suas análises cerca de 400 mirtrons e 700 miRNAs, e “*Comparing miRNA structure of mirtrons and non-mirtrons*“ utilizou 460 mirtrons e 2.400 miRNAs, as características aqui analisadas foram extraídas com base em 5.200 registros de miRNAs e 3.800 de mirtrons. A quantidade de registros de mirtrons aqui utilizados representa 8 vezes

mais do que o comparativo com de maior quantidade de mirtrons disponível na literatura. Destaca-se também que dentre os comparativos de miRNAs e mirtrons disponíveis, não há registro de comparativos de distribuição de mono, di e trinucleotídeos.

3.5. CONCLUSÃO

MiRNAs e mirtrons apesar de possuírem mesma origem, apresentam distinções em etapas biogênicas e características estruturais. Os resultados aqui apresentados tem por base análises realizadas com expressiva quantidade de registros e permitem a identificação de características relevantes para distinção de miRNAs e mirtrons.

Dentre as características analisadas é possível afirmar que a aplicação de um conjunto de características de estrutura e sequência apresenta potencial para a distinção de miRNAs e mirtrons, principalmente conteúdo GC, distribuição de mínimo de energia e destruição de mono, di e tri-nucleotideos.

Portanto, pode-se afirmar que este trabalho contribui para a evolução do tema através da (i) caracterização de miRNAs e mirtrons e pela possibilidade da utilização de tais características para (ii) criação ou aperfeiçoamento de preditores de mirtrons, (iii) associação de tais características à funcionalidades desempenhadas e questões evolutivas em organismos, (iii) desenvolvimento de algoritmos de seleção de características com o objetivo de identificação e reconhecimento de padrões, (iv) elaboração de modelo estatístico e aplicação de inteligência artificial para simulação de experimentos in-silico e (v) elaboração de comparativo e análise de alvos e rede de interações entre miRNAs e mirtrons. Este trabalho também apresenta potencial de promoção de novos e mais aprofundados comparativos em biologia computacional.

4. REFERÊNCIAS

ATIANAND, M. K.; CAFFREY, D. R.; FITZGERALD, K. A. Immunobiology of Long Noncoding RNAs. **The Annual Review of Immunology**, n. January, p. 177–198, 2017.

AXTELL, M. J.; MEYERS, B. C. Revisiting criteria for plant miRNA annotation in the era of big data. **The Plant Cell Advance Publication**, p. tpc.00851.2017, 2018.

BARTEL, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. **Cell**, v. 116, n. 2, p. 281–297, 2004.

BERARDINI, T. Z. et al. The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. **Genesis**, v. 53, n. 8, p. 474–485, 2015.

BEREZIKOV, E. et al. Mammalian Mirtron Genes. **Mol Cell**, v. 28, n. 2, p. 328–336, 2007.

BOLSER, D. et al. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. **Plant Bioinformatics**, v. 1374, p. 115–140, 2007.

BONNET, E. et al. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. **Bioinformatics**, v. 26, n. 12, p. 1566–1568, 2010.

BRAUN, J. E.; HUNTZINGER, E.; IZAURRALDE, E. A molecular link between miRISCs and deadenylases provides new insight into the mechanism of gene silencing by microRNAs. **Cold Spring Harbor Perspectives in Biology**, v. 4, n. 12, p. 1–15, 2012.

BUDAK, H.; AKPINAR, B. A. Plant miRNAs: biogenesis, organization and origins. **Functional and Integrative Genomics**, v. 15, n. 5, p. 523–531, 2015.

CAMACHO, C. et al. BLAST+: Architecture and applications. **BMC Bioinformatics**, v. 10, p. 1–9, 2009.

CHARIF, D.; LOBRY, J. R. Seqin R 1 . 0-2 : A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis.

Springer, 2007.

CHUNG, W. et al. Computational and experimental identification of mirtrons in Drosophila melanogaster and Caenorhabditis elegans. **Genome Research**, p. 286–300, 2011.

CROOKS, G. E. et al. WebLogo : A Sequence Logo Generator. **Genome Research**, 2004.

CURTIS, H. J. et al. Knockdown and replacement therapy mediated by artificial mirtrons in spinocerebellar ataxia 7. **Nucleic Acids Research**, v. 45, n. 13, p. 7870–7885, 2017.

CURTIS, H. J.; SIBLEY, C. R.; WOOD, M. J. A. Mirtrons, an emerging class of atypical miRNA. **Wiley Interdisciplinary Reviews: RNA**, v. 3, n. 5, p. 617–632, 2012.

DAI, X.; ZHUANG, Z.; ZHAO, P. X. psRNATarget : a plant small RNA target analysis server (2017 release). **Nucleic Acids Research**, v. 46, n. April, p. 7–10, 2018.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27–34, 1996.

FERDOUS, J.; HUSSAIN, S. S.; SHI, B. Role of microRNAs in plant drought tolerance. **Plant Biotechnology Journal**, 2015.

FRANCO-ZORRILLA, J. M. et al. Target mimicry provides a new mechanism for regulation of microRNA activity. **Nature Genetics**, v. 39, n. 8, p. 1033–1037, 2007.

FRANZ, M. et al. Cytoscape . js : a graph theory library for visualisation and analysis. **Bioinformatics**, n. September 2015, 2015.

GLASGOW, A. M. A.; DE SANTI, C.; GREENE, C. M. Non-coding RNA in cystic fibrosis. **Biochemical Society transactions**, p. BST20170469, 2018.

GLAZOV, E. A. et al. A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. **Genome Research**, p. 957–964, 2008.

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

GRAMATES, L. S. et al. FlyBase at 25: Looking to the future. **Nucleic Acids Research**, v. 45, n. D1, p. D663–D671, 2017.

GRIMSON, A. et al. MicroRNA Targeting Specificity in Mammals: Determinants Beyond Seed Pairing. **Mol Cell**, v. 27, 2007.

GULÌA, C. et al. Role of non-coding RNAs in the etiology of bladder cancer. **Genes**, v. 8, n. 11, 2017.

HU, C. et al. Human GRIN2B variants in neurodevelopmental disorders. **Journal of Pharmacological Science**, v. 132, n. 2, p. 115–121, 2016.

KARTHA, R. V.; SUBRAMANIAN, S. Competing endogenous RNAs (ceRNAs): New entrants to the intricacies of gene regulation. **Frontiers in Genetics**, v. 5, n. JAN, p. 1–9, 2014.

KATZ, M. G. et al. The role of microRNAs in cardiac development and regenerative capacity. **American journal of physiology. Heart and circulatory physiology**, v. 310, n. 5, p. H528-41, 2016.

KAUR, N.; SINGH, G. A Review Paper On Data Mining And Big Data. **International Journal of Advanced Research in Computer Science**, v. 8, n. 4, p. 407–409, 2017.

KIKUCHI, S. et al. Collection, Mapping, and Annotation of Over 28 , 000 cDNA Clones from japonica Rice. **Science**, 2003.

KOZOMARA, A.; GRIFFITHS-JONES, S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. **Nucleic Acids Research**, v. 42, n. D1, p. 68–73, 2013.

LADEWIG, E. et al. Discovery of hundreds of mirtrons in mouse and human small RNA data. **Genome Research**, v. 22, n. 9, p. 1634–1645, 2012.

LAGOS-QUINTANA, M. et al. Identification of novel genes Coding for RNAs of Small expressed RNAs. **Science**, v. 294, n. 5543, p. 853–858, 2001.

LAN, K. et al. A Survey of Data Mining and Deep Learning in Bioinformatics. **Journal of Medical Systems**, v. 42, n. 8, p. 139, 2018.

LAURETTO, M. DE S. Análise exploratória de dados. **Universidade de São Paulo (USP)**, 2001.

LAURYNAS, Č. et al. Splicing-dependent expression of microRNAs of mirtron origin in human digestive and excretory system cancer cells. **Clinical Epigenetics**, p. 1–11, 2016.

LE, T. D. et al. Computational methods for identifying miRNA sponge interactions. **Briefings in bioinformatics**, n. January, p. bbw042, 2016.

LELANDAIS-BRIÈRE, C. et al. Small RNA diversity in plants and its impact in development. **Current genomics**, v. 11, n. 1, p. 14–23, 2010.

LEMKE, J. R. et al. GRIN2B Mutations in West Syndrome and Intellectual Disability with Focal Epilepsy. **Annals of Neurology**, p. 147–154, 2013.

LEVENTHAL, B. An introduction to data mining and other techniques for advanced analytics. **Journal of Direct, Data and Digital Marketing Practice**, v. 12, n. 2, p. 137–153, 2010.

LI, W.; GODZIK, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, v. 22, n. 13, p. 1658–1659, 2006.

LIAO, P. et al. A comprehensive review of web-based resources of non-coding RNAs for plant science research. **International Journal of Biological Sciences**, 2018.

LORENZ, R. et al. ViennaRNA Package 2.0. **Algorithms for Molecular Biology**, 2011.

MENG, Y.; SHAO, C. Large-scale identification of mirtrons in arabidopsis and rice. **PLoS ONE**, v. 7, n. 2, p. 1–6, 2012.

MISHRA, N. et al. Chromosome 12p Deletion Spanning the GRIN2B Gene Presenting With a Neurodevelopmental Phenotype : A Case Report and Review of Literature. **Child Neurology Open**, 2015.

MITRA, A. et al. TC-PTP Dephosphorylates the Guanine Nucleotide Exchange Factor C3G (RapGEF1) and Negatively Regulates Differentiation of Human Neuroblastoma Cells. **Plos One**, v. 6, n. 8, p. 1–13, 2011.

MORAN, Y. et al. The evolutionary origin of plant and animal microRNAs. **Nature Ecology & Evolution**, v. 1, n. 3, p. 0027, 2017.

NAQVI, A. R. et al. The fascinating world of RNA interference. **International Journal of Biological Sciences**, v. 5, n. 2, p. 97–117, 2009.

NIST/SEMATECH. **e-Handbook of Statistical Methods**. [s.l: s.n.].

OKAMURA, K. et al. The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in Drosophila. **Cell**, v. 130, n. 1, p. 89–100, 2007.

PALAZZO, A. F.; LEE, E. S. Non-coding RNA: What is functional and what is junk? **Frontiers in Genetics**, v. 5, n. JAN, p. 1–11, 2015.

PAN, Y. et al. Association of genetic variants of GRIN2B with autism. **Scientific Reports**, p. 1–5, 2015.

PANIR, K. et al. Non-coding RNAs in endometriosis: a narrative review. **Human reproduction update**, v. 24, n. 4, p. 497–515, 2018.

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.

PAYTUV, A. et al. GREENC : a Wiki-based database of plant lncRNAs. **Published online 17 November 2015 Nucleic Acids Research**, v. 44, n. November 2015, p. 1161–1166, 2016.

QU, Z.; ADELSON, D. L. Evolutionary conservation and functional roles of ncRNA. **Frontiers in Genetics**, v. 3, n. OCT, p. 1–11, 2012.

RAJAGOPALAN, R. et al. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. **Genes and Development**, v. 20, n. 24, p. 3407–3425, 2006.

RAZA, K. Application of Data mining in Bioinformatics. **Indian Journal of Computer**

Science and Engineering, v. 1, n. 2, p. 114–118, 2010.

ROMAO, J. M. et al. MicroRNAs in bovine adipogenesis : genomic context , expression and function. **BMC Genomics**, p. 1–15, 2014.

RORBACH, G.; UNOLD, O.; KONOPKA, B. M. Distinguishing mirtrons from canonical miRNAs with data exploration and machine learning methods. **Scientific Reports**, 2018.

RUBY, J. G. et al. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. **Genome Research**, v. 17, n. 12, p. 1850–1864, 2007.

RUBY, J. G.; JAN, C. H.; BARTEL, D. P. Intronic microRNA precursors that bypass Drosha processing. **Nature**, 2007.

SALMENA, L. et al. A ceRNA hypothesis: the Rosetta stone of a hidden RNA language? **Cell**, v. 146, n. 3, p. 353–358, 2011.

SHAH, B. et al. C3G / Rapgef1 Is Required in Multipolar Neurons for the Transition to a Bipolar Morphology during Cortical Development. **Plos One**, p. 1–22, 2016.

SIBLEY, C. R. et al. Silencing of Parkinson ' s disease-associated genes with artificial mirtron mimics of miR-1224. **Nucleic Acids Research**, 2012.

SOEMEDI, R. et al. The effects of structure on pre-mRNA processing and stability. **Methods**, v. 125, p. 36–44, 2017.

TEO, Y. Y. Exploratory data analysis in large-scale genetic studies. **Biostatistics**, v. 11, n. 1, p. 70–81, 2010.

THOMSON, D. W.; DINGER, M. E. Endogenous microRNA sponges: evidence and controversy. **Nature Reviews Genetics**, v. 17, n. 5, p. 272–283, 2016.

TITOV, I. I.; VOROZHEYKIN, P. S. Comparing miRNA structure of mirtrons and non-mirtrons. **BMC Genomics**, 2018.

WEN, J. et al. Analysis of Nearly One Thousand Mammalian Mirtrons Reveals Novel Features of Dicer Substrates. **PLoS Computational Biology**, v. 11, n. 9, 2015.

WESTHOLM, J. O.; LAI, E. C. Mirtrons: MicroRNA biogenesis via splicing. **Biochimie**, v. 93, n. 11, p. 1897–1904, 2011.

YONES, C. A. et al. MiRNAfe: A comprehensive tool for feature extraction in microRNA prediction. **BioSystems**, v. 138, p. 1–5, 2015.

YOUSEF, M. et al. MicroRNA categorization using sequence motifs and k-mers. **BMC Bioinformatics**, v. 18, n. 1, p. 1–9, 2017.

ZERBINO, D. R. et al. Ensembl 2018. **Nucleic Acids Research**, v. 46, n. D1, p. D754–D761, 2018.

ZHANG, Y. et al. Comparison of miRNA evolution and function in plants and animals. **MicroRNA**, v. 7, n. 1, p. 4–10, 2018.

ZHU, Q. et al. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. **Genome Research**, p. 1456–1465, 2008.

ANEXO 1

ANEXO 2

ANEXO 3

ANEXO 4

ANEXO 5

ANEXO 6

ANEXO 7

ANEXO 8

ANEXO 1: Papers retrieved in data collection process.

Pubmed Id	Title	Year	Data Collected
28717225	A mammalian mirtron miR-1224 promotes tube-formation of human primary endothelial cells by targeting anti-angiogenic factor epsin2.	2017	Yes
28579988	An Expanded Role for HLA Genes: HLA-B Encodes a microRNA that Regulates IgA and Other Immune Response Transcripts.	2017	Yes
28575281	Knockdown and replacement therapy mediated by artificial mirtrons in spinocerebellar ataxia 7.	2017	
28053119	Short intron-derived ncRNAs.	2017	
27715458	Alternative splicing of a viral mirtron differentially affects the expression of other microRNAs from its cluster and of the host transcript.	2016	
27173734	Argonaute-associated short introns are a novel class of gene regulators.	2016	
27119849	An Improved microRNA Annotation of the Canine Genome.	2016	Yes
27021098	Profile of microRNA in Blood Plasma of Healthy Humans.	2016	
27019673	Splicing-dependent expression of microRNAs of mirtron origin in human digestive and excretory system cancer cells.	2016	
26848861	Tumor suppressor microRNAs are downregulated in myelodysplastic syndrome with spliceosome mutations.	2016	
26325366	Analysis of Nearly One Thousand Mammalian Mirtrons Reveals Novel Features of Dicer Substrates.	2015	Yes
26186287	Tailoring MicroRNA Function: The Role of Uridylation in Antagonizing Mirtron Expression.	2015	
26145176	Uridylation of RNA Hairpins by Tailor Confines the Emergence of MicroRNAs in Drosophila.	2015	

26145174	Selective Suppression of the Splicing-Mediated MicroRNA Pathway by the Terminal Uridyltransferase Tailor.	2015	
26089392	Functional VEGFA knockdown with artificial 3'-tailed mirtrons defined by 5' splice site and branch point.	2015	
25609829	Intronic regions of plant genes potentially encode RDR (RNA-dependent RNA polymerase)-dependent small RNAs.	2015	
25447893	Deep sequencing analyses of pine wood nematode Bursaphelenchus xylophilus microRNAs reveal distinct miRNA expression patterns during the pathological process of pine wilt disease.	2015	
25319661	Gene silencing in vitro and in vivo using intronic microRNAs.	2014	
25248950	The small RNA diversity from Medicago truncatula roots under biotic interactions evidences the environmental plasticity of the miRNAome.	2014	Yes
25055917	Experimental validation of predicted mammalian microRNAs of mirtron origin.	2014	
24835514	Expanding the annotation of zebrafish microRNAs based on small RNA sequencing.	2014	Yes
24823351	Versatile microRNA biogenesis in animals and their viruses.	2014	
24687917	Small RNA as a regulator of hematopoietic development, immune response in infection and tumorigenesis.	2014	
24548287	MicroRNAs in bovine adipogenesis: genomic context, expression and function.	2014	Yes
24443800	A comprehensive microRNA expression profile of the backfat tissue from castrated and intact full-sib pair male pigs.	2014	Yes
24330712	Genome-wide characterization of microRNA in foxtail millet (Setaria italica).	2013	Yes
24235016	Extremely complex populations of small RNAs in the mouse retina and RPE/choroid.	2013	
24166299	The pathway of miRNA maturation.	2014	

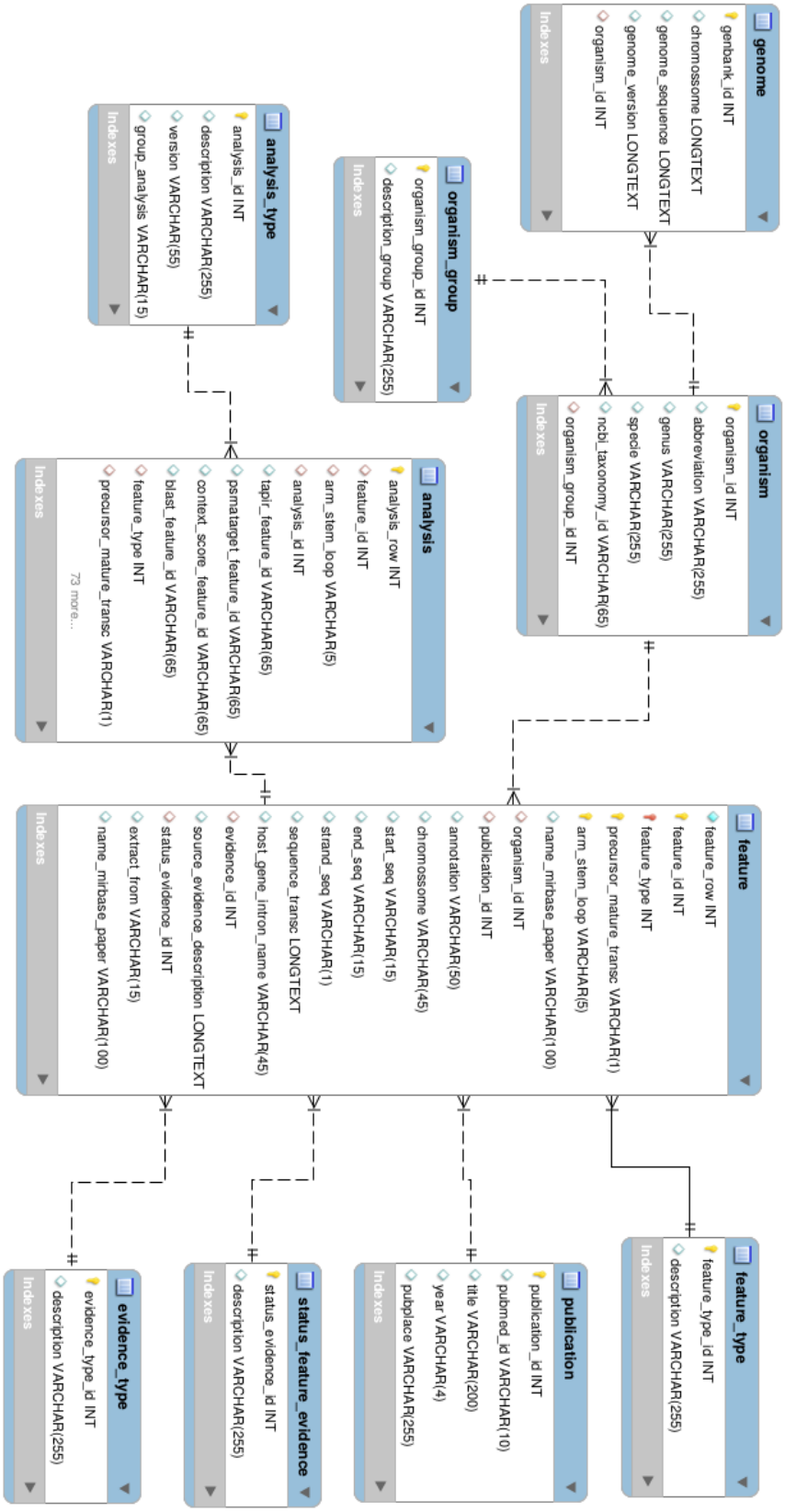
24035931	Mmu-miR-702 functions as an anti-apoptotic mirtron by mediating ATF6 inhibition in mice.	2013	
23882112	The impact of age, biogenesis, and genomic clustering on Drosophila microRNA evolution.	2013	
23396444	Global profiling of miRNAs and the hairpin precursors: insights into miRNA processing and novel miRNA discovery.	2013	
23341152	MIR846 and MIR842 comprise a cistronic MIRNA pair that is regulated by abscisic acid by alternative splicing in roots of Arabidopsis.	2013	
23325850	Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues.	2013	
23175445	Noncanonical microRNAs and endogenous siRNAs in normal and psoriatic human skin.	2013	
23018783	Human mirtrons can express functional microRNAs simultaneously from both arms in a flanking exon-independent manner.	2012	
22955976	Discovery of hundreds of mirtrons in mouse and human small RNA data.	2012	Yes
22848108	Silencing of Parkinson's disease-associated genes with artificial mirtron mimics of miR-1224.	2012	
22733569	Mirtrons, an emerging class of atypical miRNA.	2012	
22647847	Artificial mirtron-mediated gene knockdown: functional DMPK silencing in mammalian cells.	2012	
22546559	Identification of mirtrons in rice using MirtronPred: a tool for predicting plant mirtrons.	2012	Yes
22388699	Computational identification of microRNAs and their targets in cassava (Manihot esculenta Crantz.).	2013	Yes
22348048	Large-scale identification of mirtrons in Arabidopsis and rice.	2013	Yes
22270084	Biogenesis of mammalian microRNAs by a non-canonical processing pathway.	2012	

22223733	Mirtron microRNA-1236 inhibits VEGFR-3 signaling during inflammatory lymphangiogenesis.	2012	
22201644	Naive and primed murine pluripotent stem cells have distinct miRNA expression profiles.	2012	
22190743	Common and distinct patterns of terminal modifications to mirtrons and canonical microRNAs.	2012	
21947201	A Drosophila genetic screen yields allelic series of core microRNA biogenesis factors and reveals post-developmental roles for microRNAs.	2011	
21914725	The biogenesis and characterization of mammalian microRNAs of mirtron origin.	2012	
21843590	Small RNAs derived from longer non-coding RNAs.	2011	
21712401	A role for noncanonical microRNAs in the mammalian brain revealed by phenotypic differences in Dgcr8 versus Dicer1 knockouts and small RNA sequencing.	2011	
21712066	Mirtrons: microRNA biogenesis via splicing.	2011	
21518803	Plant siRNAs from introns mediate DNA methylation of host genes.	2011	
21479628	Small RNA library preparation for next-generation sequencing by single ligation, extension and circularization technology.	2011	
21420026	MicroRNAs: miRRORS of health and disease.	2011	
21228487	Solexa sequencing analysis of chicken pre-adipocyte microRNAs.	2011	Yes
21177960	Computational and experimental identification of mirtrons in Drosophila melanogaster and Caenorhabditis elegans.	2011	Yes
21138856	Hypermethylation of CpG islands and shores around specific microRNAs and mirtrons is associated with the phenotype and presence of bladder cancer.	2011	
20841420	Noncanonical cytoplasmic processing of viral microRNAs.	2010	

20713509	Canonical and alternate functions of the microRNA biogenesis machinery.	2010	
20620959	MicroRNA biogenesis via splicing and exosome-mediated trimming in Drosophila.	2010	
20129062	A mammalian herpesvirus uses noncanonical expression and processing mechanisms to generate viral MicroRNAs.	2010	
25067898	Reply to evolutionary flux of canonical microRNAs and mirtrons in Drosophila.	2010	
20037610	Evolutionary flux of canonical microRNAs and mirtrons in Drosophila.	2010	Yes
19633723	Repertoire of bovine miRNA and miRNA-like small regulatory RNAs expressed upon viral infection.	2009	
19173032	The fascinating world of RNA interference.	2009	
19148136	Glucocorticoid-regulated microRNAs and mirtrons in acute lymphoblastic leukemia.	2009	
19047376	A Drosophila pasha mutant distinguishes the canonical microRNA and mirtron pathways.	2009	
18923076	Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs.	2008	
18769156	The long and short of inverted repeat genes in animals: microRNAs, mirtrons and hairpin RNAs.	2008	
18687877	A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains.	2008	Yes
18469162	A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach.	2008	Yes
17981129	And now introducing mammalian mirtrons.	2007	
17964270	Mammalian mirtron genes.	2007	Yes

17599402	The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila.	2007	Yes
17589500	Intronic microRNA precursors that bypass Drosha processing.	2007	Yes
21085120	Formation, Regulation and Evolution of Caenorhabditis elegans 3′UTRs	2011	Yes

ANEXO 2: mirtronDB database modelling



ANEXO 3: Genomes used for similarity analysis.

SPECIE	GROUP	VERSION	GENOME
G. gallus	Chordate	5.0	https://www.ensembl.org/Gallus_gallus/Info/Index
B. taurus	Chordate	UMD3.1	https://www.ensembl.org/Bos_taurus/Info/Index
C. familiaris	Chordate	CanFam3.1	http://www.ensembl.org/Canis_familiaris/
H. sapiens	Chordate	HG38	ftp://ftp.ensembl.org/pub/release-90/fasta/homo_sapiens/
M. mulatta	Chordate	8.0.1	http://www.ensembl.org/Macaca_mulatta/Info/Index
M. musculus	Chordate	GRCm38.p6	https://www.ensembl.org/Mus_musculus/Info/Index
P. troglodytes	Chordate	CHIMP2.1.4	ftp://ftp.ensembl.org/pub/release-90/fasta/pan_troglodytes/dna/
S. scrofa	Chordate	Sscrofa10.2	ftp://ftp.ensembl.org/pub/release-89/fasta/sus_scrofa/dna/
D. rerio	Chordate	Zv9	ftp://ftp.ensembl.org/pub/release-79/fasta/danio_rerio/dna/
D. melanogaster	Invertebrate	r6.15	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.15_FB2017_02/
D. pseudoobscura	Invertebrate	r3.2	ftp://ftp.flybase.net/genomes/Drosophila_pseudoobscura/dpse_r3.2_FB2014_04/
D. simulans	Invertebrate	r2.02	ftp://ftp.flybase.net/genomes/Drosophila_simulans/dsim_r2.02_FB2017_04/
C. elegans	Invertebrate	WBcel235	https://www.ensembl.org/Caenorhabditis_elegans/Info/Index
A. thaliana	Plant	TAIR10	https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release
M. esculenta	Plant	V6.1	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Mesculenta
M. truncatula	Plant	4.0	https://plants.ensembl.org/Medicago_truncatula/Info/Index
O. sativa	Plant	IRGSP-1.0	https://plants.ensembl.org/Oryza_sativa/Info/Index
S. italica	Plant	JGIv2.0	http://plants.ensembl.org/Setaria_italica/Info/Index

ANEXO 4: Mirtron similarity analysis among species.

Mirtrons origin	Organism Genome										Total
	B. taurus	*C. familiaris*	*D. melanogaster*	*D. simulans*	*G. gallus*	*H. sapiens*	*M. mulatta*	*P. troglodytes*	*S. italica*	*S. scrofa*	
Chordates	*1*	*2*	*0*	*0*	*1*	*19*	*225*	*646*	*0*	*2*	*896*
B. taurus		1								1	2
D. rerio					1						1
H. sapiens	1	1					222	643		1	868
M. mulatta						10		3			13
P. troglodytes						9	3				12
Plants	0	0	0	0	0	0	0	0	2	0	2
O. sativa									2		2
Invertebrates	0	0	2	44	0	0	0	0	0	0	46
D. melanogaster				44							44
D. simulans			2								2
Total	1	2	2	44	1	19	225	646	2	2	944

ANEXO 5: Mature mirtrons size by stem, group and species.

a. 3’ mature mirtrons size by group and species

3' Mature	16	17	18	19	20	21	22	23	24	25	26	27	28	29	Total
Chordate		1	17	25	132	372	349	153	55	9	2	2	1		1,118
Bos taurus															
Canis familiaris				1	1	1		1							4
Danio rerio							1	1							2
Gallus gallus					1	6	7	2	1						17
Homo sapiens		1	9	13	67	196	174	69	29	6	1	2	1		568
Macaca mulatta						3	4	1							8
Mus musculus			8	10	63	165	162	79	25	3	1				516
Pan troglodytes				1		1	1								3
Sus scrofa															
Plant			3	6	7	13	8	14	29	5	3				88
Arabidopsis thaliana			2	1		1	2	2	5						13
Manihot esculenta						1									1
Medicago truncatula															
Oryza sativa			1	5	7	11	6	12	24	4	2				72
Setaria italica										1	1				2
Invertebrate					3	8	23	13	3	4	1	2			57
Caenorhabditis elegans						2	5	3	1						11
Drosophila melanogaster					1	5	8	6	2	1					23
Drosophila pseudoobscura					1	1	7	4		3	1	2			19
Drosophila simulans					1		3								4
Total	0	1	20	31	142	393	380	180	87	18	6	4	1	0	1,263

b. 5’ mature mirtrons size by group and species

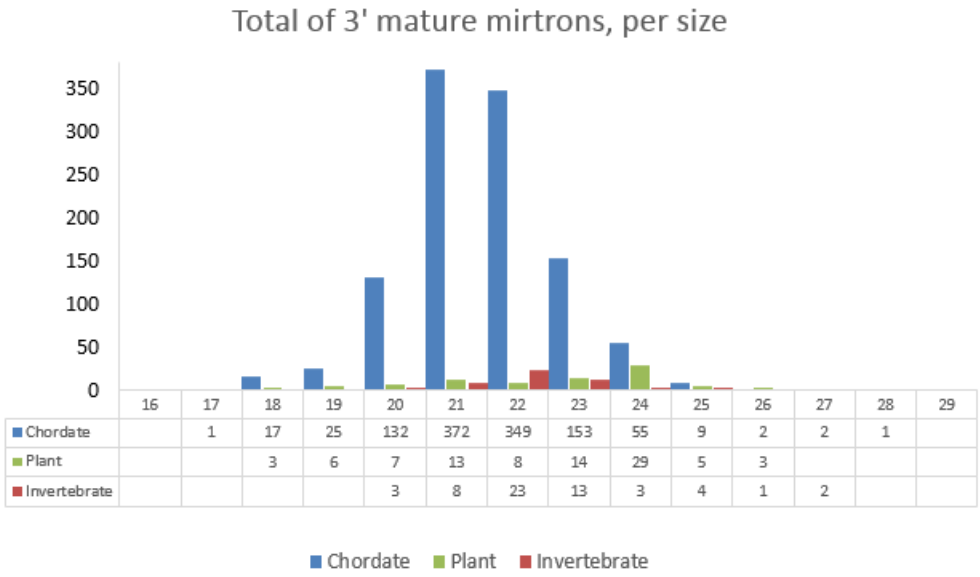
5' Mature	16	17	18	19	20	21	22	23	24	25	26	27	28	29	Total
Chordate	6	5	22	38	79	178	356	258	100	51	6		2		1,101
Bos taurus						1									1
Canis familiaris					1	1	1		1						4
Danio rerio															
Gallus gallus					1	1									2
Homo sapiens	3	4	20	24	51	78	185	126	48	24	4		1		568
Macaca mulatta				1	2	1									4
Mus musculus	3	1	2	13	23	95	168	132	49	27	2		1		516
Pan troglodytes									1						1
Sus scrofa					1	1	2		1						5
Plant				1	2	26	18	1	3						51
Arabidopsis thaliana															
Manihot esculenta															
Medicago truncatula					1	25	1								27
Oryza sativa				1	1	1	17	1	2						23
Setaria italica									1						1
Invertebrate						1	7			2				1	11
Caenorhabditis elegans						1	6			1					8
Drosophila melanogaster							1			1				1	3
Drosophila pseudoobscura															
Drosophila simulans															
Total	6	5	22	39	81	205	381	259	103	53	6	0	2	1	1,163

c. All mature mirtrons size by group and species

All mature mirtrons	16	17	18	19	20	21	22	23	24	25	26	27	28	29	Total
Chordate	6	6	39	63	211	550	705	411	155	60	8	2	3	0	2,219
Bos taurus	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Canis familiaris	0	0	0	1	2	2	1	1	1	0	0	0	0	0	8
Danio rerio	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2
Gallus gallus	0	0	0	0	2	7	7	2	1	0	0	0	0	0	19
Homo sapiens	3	5	29	37	118	274	359	195	77	30	5	2	2	0	1,136
Macaca mulatta	0	0	0	1	2	4	4	1	0	0	0	0	0	0	12
Mus musculus	3	1	10	23	86	260	330	211	74	30	3	0	1	0	1,032
Pan troglodytes	0	0	0	1	0	1	1	0	1	0	0	0	0	0	4
Sus scrofa	0	0	0	0	1	1	2	0	1	0	0	0	0	0	5
Plant	0	0	3	7	9	39	26	15	32	5	3	0	0	0	139
Arabidopsis thaliana	0	0	2	1	0	1	2	2	5	0	0	0	0	0	13
Manihot esculenta	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Medicago truncatula	0	0	0	0	1	25	1	0	0	0	0	0	0	0	27
Oryza sativa	0	0	1	6	8	12	23	13	26	4	2	0	0	0	95
Setaria italica	0	0	0	0	0	0	0	0	1	1	1	0	0	0	3
Invertebrate	0	0	0	0	3	9	30	13	3	6	1	2	0	1	68
Caenorhabditis elegans	0	0	0	0	0	3	11	3	1	1	0	0	0	0	19
Drosophila melanogaster	0	0	0	0	1	5	9	6	2	2	0	0	0	1	26
Drosophila pseudoobscura	0	0	0	0	1	1	7	4	0	3	1	2	0	0	19
Drosophila simulans	0	0	0	0	1	0	3	0	0	0	0	0	0	0	4
Total	6	6	42	70	223	598	761	439	190	71	12	4	3	1	2,426

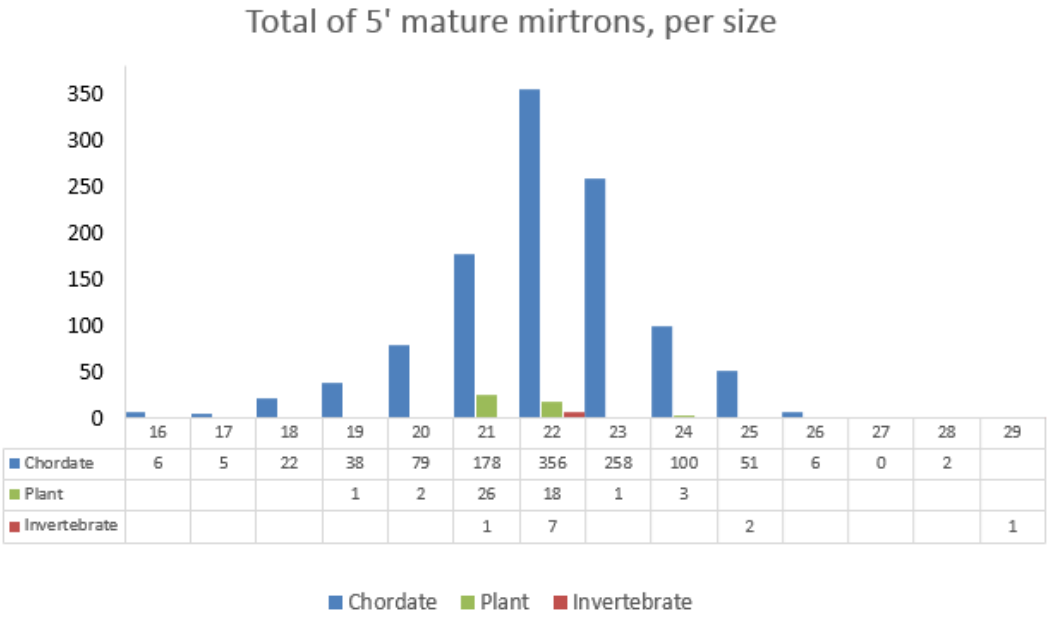
ANEXO 6: Mature mirtrons per stem and size.

a. 3’ mature mirtrons per size



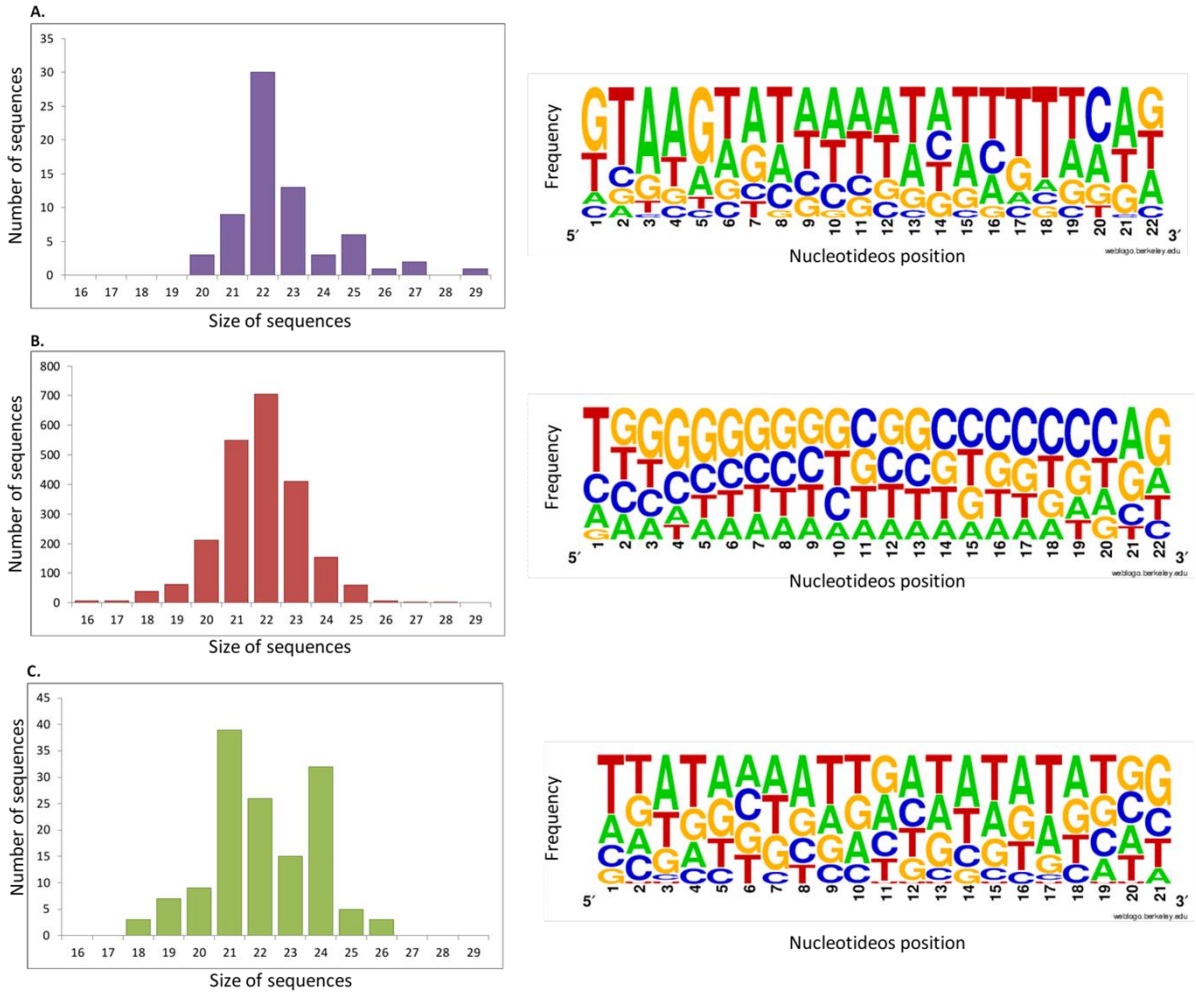
	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Chordate		1	17	25	132	372	349	153	55	9	2	2	1	
Plant			3	6	7	13	8	14	29	5	3			
Invertebrate					3	8	23	13	3	4	1	2		

b. 5’ mature mirtrons per size



	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Chordate	6	5	22	38	79	178	356	258	100	51	6	0	2	
Plant				1	2	26	18	1	3					
Invertebrate						1	7			2				1

ANEXO 7: Mature mirtrons bases distribution by nucleotides number and frequency. A. Invertebrates, B. Chordates and C. Plants.



ANEXO 8: Total of potential ceRNAs per organism:

lncRNAs from	mirtrons from				Total
	A. thaliana	*M. esculenta*	*M. truncatula*	*O. sativa*	*S. italica*	
Amborela trichopoda	4		5	12		21
Ananas comosus	1			18		19
Arabidopsis lyrata				18		18
Arabidopsis thaliana	5		1	4		10
Brachypodium distachyon	9		6	79	1	95
Capsella grandiflora	1		1	8		10
Capsella rubella	1			2		3
Carica papaya	1		2	16		19
Chlamydomonas reinhardtii	2			32		34
Citrus clementina				3		3
Citrus sinensis			2	14		16
Coccomyxa subellipsoidea				1		1
Cucumis sativus	2		1	11		14
Eucalyptus grandis		1		16		17
Eutrema salsugineum				6		6
Fragaria vesca	2		4	14		20
Glycine max			2	34		36
Gossypium raimondii	3		1	6		10
Linum usitatissimum			2	6		8
Malus domestica	4		4	12		20
Manihot esculenta	6		1	8		15
Medicago truncatula	1		3	16		20
Micromonas pusilla CCMP	4			13		17
Micromonas pusilla RCC	2			3		5
Mimulus guttatus	2			6		8
Musa acuminata	2		1	12	1	16
Oryza Sativa				37		37
Ostreococcus lucimarinus	1			10		11
Phaseolus vulgaris				12		12
Physcomitrella patens	11		13	86		110
Populus trichocarpa	3		3	16		22
Prunus persica	3		1	22		26
Ricinus communis	7		1	10		18
Selaginella moellendorffii				6		6
Setaria italica	8		1	54		63
Solanum lycopersicum	3		5	20		28
Solanum tuberosum			1	36		37
Sorghum bicolor	15		1	99	1	116
Spirodela polyrhiza	2			4		6
Theobroma cacao	8		3	22	1	34
Triticum aestivum	76		12	423		511
Vitis vinifera	2			6		8
Volvox carteri	1		2	26		29
Zea mays	18		5	166		189
Zostera marina				14		14
Total	**210**	**1**	**84**	**1,439**	**4**	**1,738**