UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

RICARDO PIUCO

Estudo dos *PIWI-interacting* RNAs (piRNAs): do desenvolvimento de uma base de dados à análise sistemática da expressão gênica em tecidos normais e tumorais

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

São Paulo
2023

RICARDO PIUCO

Estudo dos *PIWI-interacting* RNAs (piRNAs): do desenvolvimento de uma base de dados à análise sistemática da expressão gênica em tecidos normais e tumorais

Versão para entrega na biblioteca.

Tese apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Área de Concentração: Bioinformática

Orientador: Dr. Pedro A. F. Galante

São Paulo
2023

Ficha Catalográfica

Ficha catalográfica elaborada com dados inseridos pelo(a) autor(a)
Biblioteca Carlos Benjamin de Lyra
Instituto de Matemática e Estatística
Universidade de São Paulo

Piuco, Ricardo
Estudo dos PIWI-interacting RNAs (piRNAs): do desenvolvimento de uma base de dados à análise sistemática da expressão gênica em tecidos normais e tumorais / Ricardo Piuco; orientador, Pedro Alexandre Favoretto Galante. - São Paulo, 2023.
67 p.: il.

Tese (Doutorado) - Programa Interunidades de Pós-Graduação em Bioinformática / Instituto de Matemática e Estatística / Universidade de São Paulo.
Bibliografia
Versão original

1. EXPRESSÃO GÊNICA. 2. RNA. 3. Genômica. 4. NEOPLASIAS. I. Galante, Pedro Alexandre Favoretto. II. Título.

Bibliotecárias do Serviço de Informação e Biblioteca Carlos Benjamin de Lyra do IME-USP, responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2: Maria Lúcia Ribeiro CRB-8/2766; Stela do Nascimento Madruga CRB 8/7534.

Folha de Avaliação

(Esta página foi intencionalmente deixada em branco)

Agradecimentos

Agradeço muito aos meus pais primeiramente por sempre estarem ao meu lado, proporcionando tantos privilégios, não só materiais mas incentivos e sua própria história de vida que me trouxeram aqui hoje.

Agradeço também ao meu orientador do Doutorado-Direto, Dr. Pedro Galante, que acreditou em um garoto recém formado e com quase nenhuma bagagem de vivência na bioinformática. Estendo esse agradecimento a todos que me apoiaram na minha trajetória científica, seja sendo quase um co-orientador ou apenas ouvindo eu reclamar de tudo. Não esquecendo, é claro, da Dra. Cibele Masotti, Dra. Raquel Chacon, Dra. Anamaria Camargo, além dos meus amigos, que em boa parte do tempo faziam esse papel, Fernanda Orpinelli, Thiago Miller e tantos outros que compartilharam o ambiente do laboratório de bioinformática do IEP comigo.

Deixo um agradecimento também à minha orientadora do TCC, Dra. Angélica Maris, que me deu um leve empurrãozinho para a área de bioinformática e aqui hoje estou.

Agradeço muito a minha namorada Fabiana por me aguentar nos piores momentos de estresse e também nos piores momentos que tentava ser engraçado.

Agradeço ao Daniel, nosso amigo que sempre está disposto a nos ajudar em qualquer empecilho tecnológico que venha ocorrer e por manter tudo em ordem para realizarmos nossas tarefas.

Ao Programa Interunidades em Bioinformática, do Instituto de Matemática e Estatística e todos os seus membros por me proporcionarem um mundo novo de conhecimento e oportunidades.

Ao Instituto de Ensino e Pesquisa do Hospital Sírio-Libanês por proporcionar um ambiente onde não faltava nada para realizarmos nossa pesquisa.

À Banca de Qualificação, Banca Examinadora do Exame de Progresso em Pesquisa e à Banca Examinadora da Defesa da tese pela disponibilidade em ouvir um pouco sobre este tema pelo qual aprendi a gostar cada vez mais, pelas correções e pelos conselhos nestas etapas finais do doutorado.

Agradeço a tudo e a todos na verdade, pois ao final da graduação não imaginava que tantas pessoas poderiam ter influência na minha trajetória, fornecendo desde conselhos sobre congressos até as menores dicas que facilitam estar onde estou hoje e com toda a carga de conhecimento que possuo.

À CAPES, pela concessão de um Bolsa de Doutorado ao longo dos 4 anos de pesquisa.

Resumo

PIUCO R. Estudo dos PIWI-interacting RNAs (piRNAs): do desenvolvimento de uma base de dados à análise sistemática da expressão gênica em tecidos normais e tumorais. 2023. []55p. Programa Interunidade em Bioinformática. Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Os RNAs curtos, por definição, são moléculas de RNA que possuem comprimentos que variam de 20 a 200 nucleotídeos. Esses RNAs possuem funções relacionadas a uma ampla variedade de processos biológicos, incluindo a regulação da expressão gênica, desenvolvimento embrionário e diferenciação celular. Entre as classes de RNAs curtos estão os PIWI-interacting RNAs (piRNAs). Os piRNAs, RNAs de cerca de 24-32 nucleotídeos de comprimento, desempenham um papel fundamental na regulação da expressão gênica. Trabalhos recentes mostram que os piRNAs também apresentam funções relacionadas à carcinogênese (proliferação celular e invasão). Apesar desse conhecimento, existem poucos estudos sistemáticos sobre os piRNAs, mesmo em espécies com genomas e transcriptomas muito bem estudadas como *Homo sapiens*. Aqui, nesta tese, nós realizamos um estudo de amplo espectro dos piRNAs: iniciamos com a catalogação dos piRNAs e organização de um banco de dados, exploramos características marcantes desses RNAs curtos e, por fim, realizamos um estudo sistemático da expressão dos piRNAs em tecidos normais e tumorais. Para a etapa inicial, analisamos 16 *conjuntos de dados* de *small RNASeq e RIPseq* e obtivemos 200.123 sequências de piRNAs. Identificamos 1.162.670 eventos de pares alvo-piRNA segundo as mais recentes metodologias conhecidas, sendo que a maioria destes genes são do biotipo codificador de proteína. Ao final desenvolvemos um banco de dados, piRNAdb, para armazenar todas as informações processadas. Atualmente o piRNAdb está no topo da lista de ferramentas de buscas e é visitado cerca de 1000 vezes por mês. Sobre a expressão dos piRNAs, obtivemos dados de expressão (*small* RNA sequencing) de 2.046 amostras de tumores de mama, cólon, próstata e pâncreas. Encontramos que existem diversos piRNAs diferencialmente expressos neste tumores. Além disso, descobrimos que diversos *alvos* destes piRNAs, como o gene DPF2, estão descritos como tendo algum papel oncogênico ou com expressão alterada em cânceres. Portanto, neste trabalho nós estudamos um conjunto de RNAs curtos, os piRNAs, que são pouco estudados, mas de extrema importância para a regulação de diversos processos celulares. Acreditamos que estamos dando contribuições significativas para a área de piRNAs, com um banco de dados completo e também com essa análises detalhadas da expressão dos piRNAs.

Palavras-chave: RNAs curtos, *PIWI-interacting* RNA; piRNAs; Expressão Gênica; Câncer; Genômica; Bioinformática.

Abstract

Short RNAs are RNA molecules that range in length from 20 to 200 nucleotides. These RNAs have functions related to a wide variety of biological processes, including gene expression regulation, embryonic development and cell differentiation. Among the classes of short RNAs are the PIWI-interacting RNAs (piRNAs). PiRNAs, approximately 24-32 nucleotides in length, play a fundamental role in gene expression regulation. Recent studies have shown that piRNAs also have functions related to carcinogenesis. Despite this knowledge, there are few systematic studies on piRNAs, even in species with well-studied genomes and transcriptomes such as Homo sapiens. In this thesis, we conducted a comprehensive study of piRNAs. We began by cataloging piRNAs and organizing a database, explored distinctive characteristics of these short RNAs, and finally performed a systematic study of piRNA expression in normal and tumor tissues. For the initial step, we analyzed 16 sets of small RNASeq and RIPseq data and obtained 200,123 piRNA sequences. We identified 1,162,670 target-piRNA pairing events using the latest known methodologies. Finally, we developed a database, piRNAdb, to store all the processed information. Currently, piRNAdb is at the top of the search tools list and receives approximately 1000 visits per month. Regarding piRNA expression, we obtained expression data (small RNA sequencing) from 2,046 samples of breast, colon, prostate, and pancreatic tumors. We found that there are several differentially expressed piRNAs in these tumors. Furthermore, we discovered that several targets of these piRNAs, such as the DPF2 gene, are known to have an oncogenic role or altered expression in cancers. Therefore, in this work, we studied a group of short RNAs, piRNAs, which are poorly understood but extremely important for regulating various cellular processes. We believe that we are making significant contributions to the field of piRNAs with a comprehensive database and detailed analyses of piRNA expression.

Keywords: Short RNAs, *PIWI-interacting* RNA, piRNAs, Gene expression, Cancer, Genomics, Bioinformatics

Lista de Ilustrações

Lista de Tabelas

Sumário

1 Introdução

1.1 Os RNAs curtos

Os RNAs curtos ou pequenos (tamanho entre 20 e 200nt) são encontrados em organismos eucarióticos com diversas atividades na regulação da expressão gênica por induzirem a inibição da tradução e/ou a degradação do RNA mensageiro (de genes codificadores) (GHILDIYAL e ZAMORE, 2009). Existem diversos tipos de RNAs curtos, incluindo os microRNAs (miRNAs), os pequenos RNAs nucleolares (snoRNAs), os pequenos RNAs nucleares (snRNAs), os siRNAs (*short interfering* RNAs) e os os PIWI-interacting RNAs (piRNAs). Cada tipo de RNA curto desempenha funções diferentes, como iremos detalhar a seguir.

1.2 siRNAs

siRNAs foram descobertos inicialmente em plantas (HAMILTON e BAULCOMBE, 1999), porém mais tarde encontrado em amostras de organismos animais, esta classe de RNA curto está relacionada principalmente a servir como uma sequência guia indicando onde possivelmente ocorrerá a quebra do RNA alvo. Mais usada por facilitar as distinções biológicas inerentes entre os ncRNAs sua classificação pode se basear na identificação das moléculas que desencadeiam sua produção (HAMMOND et al., 2000). Os primeiros são desencadeados por sequências de dsRNA exógenos longos, onde são clivados em siRNA de fita dupla pela proteína Dicer. Cada fita desempenha um papel específico na regulação do seu alvo (GHILDIYAL e ZAMORE, 2009). Já a segunda ocorre em plantas, provavelmente pela intensa necessidade de defender-se de estresses causados por outros organismos ou pelo ambiente que a rodeia. Nestas, as fontes para a formação de siRNAs não fica

restrita aos dsRNAs, onde transcritos de fita simples gerados de regiões com repetições *in-tandem* e transgenes de cópia simples que são altamente expressados também podem fornecer substrato para a clivagem em dsRNA e ao final para siRNA (QU et al. 2005).

1.3 miRNAs

Descobertos inicialmente em nematóides *Caenorhabditis elegans* (BARTEL, 2004) e posteriormente também relacionados a importantes funções na regulação da expressão gênica e em outros processos celulares como a diferenciação celular (CHENG e MAHATO, 2011). Os miRNAs são similares aos outros RNAs curtos e apresentam entre 18 e 25 nucleotídeos em sua sequência (LUDVÍKOVÁ et al., 2015). Porém a quantidade diversa de sequências é considerável em relação aos outros componentes desta classe de RNAs curtos (VILELA et al. 2012).

A biogênese do miRNAs começa pela transcrição do gene (de miRNA), chamado de pri-miRNA, caracterizado pela grande abundância de estruturas em forma de *hairpin*. Em seguida, a proteína Drosha realizará as modificações necessárias ao transcrito que então será chamado de pré-miRNA, sendo então transportado para o citoplasma ainda com sua conformação de cadeia. A etapa seguinte consiste na atuação de um complexo formado, basicamente, pelas proteínas Drosha e Helicase e, que clivam a estrutura em um *duplex* de miRNA onde pouco tempo depois formará duas cadeias de miRNA individuais. Na etapa seguinte uma das cadeias formará o miRNA maduro que será ligado ao complexo RISC (*RNA-induced silencing complex*) e então ficará pronto para executar sua função previamente descrita enquanto que a fita de miRNA remanescente é, na maiorias das vezes, degradada (HAYES et al., 2014; LUDVÍKOVÁ et al., 2015).

1.4 snRNAs e snoRNAs

Os snRNAs (*small nuclear RNAs*) e snoRNAs (*small nucleolar RNAs*) são duas classes de RNAs curtos encontrados no núcleo celular e envolvidos na regulação da expressão gênica (para uma revisão, ver (KUFEL e GRZECHNIK 2019)). Em termo de funções, em resumo, os snRNAs são componentes importantes do spliceossomo, o complexo de proteínas e RNA que realiza o *splicing*. Neste processo, os snRNAs se associam a proteínas específicas para formar ribonucleoproteínas (snRNPs) que reconhecem e ligam-se aos exons e íntrons do pré-mRNA. Os snRNAs catalisam a remoção dos íntrons e a ligação dos exons, levando à formação do mRNA maduro, isto é, o *splicing* (KUFEL e GRZECHNIK, 2019).

Já os snoRNAs são encontrados no nucléolo e desempenham funções relacionadas, sobretudo, à modificação química de RNAs. Os snoRNAs se associam a proteínas específicas para formar ribonucleoproteínas (snoRNPs) que reconhecem e ligam-se a RNAs específicos. Os snoRNPs modificam a estrutura química desses RNAs, adicionando grupos metil ou pseudouridina em nucleotídeos específicos. Essas modificações são importantes para estabilizar os RNAs, regulando a expressão gênica e processando RNAs não codificantes. Os snoRNAs podem ser divididos em dois grupos principais, os snoRNAs guia e os snoRNAs de processamento. Os snoRNAs guia (ou snoRNAs de cajado) são responsáveis pela modificação química de RNAs ribossômicos, enquanto os snoRNAs de processamento são responsáveis pela modificação química de RNAs pequenos nucleares (snRNAs) e RNAs nucleares longos (lncRNAs) (KUFEL e GRZECHNIK, 2019).

1.5 piRNAs

A subfamília PIWI das proteínas argonautas, formada por AGO3, Aub e PIWI, é mais especificamente encontrada ligada aos piRNAs. Sua função na regulação de genes é significante pelo fato de estar relacionado com os processos biológicos de renovação e diferenciação de células células-tronco (GANGARAJU e LIN, 2009), desenvolvimento de células da linhagem germinativa (SAXE e LIN, 2011) e alguns tipos de câncer em humanos (ESTELLER, 2011), por exemplo.

Especificamente, os piRNAs são RNA curtos de 21 a 35 nucleotídeos (ARAVIN et al., 2006) encontrados exclusivamente em animais. A ausência de estrutura secundária e a metilação específica do oxigênio 2` na porção terminal 3` da sequência são outras características que distinguem os piRNAs de outros RNAs curtos não codificantes (MOAZED, 2009; SIOMI et al., 2011; OZATA et al., 2018; ARAVIN et al., 2006; OHARA et al., 2007).

Entre as funções mais conhecidas do piRNAs, estão o controle da expressão de elementos transponíveis (ETs) através da indução da degradação, alteração da cromatina e possível inibição da tradução daqueles ETs com regiões codificadoras (WATANABE e LIN, 2014). Como cerca de um quarto dos RNAs mensageiros humanos presentes em banco de referência com o RefSeq (PRUITT et al., 2005) possuem sequências derivadas de elementos transponíveis na região 3'UTR (e mesmo codificadoras), em teoria, tais genes podem ser regulados por piRNAs, porém há poucos relatos de tais fenômenos na literatura (GHILDIYAL e ZAMORE 2009). Linhagens germinativas e repressão de elementos transponíveis são o maior foco dos estudos em piRNAs hoje e isto se deve, em parte, pela identificação incial dos piRNAs neste tipo celular, além da primeira função definida para tais RNAs curtos, o da regulação destes ETs (SIOMI et al., 2011; ARAVIN et al., 2006). No entanto, o

interesse no estudo dos miRNAs no contexto de outros tecidos (somáticos) vem aumentado gradativamente. Hoje há caracterização de piRNAs atuando na regulação de outros processos biológicos, tais como o controle da expressão de mRNA e lncRNA (WATANABE e LIN, 2014; WATANABE et al., 2015), em processo de invasão, diferenciação e crescimento de células tumorais (cânceres de mama, gástricos e hepatocelulares (WATANABE e LIN, 2014; TAN et al., 2015)).

Em relação a origem do piRNAs, atualmente sabemos que tais RNAs curtos originam-se de regiões genômicas com alta densidade de piRNAs, chamados de *clusters*, e podem ser transcritos de apenas uma das fitas ou de ambas (uni- e bi-direcional, respectivamente) (YAMANAKA et al., 2014). Estes transcritos são selecionados e processados, sem uma ordem pré-definida, dentro das células e, como descrito, em uma grande quantidade (ISHIZU et al, 2012). Nishida e seus colaboradores em 2009 sugeriram que uma proteína, chamada *Tudor*, se liga com Aub ou AGO3 para facilitar a produção dos piRNAs, além de atuarem em uma espécie de controle de qualidade desse RNAs. Além disso, sabemos que a posição genômica de um *cluster* (de piRNAs) é conservada entre as espécies próximas, porém a composição dos piRNAs nesses *clusters* pode variar entre espécie e, até mesmo, entre indivíduos da mesma espécie (ROOVERS et al., 2015).

As funções desempenhadas pelo complexo piwi-piRNA foram descritas também em diversas outras espécies (por exemplo, porcos, saguis, *Drosophilas,* entre outras), indicando que sãi conservadas entre as espécies (GEBERT et al., 2015; HIRANO et al,. 2014). Interessantemente, cada proteína da família piwi apresenta a preferência por se ligar a piRNA de tamanhos específicos. Por exemplo, piRNAs de camundongos de aproximadamente 26 nucleotídeos se ligam preferencialmente a

MILI, já sequências com cerca de 28 nucleotídeos se ligam a MIWI2 e sequências com aproximadamente 30 nucleotídeos se ligam a MIWI (BORTVIN, 2013). Em *Drosophilas* sp. foi descrito um fenômeno semelhante (SIOMI et al., 2011).

Atualmente existem dois modelos que descrevem a produção de piRNAs. No primeiro, o transcrito de um *cluster* é ligado a proteína piwi gerando o terminal 5' contendo bases extras no terminal 3'. Este então será processado para remoção destas bases e adição da metilação característica deste RNA (KAWAOKA et al. 2011). O *cluster* transcrito pode sofrer essa primeira quebra em posições distintas, existindo o viés para o tamanho do piRNA dependente da proteína piwi (BORTVIN, 2013), como descrito anteriormente.

O segundo modelo de formação do repertório de piRNAs de um organismo é conhecido como "ciclo de amplificação *ping-pong*". Em resumo, esse ciclo começa com a formação de uma grande quantidade de piRNAs primários. No momento em que o complexo piwi-piRNA encontra seu alvo, a endonuclease Slicer realiza uma quebra do RNA alvo distante cerca de 10 bases do terminal 5' do piRNA (ISHIZU et al., 2012). Esse processo não somente inativa o RNA alvo, mas também fornece um transcrito com terminal 5' que se ligará a proteína AGO3, em drosófilas, ou MIWI2, em camundongos, para ser processado do mesmo modo que um piRNA primário. Por fim é gerado um piRNA maduro ativo que pode ser classificado como gênico (ZUO et al., 2016). Esse novo complexo poderá se ligar e quebrar os elementos complementares ao RNA guia, como transcritos de *clusters* de piRNA. O resultado dessa segunda quebra são vários novos fragmentos idênticos aos piRNAs originais da via primária. Possibilitando um ciclo de amplificação que produz muitos piRNAs a partir de poucas sequências iniciais (ARAVIN et al,. 2007).

Elementos transponíveis (ETs) são elementos repetitivos de DNA capazes de se mobilizar para uma nova localização genômica e encontram-se em frações grandes dos genomas dos eucariotos. A ativação e realocação de elementos transponíveis são permitidas e facilitadas pela perda da heterocromatina e hipometilação do genoma. O estado de repressão desses ETs requer a atividade contínua do complexo piwi-piRNA já que a atividade dos piRNAs rapidamente causa desrepressão dos ETs (YAMANAKA et al., 2014). A função dos piRNAs no silenciamento do ETs em *Drosophila* sp é semelhante ao já descrito acima, porém foi percebido que existe também uma deposição maternal de piRNAs nos ovos desse inseto, o que indica um sistema de silenciamento que pode ser herdado (MALONE et al., 2009). Le Thomas e seus colaboradores em 2014 identificaram que essa deposição de fatores citoplasmáticos é suficiente para ativar a geração de piRNA mesmo que não tenha sido herdado o *loci* ativo.

Existe um modelo proposto no qual os *clusters* de piRNA atuam como “armadilhas” para os elementos transponíveis. Devido à grande presença destes elementos truncados e porque a maioria dos piRNA que derivam desses locais, controlam a atividade dos elementos homólogos (YAMANAKA et al., 2014). Quando um elemento transponível se insere dentro de um desses *clusters* ele, em um processo de complexo de "auto controle", passa a contribuir com sequências para o repertório de piRNAs (CLARK e LAU, 2014). Este evento aumenta o repertório e a diversidade de piRNAs (YAMANAKA et al., 2014).

Em um contexto de câncer, estudos com linhagens de células tumorais estão sendo realizados e mostram função biológica que pode estar indiretamente associada aos piRNAs e processos de carcinogênese. Além disso, as proteínas piwi estão sendo

associadas com algumas outras marcas fenotípicas de câncer, chamadas *hallmarks*: como sustentar sinalizações para proliferação, defesa contra supressores de crescimento e ativação de invasão e de metástases em tumores (TAN et al., 2015). Apesar desses avanços, os detalhes moleculares que levam as atividades oncogênicas das piwi ainda precisam ser melhor esclarecidos. Em tumores de mana já foram realizadas pesquisas que identificaram piRNAs associados à metástase (ZHANG et al., 2013). E em tumores gástricos o crescimento tumoral foi inibido utilizando antagonistas de dois piRNAs in vivo (CHENG et al., 2012). Além do mais, os piRNAs vem sendo apontados como possíveis biomarcadores para diagnóstico e prognóstico menos invasivo (MARTINEZ et al., 2015). Por serem pequenos fragmentos com poucos nucleotídeos, não são degradados tão facilmente, como os RNAs longos, e também por suas habilidades para passar por membranas celulares. Esses piRNAs podem ser identificados em amostras de pacientes obtidas de uma maneira mais acessível em fluidos corporais como sangue e urina. Estudos utilizando essas metodologias já começam a ser realizados e com resultados promissores, como os possíveis biomarcadores para estadiamento de tumor gástrico piR-651 e piR-823 obtidos de amostras de sangue dos pacientes (CHENG et al., 2012). Essas características são benéficas para se desenvolver novas técnicas menos invasivas e melhorar a acurácia do diagnóstico de câncer em humanos (MEI et al., 2013). Além disso, células saudáveis podem expressar padrões de piRNAs mais similares a células tumorais do que células normais, e podem ser utilizados para detectar a doença até mesmo antes de manifestar-se (ASSUMPÇÃO et al., 2015).

1.6 Bancos de dados e estudos sistêmicos

O aumento na quantidade de sequências de piRNAs e caracterização dos mecanismos envolvidos na origem e função mostram a fundamental importância de possuir um banco de dados *web* estável, escalável e confiável. Existem outras iniciativas que procuram possuir estas qualidades, como piRNABank (LAKSHMI e AGRAWAL, 2008), RNAdb (PANG et al., 2005) e piRNAQuest (SARKAR et al., 2014), mas a falta de atualizações e acesso limitado da interface *web* destas ferramentas são problemas comuns para os usuários. O piRNA cluster database (ROSENKRANZ, 2016) foca seu sistema na entidade *cluster*, provendo mais uma fonte de informação a ser usada pelos pesquisadores, porém ainda restrita. Além disso, apesar da ampla utilização e disponibilidade de dados de sequenciamento de RNAs curtos, estes dados ainda não foram correlacionados de maneira organizada em bancos de dados de piRNAs. Por fim, na nossa opinião, existem poucas ou nenhuma real implementação de técnicas que fornecem ao usuário uma interface *user-friendly* nestes bancos de dados de piRNAs.

Ao final da revisão dos bancos de dados existentes, percebemos que havia espaço (na época) para se desenvolver um banco de dados estável, escalável e que contenha informações relevantes ao estudo dos piRNAs é necessário para o embasamento de outras pesquisas futuras. O acúmulo e organização sistemática de informações sobre a biogênese, função, modificação, expressão e conservação dos piRNAs é essencial para que se tenha uma maior clareza com relação as vias de ação, interações com as proteínas piwi e regulação de elementos transponíveis e não transponíveis pelos piRNAs.

Adicionalmente, apesar de existirem muitos dados de sequenciamento realizado por *small* RNA-seq ou miRNA-seq estarem disponíveis publicamente, muito pouco foi explorado para piRNAs. Desenvolver uma metodologia que selecione, dentre todos os RNAs curtos, aqueles que apresentam padrões conhecidos de piRNAs pode fornecer uma ferramenta muito poderosa para mineração de dados. Caracterizar melhor as funções de piRNA, bem como a sua associação com proteínas Piwi, permite o entendimento dos seus papéis nas vias biológicas de fertilidade, desenvolvimento de células-tronco, evolução dos genomas e biologia de cânceres (SUZUKI et al., 2012; SANA et al., 2012).

1.7 Classificação dos Tumores

A correta classificação do tecido canceroso desempenha um papel importante na batalha contra suas desenvolvimento e expansão de suas células, pois fornece uma ferramenta que pode ser utilizada pelo médico e seu paciente para definir certos aspectos da doença, como expectativa de prognose, probabilidade de superar a doença e, principalmente, determinar a melhor abordagem para o tratamento da doença do paciente (AMIN et al., 2017). Ainda além, realizar o estadiamento do tumor corretamente define em qual grupo o paciente será alocado em um estudo de triagem clínica e auxilia na análise de resultados em estudos clínicos (AMIN et al., 2017). Fornece ainda uma nomenclatura consistente desde a biologia do tumor até a apresentação clínica do mesmo para os membros envolvidos em pesquisas (AMIN et al., 2017).

Refinar os padrões a fim de fornecer o possível melhor sistema de estadiamento é um processo que nunca termina, segundo o American Joint Committee on Cancer (AJCC), que tem liderado esses esforços nos Estados Unidos

da América (EUA) desde 1959. Contanto com a colaboração do Union for International Cancer Control (UICC), o AJCC mantém um sistema que é utilizado em todo o mundo (AMIN et al., 2017). Esse sistema organiza a extensão da doença principalmente em características como a extensão do tumor primário, linfonodos regionais e presença de metástases distantes (EDGE e COMPTON, 2010).

1.7.1 Sistema TNM (AJCC e UICC)

Desenvolvido pelo AJCC em cooperação com a UICC. Essas duas organizações trabalharam juntas em todos os níveis para criar um esquema de estadiamento que é amplamente idêntico entre as duas organizações, embora existam algumas diferenças. (AMIN et al., 2017). É um processo para determinar a quantidade de células cancerosas bem como onde estão localizadas. Utiliza de suas características para descrever a gravidade da doença de um indivíduo com base na magnitude do tumor original (primário), bem como na extensão da disseminação do câncer no corpo. Essa compreensão do estágio tumoral auxilia os médicos a desenvolver um prognóstico e a elaborar um plano de tratamento para cada paciente. Tem por característica primordial uma linguagem comum para que os médicos se comuniquem de forma eficaz em relação à doença do paciente e colaborem com os melhores cursos de tratamento. Alguns dos métodos utilizados para obtenção de características para realizar o estadiamento segundo a metodologia proposta estão: i) Exames físicos, que determinam a localização e o tamanho do(s) tumor(es); ii) Exames de imagem, que também podem mostrar a localização do câncer, o tamanho do tumor além da situação do espalhamento tumoral; iii) Exames laboratoriais, fornecendo informações adicionais sobre sangue, urina e outros fluidos do tumor; iv)

Relatórios de patologia, geralmente confirmam o diagnóstico de câncer, bem como o estágio; v) Relatórios cirúrgicos, descrevem o tamanho e a aparência de um tumor e fornecem informações sobre o envolvimento de linfonodos e outros órgãos. Utiliza de, principalmente, informações anatômicas relacionadas à extensão do tumor primário, ao status dos linfonodos regionais e à presença ou ausência de metástases distantes para classificar a extensão da doença, chamado na maioria das vezes de "classificação TNM". (AMIN et al., 2017).

Especificamente o órgão AJCC considera para a classificação TNM como cada tipo de câncer é construído pela definição da extensão anatômica do câncer para o tumor (T), linfonodos (N) e metástases distantes (M), complementada em alguns casos com fatores não anatômicos. Para cada um dos T, N e M, há um conjunto de categorias, geralmente definidas por um número (por exemplo, T1, N2). Ao final, esses elementos são então combinados, de uma forma definida para cada tipo de câncer, e um estágio geral de 0, I, II, III, IV é atribuído (AMIN et al., 2017). Às vezes, esses estágios também são subdivididos, usando letras como IIIA e IIIB.

1.8 As oportunidades para se estudar os piRNAs

Portanto, fica claro que há certa carência de estudos de piRNAs e, sobretudo, que estes RNAs são extremamente importantes em diversos contextos fisiológicos normais e patológicos. Temos a meta inicial de unificar dados de diversos segmentos e estudos, abordar o seu contexto de conservação e, potencialmente funcional dos piRNAs, além do seu papel no câncer. Estamos aliando o poder computacional de integrar dados em larga escala à investigação de aspectos funcionais importantes piRNAs. Temos atualmente na literatura diversos trabalhos que demonstram a função e a importância de se estudar os piRNAs, entretanto, há poucos trabalhos mostrando

de forma ampla, integrando e analisando diversos dados desses RNAs curtos. Por exemplo, das poucas bases de dados existentes sobre os piRNAs, muitas são incompletas, desatualizadas ou apresentam informações conflitantes.

2 Objetivos

2.1 Objetivo geral

Coletar, processar e organizar dados diretos e indiretos de piRNAs em um banco de dados e investigar de forma sistemática a expressão de piRNAs e seus alvos em amostras normais e de diversos tipos de cânceres.

2.2 Objetivos específicos

i) Usar repositórios públicos de dados genômicos e de transcriptomas, além de dados funcionais e de anotação, para coletar informações sobre os piRNAs de humanos, roedores (*Mus musculus, Rattus norvegicus e Cricetulus griseus*), *Drosophila melanogaster* e *Caenorhabditis elegans*;

ii) Desenvolver um banco de dados relacional acessível pela *web* para armazenar e disponibilizar de forma eficiente e uso intuitivo informações diversas dos piRNAs, tais como localização genômica, expressão, alvos e integração com outros bancos de dados.

iii) Analisar diversas características destes piRNAs, tais como sua localização genômica, co-localização com regiões codificadoras e/ou elementos transponíveis, predição de alvos e padrões associados;

iv) Analisar a expressão e expressão diferencial dos piRNAs em amostras normais, tumores de mama, cólon, pâncreas e próstata através de dados de *miRNA-seq*, assim como investigar os possíveis genes alvo desses piRNAs nesses tumores.

3 Materiais e métodos

3.1 Dados públicos

Presente na figura 1 estão descritas todas as etapas realizadas para a construção do banco de dados, onde primeiro executamos o *download* de sequências de RNA curtas não codificadoras classificadas como piRNAs e com no máximo 60 bases no banco de dados NCBI Nucleotide e ENA de 6 organismos: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Cricetulus griseus*, *Drosophila melanogaster* e *Caenorhabditis elegans*. Também buscamos sequências providas como dados suplementares na literatura de piRNAs. *Datasets* de um mesmo organismo foram concatenadas e colapsadas, porém mantendo a referência para todos os autores. Para construir a funcionalidade "*CrossCodes*" coletamos códigos de acesso de piRNA e nomenclaturas dos bancos de dados acima citados, RNACentral (THE RNACENTRAL CONSORTIUM, 2017), RNAdb, piRNABank, piRBase e piRNAQuest.

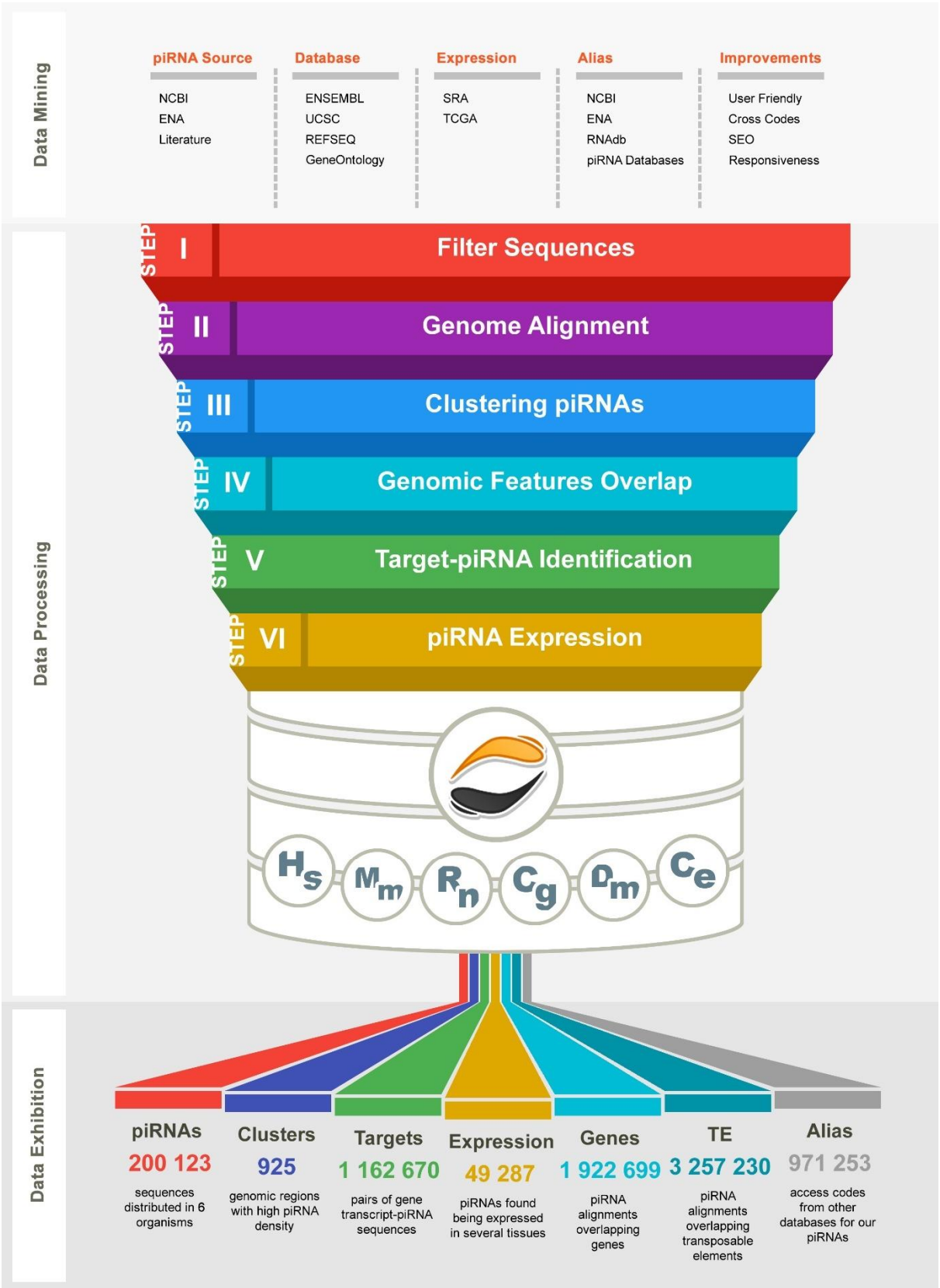

Figura 1 - Visão geral das etapas, fontes de informação e resultados encontrados durante o desenvolvimento do banco de dados piRNAdb. As sequências classificadas como piRNA foram obtidas e processadas pelas etapas da seção "Data Processing" (Processamento de dados). As anotações genômicas do genoma de referência, dos elementos gênicos e transponíveis foram obtidas na seção "Banco de dados". Além disso, obtivemos o

transcriptoma das espécies para a etapa de predição de alvos, bem como seus termos de genes do GeneOntology. As amostras para a avaliação da expressão foram coletadas nas fontes indicadas na seção "Expressão". Os códigos de acesso aos piRNAs de outros bancos de dados foram obtidos nos itens da seção "Alias". Na posição extrema direita estão atributos como responsividade, que proporcionam uma interface amigável aos nossos usuários. Dados obtidos para 6 organismos: Homo sapiens, Mus musculus, Rattus norvegicus, Cricetulus griseus, Drosophila melanogaster e Caenorhabditis elegans. A seção inferior "Data Exhibition" contém os resultados encontrados no processamento de todas as informações coletadas por nosso protocolo de etapas e disponíveis para os usuários do piRNAdb.

Para executar o alinhamento e anotação do genoma obtivemos, para *H. sapiens*, *M. musculus*, *R. norvegicus*, *C. griseus*, *D. melanogaster* and *C. elegans,* o genoma e transcriptoma do UCSC (RELEASE 84; build: hg38, mm10, rn6, criGri1, dm6 e ce11, respectivamente) (CASPER et al., 2018) and RefSeq (O'LEARY et al., 2016). Coordenadas genômicas foram obtidas do Ensembl (RELEASE 93; build: GRCh38, GRCm38, Rnor6.0, CriGri_1.0, BDGP6 e WBcel235, respectivamente) (ZERBINO et al., 2018). Também do UCSC nós obtivemos as anotações de coordenadas de elementos transponíveis e repetitivos.

Por fim, para avaliar a expressão de piRNAs, nós baixamos *datasets* públicos de *miRNA-seq* e *ncRNA-seq* do banco de dados SRA (NCBI RESOURCE COORDINATORS, 2017) para *M. musculus*, *R. norvegicus*, *C. griseus*, *D. melanogaster* e *C. elegans*. Enquanto que para humanos, obtivemos dados de *miRNA-seq* de amostras normais e tumorais de 4 tecidos do The Cancer Genome Atlas (TCGA).

3.2 Mapeando e caracterizando os piRNAs

Utilizamos o software de alinhamento BWA aln (LI e DURBIN, 2010) (parâmetro: -n 0), não permitindo *mismatch* ou *gap* e rejeitando alinhamentos parciais do piRNA. Devido à grande quantidade de alinhamentos em comparação a quantidade

inicial de sequências de piRNA, investigamos quanto cada piRNA contribui para a enorme quantidade de alinhamentos. Distribuímos os piRNAs em 3 grupos: I) alinhamento único; II) entre 2 e 49 alinhamentos e III) 50 alinhamentos ou mais.

Aplicamos a função IntersectBed do *software* BedTools (QUINLAN e HALL, 2010) para quantificar a sobreposição entre alinhamentos de piRNA e anotações genômicas de: I) genes e II) elementos transponíveis. Parâmetros da função foram: -s -f 1.00 -r -wao. Os resultados foram processados por scripts em PERL e enviados para o banco de dados piRNAdb. Então, identificamos quais eram os biotipos das anotações gênicas (*protein-coding*, *long* e *short non-coding* RNA, pseudogene e outros) que ocorreram com maior frequência.

A classificação em *clusters* proporciona evidências sobre a transcrição e auxilia na avaliação da conservação entre as espécies. Nós aplicamos a metodologia de janela deslizante (Tamanho da janela: 20kb; deslizamento: 1kb; Densidade 7 piRNAs). A fim de evitar problemas para definir a real localização de origem do piRNA, nós selecionamos apenas aqueles que se alinham unicamente ao genoma. Finalmente, observamos a sobreposição dos *clusters* e os identificamos em uni- ou bi-direcional.

Utilizando as regras recentes de refinamento e análises implicadas na complementaridade entre um piRNA e o transcriptoma do organismo, nós pudemos predizer os transcritos alvos e proporcionar maior quantidade de informação biologicamente relevante ao piRNAdb. Entre essas regras usadas estão o requerimento de uma porção *seed* (SHEN et al., 2018), no máximo 6 *mismatches* fora da região de *seed* (WU et al., 2018) e é aceitável a ocorrência de pares GU (ZHANG et al., 2018). Nesta etapa, utilizamos uma versão modificada do alinhador sRNAmapper (ROSENKRANZ et al., 2015) conhecida como sRNAtargetmapper.pl (parâmetros: -a all -n 0 -m 6 -s 8 -S 0 -c 1 -G 0 -p). Ainda além, nós requeremos que

o número de bases complementares cobrisse pelo menos 75% do tamanho do piRNA para evitar o viés ocorrido pela variabilidade no tamanho do piRNA. Por fim, identificamos quais eram os biótipos de *targets* (*protein-coding*, *long* e *short non-coding* RNA, pseudogene e outros) que mais ocorrem em nossos dados, além de buscar avaliar quantos alinhamentos que cada classe possui.

As informações relativas a identificação de piRNA dos 16 *datasets* foram adicionadas ao banco de dados do piRNAdb e estão acessíveis aos usuários. Buscaremos colaborações para validar experimentalmente alguns destes candidatos a piRNAs e *clusters*. Além de buscar parcerias com outros bancos de dados que podem fornecer metodologias para seleção de *clusters*, por exemplo, onde trabalhamos em conjunto desenvolvendo os dois sistemas ao mesmo tempo.

3.3 Análise do perfil de expressão de piRNAs

Para avaliação da expressão de piRNAs em tecidos tumorais obtivemos dados de sequenciamento de RNAs curtos de 465 amostras de adenocarcinoma de cólon (COAD), 1148 de carcinoma invasivo de mama (BRCA), 435 de adenocarcinoma de próstata e 145 de adenocarcinoma de pâncreas vindos do TCGA. Nos tecidos de mama, cólon e pâncreas foram obtidas as informações de classificação patológica em estágios do tumor: estágio I, estágio II, estágio III e estágio IV. Apenas para o tecido de próstata esta informação não foi obtida por não constar diretamente nos dados fornecidos pelo TCGA. Este método classifica cada amostra de tumor segundo tamanho do tumor (T), invasão de linfonodos (N) e presença de metástase em outros tecidos (M). Ao final, todas as características de TNM são compiladas e a amostra é classificada com os estágios apresentados.

Para humanos, comparamos e contamos as sequências obtidas com uma lista de piRNAs já anotados e aplicamos um filtro removendo as entidades que possuem menos que um *read* em menos de 10% das amostras (MARTINEZ et al., 2015), etapa feita para cada tecido separadamente. Então, usamos o pacote edgeR (ROBINSON et al., 2010) do *software* R, para normalizar a quantidade de *reads* segundo o tamanho da biblioteca do sequenciamento com a metodologia *Trimmed Mean of M-values* (TMM).

A fim de realizar comparações em relação aos estágios, realizamos a soma da quantificação de piRNAs de todas as amostras em seus respectivos estágios. Então as seguintes categorias de amostras foram separadas: I) Normal *versus* Tumor; II) Normal *versus* Estágio I; III) Estágio I *versus* Estágio II; IV) Estágio II *versus* Estágio III e V) Estágio III *versus* Estágio IV. Após a seleção destes piRNAs continuaremos utilizando o edgeR para quantificação da expressão diferencial dos piRNAs entre as condições analisadas. Apenas os piRNAs que apresentarem logFC maior ou igual a 1 ou menor ou igual a -1, além de p-valor ajustado menor do que 0.05 serão considerados diferencialmente expressos. Por fim, os piRNAs anotados e diferencialmente expressos foram pesquisados no banco de dados piRNAdb para identificação de quais *targets* os alinhamentos do mesmo estão associados.

3.4 Desenvolvimento do banco de dados e ferramenta web

O banco de dados piRNAdb é uma plataforma aberta de uso gratuito e *user-friendly,* foi projetado e construído ao longo dos passos iniciais desta pesquisa. Como ainda não existe uma nomenclatura padrão que pode ser utilizada para identificar cada piRNA, adotamos um padrão de nomenclatura similar a dos microRNAs, a qual já está bem estabelecida. Então, todas as sequências de piRNA do nosso banco de dados receberam um código de três letras identificando o organismo, o termo ‘piR’ e um

código numérico único: hsa-piR-9, por exemplo. Já os *clusters* receberam algo muito similar: hsa-clu-12, por exemplo.

Inovações que não estão presentes em outros bancos de dados e podem ajudar na melhor fundamentação de ideias e possibilitar a comunicação entre pesquisadores foram desenvolvidas. Uma delas é a possibilidade de interagir com o banco de dados por um sistema de *Feedback,* onde poderá ser votado por cada usuário se um piRNA pode ser considerado verdadeiro ou não, e contando ainda com um sistema de comentários que podem ser publicados explicitando a opinião de forma textual.

piRNAdb foi desenvolvido utilizando a plataforma MySQL devido a sua garantia de integridade de dados e relacionamento entre tabela. A interface gráfica foi desenvolvida usando *HyperText Markup Language 5* (HTML5), com funções em *JavaScript* e *Asynchronous JavaScript* (AJAX) a fim de reduzir o tempo de carregamento das páginas e proporcionar interatividade ao usuário. Desenvolvemos o *layout* do piRNAdb usando *Cascading Style Sheets 3* (CSS3) e *Grids*, que instruem o navegador a renderizar a página *web* de acordo com a resolução de tela do usuário. A comunicação entre o banco de dados e o processamento dos dados é controlado usando a linguagem *PHP Hypertext Preprocessor* (PHP). Todas as linguagens mencionadas acima são conectadas pela arquitetura *Model-View-Controller* (MVC). Nós também tomamos cuidado em proporcionar técnicas de *Search Engine Optimization* (SEO) e *HTML* semântico para fornecer aos mecanismos de buscas, como Google e Bing, instruções para considerar alguns tópicos específicos, seções e páginas mais significantes para o usuário. No sentido de universalizar e agilizar o acesso pelos usuários, o piRNAdb foi desenvolvido seguindo regras de

responsividade aos principais navegadores, sendo disponível também para *tablets* e *smartphones*.

4 Resultados

4.1 Banco de dados piRNAdb

4.1.1 Características gerais

A publicação do *RELEASE* 1 do piRNAdb, acessível em https://www.pirnadb.org, conta com 16 *datasets* de piRNAs para 6 organismos diferentes (Tabela 1). Agrupando-os, possuem 236.183 sequências de piRNAs identificadas pelos autores e seus colaboradores (Tabela 1). piRNAs do mesmo organismo encontrados em *datasets* diferentes foram colapsados, mas com as referências mantidas para todos os autores.

Tabela 1 - Lista dos conjuntos de dados obtidos de bancos de dados públicos e literatura especializada contendo sequências de RNAs curtos classificados pelos autores como piRNAs. Dados disponíveis, após filtragem, no banco de dados piRNAdb.org.

Organism	Methodology	Author	PubMed ID
H.sapiens	Immunoprecipitation	Girard et al., 2006	16751776
H.sapiens	Immunoprecipitation	Aravin et al., 2006	16751777
H.sapiens	Small RNA-seq	Krawetz et al., 2011	21989093
H.sapiens	Small RNA-seq	Li et al., 2012	22313525
H.sapiens	Small RNA-seq	Mei et al., 2015	26095918
H.sapiens	Small RNA-seq	Rizzo et al., 2016	27429044
H.sapiens	Small RNA-seq	Law et al., 2013	23376363
M. musculus	Immunoprecipitation	Girard et al., 2006	16751776
M. musculus	Small RNA-seq	Lau et al., 2006	16778019
R. norvegicus	Immunoprecipitation	Girard et al., 2006	16751776
R. norvegicus	Small RNA-seq	Lau et al., 2006	16778019
C. griseus	Small RNA-seq	Gerstl et al., 2013	23639388
D. melanogaster	Immunoprecipitation	Nishida et al., 2007	17872506
D. melanogaster	Immunoprecipitation	Brennecke et al., 2007	17346786
C. elegans	RNA-seq	C. elegans Sequencing Consortium, 1998	9851916
C. elegans	Small RNA-seq	Kato et al., 2009	19460142

Fonte: Elaborada pelo autor

Após alinhar estas sequências ao genoma de referência em cada espécie, obtivemos 8.556.265 alinhamentos no total, onde *R. norvegicus* e *M. musculus* resultaram na maior quantidade de alinhamentos, aproximadamente 3,4 e 3,1 milhões, respectivamente. Enquanto que *C. elegans* mostrou a menor quantidade, apenas 15994. *H. sapiens* tem 812384 alinhamentos de piRNAs ao seu genoma. Ao aprofundar as análises relativas às características de posicionamento genômico, encontramos na tabela 2 que cerca de 70% das sequências de piRNA alinham apenas uma vez ao genoma, exceto para *D. melanogaster* com 23,54%. Em *M. musculus* e *R. norvegicus* este valor aumenta para perto de 90% e *C. elegans* tem quase todos os piRNAs alinhando apenas uma vez ao genoma (99,57%). Enquanto que *H. sapiens* e *M. musculus* 1,29% e 1,12%, respectivamente, dos piRNAs alinham 50 vezes ou mais, para *R. norvegicus* a quantidade é ainda menor, apenas 0,49%. *C. elegans* não apresentou nenhum piRNA com 50 alinhamentos ou mais nesta análise. Opostamente a *D. melanogaster*, que apresentou a maior quantidade de piRNAs com esta característica, 12,36% dos seus piRNAs.

Tabela 2 - Proporção da quantidade de piRNAs que possuem apenas um alinhamento, de 2 a 49 alinhamentos e 50 ou mais alinhamentos ao genoma de referência de cada espécie.

piRNA genome mapping amount	Build Version (% of total)					
	hg38	mm10	rn6	crigri1	dm6	ce11
Unique	22464 (81.10)	49315 (89.88)	47630 (87.43)	18223 (71.18)	5079 (23.55)	15845 (99.57)
2-49	4876 (17.60)	4932 (8.99)	6573 (12.07)	6881 (26.88)	13824 (64.09)	68 (0.43)
50+	360 (1.30)	618 (1.13)	272 (0.50)	496 (1.94)	2667 (12.36)	0 (0.00)

Fonte: Elaborada pelo autor

Ao final da análise, encontramos 36.060 sequências de piRNA sem nenhum alinhamento ao genoma de referência. Após remover estas sequencias não alinhadas, o nosso banco de dados (piRNAdb) ficou com 200.123 sequências de piRNA (Tabela 3). Roedores contribuem em maior quantidade (*M. musculus*: 54.865;

R. norvegicus: 54.475 e *C. griseus*: 25.600). *H. sapiens* possuem 27.700 piRNAs encontrados em 7 conjuntos de dados após todas as etapas de filtragem. Enquanto que *D. melanogaster* apresentou 21.570 sequências armazenadas e *C. elegans* tem (apenas) 15.913 elementos.

Tabela 3 - Quantidade de sequências de piRNAs encontrados antes e depois da remoção das entidades sem alinhamento ao genoma. Dados separados por espécie.

Organism	piRNA Sequences Collected	Available at piRNAdb
Homo sapiens	32804	27700
Mus musculus	72747	54865
Rattus norvegicus	66758	54475
Cricetulus griseus	25626	25600
Drosophila melanogaster	22333	21570
Caenorhabditis elegans	15915	15913
Total	236183	200123

Fonte: Elaborada pelo autor

Com os dados dos alinhamentos disponíveis, foi possível executar os próximos dois passos da metodologia da construção do piRNAdb. O primeiro relacionado à identificação dos *clusters* de piRNA. Onde identificamos 160 *clusters* para *H. sapiens* (Tabela 4). *C. elegans* mostrou a menor quantidade de *clusters* de piRNA, apenas 40, enquanto que *C. griseus* obteve a maior quantidade, 379 *clusters*. Este último organismo também possui a maior quantidade de *clusters* bi-direcionais, 73, em comparação com os outros organismos, mas quando analisada a quantidade relativa ao total de *clusters* em cada organismo, *C. elegans* dispara com 33 de seus 40 *clusters* sendo bi-direcionais.

Tabela 4 - Distribuição da quantidade de clusters identificados pela metodologia de janela deslizante, sendo separado pela característica de ocorrência nas fitas genômicas.

Organism	Uni-Directional		Bi-Directional	Total
	Plus Strand	Minus Strand		
H. sapiens	62	69	29	160
M. musculus	66	75	12	153
R. norvegicus	53	70	14	137
C. griseus	147	159	73	379
D. melanogaster	31	15	10	56
C. elegans	3	4	33	40
Total	362	392	171	925

Fonte: Elaborada pelo autor

O segundo passo foi buscar os genes associados aos piRNAs, etapa executada buscando alinhamentos que estejam se sobrepondo ao posicionamento de genes anotados no Ensembl. Próximo de 22,5% (1.922.699) de todos os alinhamentos foram encontrados em sobreposição a coordenadas gênicas (Tabela 5). Onde *M. musculus* e *R. norvegicus* mostraram a maior quantidade de genes associados, 815718 e 549839, respectivamente. *H. sapiens* tem 277.704, enquanto que *C. elegans* exibe apenas 3.111 alinhamentos de piRNAs se sobrepondo a anotações gênicas. Indo além nas análises, agrupamos os genes associados segundo sua classificação de biotipo. Enfatizamos aqui os genes codificadores de proteína que mostraram estar em maior quantidade quando observado os alinhamentos de piRNAs em todos os organismos. Em relação ao segundo mais abundante biotipo existe uma divisão, onde *H. sapiens*, *M. musculus*, *R. norvegicus* e *D. melanogaster* possuem um *long non-coding* RNA (lncRNA) com maior frequência, enquanto que *C. elegans* possui pseudogenes e *C. griseus* possui *short non-coding* RNA (ncRNA).

Tabela 5 - Avaliação da quantidade de genes sobrepostos a alinhamentos de piRNAs segundo a classificação de biotipo do gene.

	Organism Build Version						Total
Biotype	**hg38**	**mm10**	**rn6**	**crigri1**	**dm6**	**ce11**	
lncRNA	54987	127369	29937	5425	6709	74	224501
ncRNA	761	731	321	11491	3645	11	16960
Pseudogene	19737	16078	1201	327	1131	123	38597
Protein Coding	202219	671540	518380	132096	115503	2903	1642641
Total	277704	815718	549839	149339	126988	3111	1922699

Fonte: Elaborada pelo autor

A mesma metodologia foi empregada para realizar a sobreposição dos elementos transponíveis, entretanto os grupos seguem a classificação fornecida pela UCSC (Tabela 6). *C. griseus* não foi avaliado nesta etapa devido à falta de anotação oficial para este tipo de elemento. Para *H. sapiens* e roedores, há mais sobreposição de alinhamento a elementos transponíveis quando comparados a coordenadas gênicas. Em *H. sapiens*, 79,6% de 380.948 alinhamentos sobrepostos localizam-se em regiões de *Short Interspersed Nuclear Element* (SINE) (30.3351 alinhamentos), seguidos por *Long Interspersed Nuclear Element* (LINE) (35023 alinhamentos). Para *M. musculus* e R*. norvegicus* a família TE mais frequente nesta análise é a "*Simple Repeats*", que mostrou aproximadamente 1 milhão de sobreposições de alinhamento em cada organismo. Apenas alguns alinhamentos em elementos transponíveis foram encontrados em *C. elegans* e *D. melanogaster*, 95 e 8626, respectivamente.

Ao concatenar os resultados encontrados para genes e elementos transponíveis, buscamos aqueles piRNAs que possuem alinhamentos tanto sobre genes quanto TE ao mesmo tempo (Tabela 7). Encontramos que os humanos possuem a maior quantidade desta categoria (1.620), seguido por camundongo e rato

(1242 e 489, respectivamente) e ao final com a menor quantidade *C. elegans* com apenas 11 piRNAs com essa característica.

Tabela 6 - Avaliação da quantidade de elementos transponíveis sobrepostos a alinhamentos de piRNAs segundo a classificação de famílias. Asteriscos (*) indicam a ausência de elementos encontrados nos resultados.

	Build Version				
Family	**hg38**	**mm10**	**rn6**	**dm6**	**ce11**
DNA	491	294	330	3443	40
LINE	35023	340409	170056	1470	2
Low_complexity	1	37	4471	*	7
LTR	10104	150929	67418	3570	1
RC	1	*	*	*	14
Retroposon	28103	*	*	*	1
RNA	65	19	*	*	*
rRNA	332	141	202	*	*
Satellite	2402	77	533	130	*
scRNA	157	42	*	*	*
Simple_repeat	38	809295	1045244	*	27
SINE	303351	104583	169780	*	*
snRNA	86	72	*	*	*
srpRNA	169	12	68	*	*
tRNA	623	1	305	*	*
Unknown	1	134	2936	*	2
Other	*	38	135	13	*
Total	380947	1406083	1461478	8626	94

Fonte: Elaborada pelo autor

Tabela 7 - Quantificação de piRNAs que possuem alinhamentos sobrepostos a genes e elementos transponíveis ao mesmo tempo.

	Build Version				
	hg38	**mm10**	**rn6**	**dm6**	**ce11**
Total	1620	1242	489	149	16

Fonte: Elaborada pelo autor

Nós encontramos 1.162.670 eventos de pares alvo-piRNA preditos, nos quais quase 43% (494.899) ocorrem em *H. sapiens*, seguido por *C. elegans* com 280.492

pares. O roedor *C. griseus* apresenta o menor número de pares, apenas 1.056 (Tabela 8). Na próxima etapa, avaliamos o biótipo de genes alvo de piRNAs. Para todos os organismos, o biótipo mais frequente foi o codificador de proteínas, sendo uma ordem de magnitude maior que o *lncRNA*, o segundo mais frequente.

Tabela 8 - Distribuição da quantidade de genes alvos de piRNAs preditos separados segundo a classificação em biótipo dos genes. Asteriscos (*) indicam a ausência de elementos encontrados nos resultados.

Gene Biotype	Amount of targets					
	H. sapiens	*M. musculus*	*R. norvegicus*	*C. griseus*	*D. melanogaster*	*C. elegans*
protein coding	10538	14809	14298	203	8137	18732
lncRNA	2592	1729	108	10	921	5112
smallRNA	79	8	3	3	11	31
pseudogene	572	223	18	*	*	*
unknown	2	*	1	*	2	*
Total	13783	16769	14428	216	9071	23875

Fonte: Elaborada pelo autor

No final, utilizando os dados obtidos de sequenciamentos de RNA curtos obtidos de bancos de dados públicos encontramos informações relativas à abundância de 49.287 seqüências piRNA anotadas no piRNAdb, onde *R. norvegicus* exibem a maior quantidade de piRNAs encontrados na etapa de expressão (28.795), seguido por *D. melanogaster* (13.738). Poucos foram encontrados nos tecidos de *H. sapiens* e *M. musculus*, 259 e 114, respectivamente. Um *boxplot* mostrando a abundância em cada tecido e condição é exibido para os usuários abaixo da caixa de alinhamento.

4.1.2 Acesso e o sistema *web* do piRNAdb

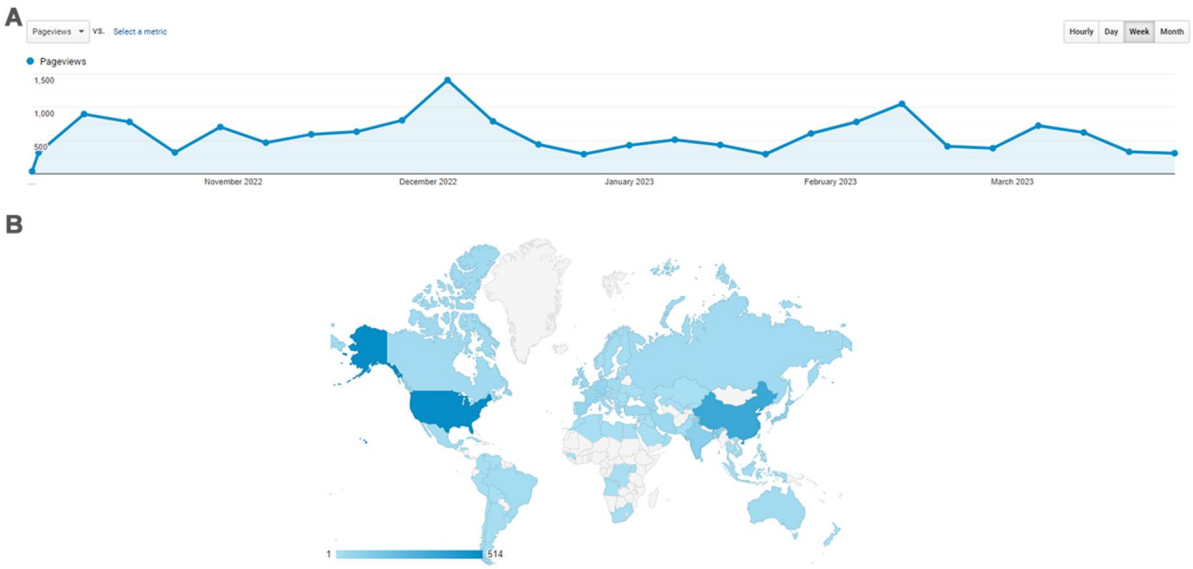

Figura 2 - Métricas de avaliação dos acessos nos últimos 6 meses (outubro de 2022 a abril de 2023) do banco de dados piRNAdb. (A) Distribuição da quantidade de pageviews do banco de dados agrupado por semana. (B) Distribuição geográfica dos acessos ao nosso banco de dados.

O sistema *web* do piRNAdb oferece uma ferramenta poderosa de armazenamento e recuperação de informações relativas aos piRNAs. A utilização em larga escala de técnicas de melhoramento de posicionamento em motores de busca e semântica do conteúdo ajudou na indexação do conteúdo e posicionamento. Já estamos na segunda posição da busca orgânica (resultados que descontam os links patrocinados) do Google. Sendo que alternadamente somos posicionados como o primeiro colocado. Recebemos cerca de 1.800 novos usuários nos últimos 6 meses e (Figura 2A) com uma média de 580 *pageviews* por semana, 15.337 *pageviews* no total do período avaliado, sendo que estas visitas são provenientes de diversos países (Figura 2B) como Estados Unidos da América, China, Índia, Japão e Alemanha (514, 347, 100, 83 e 65 usuários no período, respectivamente).

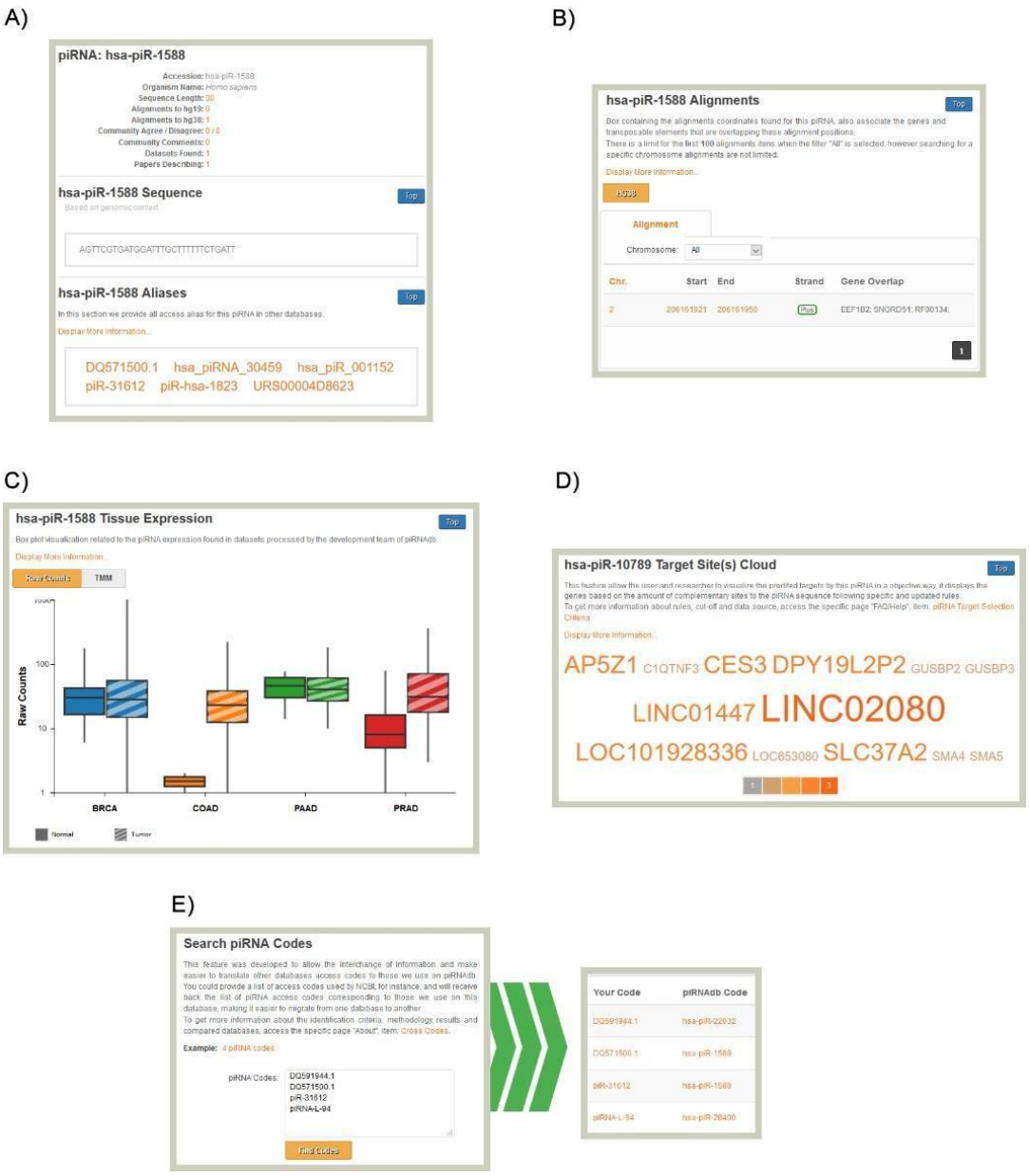

Figura 3 - Seções da página de informações do piRNA e do recurso “CrossCodes”. (A) exibe várias informações sobre o piRNA específico (has-piR-1588, neste caso) como o comprimento da sequência, a quantidade de alinhamento e o alias para outros códigos de acesso a bancos de dados. (B) quantidade, localização genômica e possíveis características de sobreposição associadas ao alinhamento do piRNA ao genoma disponível (hg38, neste caso). Um link para o UCSC Genome Browser está disponível em cada alinhamento. (C) Visualização de abundância de boxplot, em valores de TMM, mas também contagens brutas disponíveis, encontradas para este piRNA específico em vários tecidos e condições. Por exemplo, em humanos a abundância é mostrada para tecidos normais e tumorais de 4 tipos de tumor TCGA. (D) Formato de nuvem de tags exibindo os alvos putativos para o piRNA has-piR-10789. Um tamanho de fonte maior e tom laranja forte indicam mais locais de complementaridade preditos por nossas etapas de processamento. (E) seção de recurso “CrossCodes”, encontrada na página “Search”. Vários códigos de acesso a bancos de dados de piRNA podem ser traduzidos para códigos de acesso oficiais piRNAdb e redirecionamento para a página piRNA, fornecendo informações mais diversificadas para o usuário que migra

para nosso banco de dados. Um exemplo de códigos de acesso à esquerda e o resultado, códigos oficiais piRNAdb, à direita. Fonte: Elaborada pelo autor

Atualmente, o campo não possui uma nomenclatura padrão para piRNAs. A fim de contribuir para uma padronização, nomeamos piRNAs seguindo um padrão similar usado em miRBase (KOZOMARA e GRIFFITHS-JONES, 2014). Atribuímos aos piRNAs um código único formado pela inicial relacionada ao organismo (por exemplo: hsa, para humanos), a palavra "piR" e um código numérico (por exemplo, hsa-piR-5). Os *clusters* identificados também receberam códigos semelhantes formados pelo inicial do organismo, a palavra "clu" e um código numérico (por exemplo, hsa-clu-17).

Existem 3 categorias de páginas principais que fornecem informações piRNAs: I) cluster; II) *datasets* e III) piRNA. O último contém subseções, principalmente elementos gráficos, que exibem informações relevantes sobre os piRNAs, como: I) seqüências de nucleotídeos (Figura 3A); II) aliases do mesmo piRNA em outras bases de dados (Figura 3A); III) Caixa de alinhamento e sua sobreposição a elementos genômicos (Figura 3B); IV) *boxplot* de expressão do piRNA (Figura 3C); V) gene alvo predito exibido no formato de nuvem de tags (Figura 3D); VI) gráfico de barras dos termos de geneOntology associado aos genes alvos mencionados; VII) formulário de *feedback*. Páginas específicas para *clusters* e *datasets* incluem informações sobre características gerais, mas também fornecem uma lista de piRNAs que ocorrem em cada categoria.

Páginas de informações específicas do piRNA podem ser acessadas usando a seção “*Browse*” ou pelos mecanismos de busca fornecidos pela seção “*Search*”. As opções de busca incluem (Figura 3C) o uso de código de acesso, coordenada genômica, código do gene associado e gene alvo de interesse. Especificamente, na busca de *clusters*, existe a possibilidade de procurar aqueles que ocorrem em uma

coordenada do genoma desejado. Para facilitar a migração e o acesso das informações pelo usuário, desenvolvemos uma ferramenta conhecida como "*CrossCodes*" (Figura 3E), disponível na página "*Search*". Nele, o usuário é capaz de fornecer os códigos de acesso do piRNA de outros bancos de dados ou publicações e, em seguida, a ferramenta converterá o código fornecido no código oficial usado no piRNAdb, também um link para a página de informações do piRNA.

Todos os dados armazenados nesse sistema podem ser baixados dentro da página "*Downloads*" em diversos formatos. Informações com o intuito de auxiliar a navegação e esclarecer dúvidas sobre o funcionamento do sistema estão em tópicos dentro da seção "*FAQ/Help*".

4.2 Identificação do perfil de expressão

Obtivemos acesso a 2.198 amostras de sequenciamento de RNAs curtos de 166 amostras normais e 2.026 tumorais (Tabela 9). Destas a maior parte é vinda de amostras de mama (BRCA), seguida por cólon (COAD) e próstata (PRAD) com quantidade similares e pâncreas (PAAD) em menor quantidade de amostras (1148, 465, 434 e 145 amostras, respectivamente). De acordo com informações patológicas obtidas do mesmo banco de dados para as amostras coletadas, realizamos a separação das amostras em estágios. Neste momento, as amostras de próstata não puderam ser classificadas por não possuírem esta classificação explícita no banco de dados TCGA. Entretanto, nos tecidos restantes podemos perceber que a maioria das amostras é classificada em estágio II. Para mama e cólon a segunda maior categoria é o estágio III (235 e 129 amostras, respectivamente), contudo para pâncreas a

segunda maior categoria são amostras de estágio I. Nos 3 tecidos a menor quantidade de amostras está no estágio IV.

Tabela 9 - Distribuição da quantidade de amostras tumorais e normais obtidas do banco de dados TCGA para 4 tecidos separados segundo a classificação patológica de estágios disponível no mesmo banco de dados. BRCA: adenocarcinoma de mama; COAD: adenocarcinoma de cólon; PRAD: adenocarcinoma de próstata; PAAD: adenocarcinoma de pâncreas. Asterisco (*) indica que não foram encontradas informações para determinado estágio nos dados do TCGA.

Sample Stage	BRCA	COAD	PRAD	PAAD
Normal	103	8	52	3
I	173	75	*	13
II	594	175	*	124
III	235	129	*	2
IV	17	65	*	3
Tumor	1045	457	382	142

Fonte: Elaborada pelo autor

Após realizar a contagem de piRNAs já conhecidos, anotados e disponíveis no banco de dados piRNAdb (Tabela 10), encontramos que todos os tecidos possuem maior quantidade de piRNAs anotados nas amostras tumorais em relação às amostras normais. Quando comparado os estágios entre si e com amostras tumorais, para mama e cólon, os valores ficam similares, porém, para pâncreas a quantidade de piRNAs anotados encontrados é variável, visto que os estágios III e IV possuem menor quantidade quando comparados aos estágios I, II e tumoral, possuindo inclusive menor quantidade que as amostras normais. Já quando comparado às amostras tumorais entre os 4 tecidos, houve uma grande igualdade, ficando todos com valores próximos, porém ao fazer a mesma comparação com as amostras normais houve uma menor igualdade nos valores de piRNAs, principalmente ao considerar as amostras de próstata.

Tabela 10 - Distribuição da quantidade de piRNAs conhecidos e anotados encontrados nas amostras de tumores obtidas. BRCA: adenocarcinoma de mama; COAD: adenocarcinoma de cólon; PRAD: adenocarcinoma de próstata; PAAD: adenocarcinoma de pâncreas. Asterisco (*) indica que não foram encontradas informações para determinado estágio nos dados do TCGA.

Sample Stage	BRCA	COAD	PRAD	PAAD
Normal	143	139	118	130
I	168	172	*	150
II	169	174	*	154
III	173	171	*	95
IV	172	178	*	119
Tumor	167	176	165	155

Fonte: Elaborada pelo autor

O passo seguinte foi analisar a expressão gênica dos piRNAs. Ao avaliar as informações contidas na tabela 11 podemos notar que a maioria dos tecidos analisados possuem maior quantidade de piRNAs em expressão diferencial na comparação entre a condição: Normal *versus* Estádio II com 55 piRNAs sobre-regulados e apenas 2 sub-regulados, sendo que ao avaliar as outras comparações realizadas contra o tecido normal percebemos uma grande quantidade acumulada de piRNAs diferencialmente expressos em comparação as análises que foram feitas utilizando apenas estádio tumoral versus estádio tumoral, I vs II por exemplo. Esta característica se repete para todos os tecidos tumorais analisados, com exceção de PRAD além do tecido tumoral de PAAD, onde a análise não resultou em nenhum piRNA encontrado em expressão diferencial. Outra peculiaridade é a maior concentração de piRNAs encontrados sobre-regulados nos tecidos analisados, onde 13 das 15 comparações que retornaram pelo menos um piRNAs em expressão

diferencial nesta característica. Além das 17 análises de comparação da expressão não terem resultado em nenhum piRNA estando diferencialmente expresso.

Ao analisar mais a fundo a Tabela 11, o tecido oriundo da mama (BRCA) apresenta a maior quantidade de piRNAs em expressão diferencial, no qual a maioria se concentra nas comparações que utilizam o tecido normal, por exemplo a comparação Normal vs Estágio II com 57 piRNAs. Porém apesar de ter o total de piRNAs menor, o tecido COAD possui a segunda maior quantidade de piRNAs up- ou down-regulados. Na comparação Normal vs Tumor encontramos o total de 54 piRNAs diferencialmente expressos. Contudo, inversamente ao que ocorre nas outras análises realizadas, os resultados encontrados no tecido de pâncreas possui maior total de piRNAs, aliás a sua totalidade, nas comparações entre os estágios tumorais fornecidos e nenhum nas comparações que utilizam o tecido normal.

Tabela 11 - Distribuição da quantidade de piRNAs encontrados em expressão diferencial avaliando as comparações inicialmente planejadas na metodologia. Up é a abreviatura para *sobre*-regulado que indica maior quantidade de piRNAs encontrados na primeira condição em relação a segunda. Down é a abreviatura de *sub*-regulado onde indica maior quantidade de piRNAs encontrados na segunda condição versus a primeira. BRCA: adenocarcinoma de mama; COAD: adenocarcinoma de cólon; PRAD: adenocarcinoma de próstata; PAAD: adenocarcinoma de pâncreas. Asterisco (*) indica que não foram encontradas informações para determinado estágio nos dados do TCGA. Para ser considerado diferencialmente expresso o piRNA precisa apresentar no resultado da análise logFC maior ou igual a 1 ou menor ou igual -1 e p-valor ajustado menor que 0.05.

Stages Comparison	BRCA		COAD		PRAD		PAAD	
	Up	Down	Up	Down	Up	Down	Up	Down
Nor vs I	31	8	37	11	*	*	0	0
Nor vs II	55	2	39	10	*	*	0	0
Nor vs III	47	4	0	0	*	*	0	0
Nor vs IV	43	5	38	12	*	*	0	0
Nor vs Tumor	49	3	32	22	33	2	0	0
I vs II	1	2	12	0	*	*	0	5
II vs III	0	0	0	0	*	*	1	0
III vs IV	4	0	0	0	*	*	0	0

Fonte: Elaborada pelo autor

Ao avaliar os targets destes piRNAs que estão em maior quantidade em sua determinada categoria, podemos citar o piRNA hsa-piR-32933 que ocorre em maior quantidade nas amostras normais de BRCA ao compará-las com as amostras de estágio II. Este piRNA ao estar ativo no complexo RISC tem a capacidade de ligar-se ao mRNA do gene SLC39A12 e mRNA do gene MAST2. Destacando outro piRNA, porém que se encontra down-regulado, ou seja, em maior quantidade nas amostras de estágio II ao ser comparado com o tecido normal de COAD. O piRNA hsa-piR-2980 possui apenas um mRNA que consegue preencher os pré-requisitos de ligação ao complexo RISC, o mRNA do gene ASGR1.

Como citado na metodologia, o banco de dados TCGA não fornece diretamente as classificações patológicas em estágios para o tumor de próstata, portanto, o mesmo não é avaliado neste momento. Porém foi possível avaliar a expressão entre amostras normais e tumorais, que revelaram uma grande quantidade de piRNAs mais expressos nas amostras normais (33 piRNAs) em relação às tumorais, onde foi encontrado apenas 2 piRNAs.

Por fim, avaliamos o tecido do pâncreas segundo as condições mencionadas. Curiosamente, foi o único dos 4 tecidos que apresentou piRNAs em expressão diferencial para a condição estágio II *versus* estágio III, apesar de ser apenas 1 piRNA. Distanciando ainda mais este tecido dos outros analisados, nenhum piRNA foi encontrado em expressão diferencial nas análises entre amostras de estágio I, II, III, IV e tumor quando comparados a amostras de tecido normal, mesmo efeito ocorre quando analisado as amostras de tecido de estágio III *versus* estágio IV. Apenas a

condição estágio I *versus* estágio II apresentou maior quantidade de piRNAs em expressão diferencial, onde a maioria está em maior expressão nas amostras de estágio II, com 5 piRNAs. Como exemplo podemos citar o piRNA hsa-piR-26397, que além das suas características normais, teria um sítio alvo no mRNA do gene RALBP1.

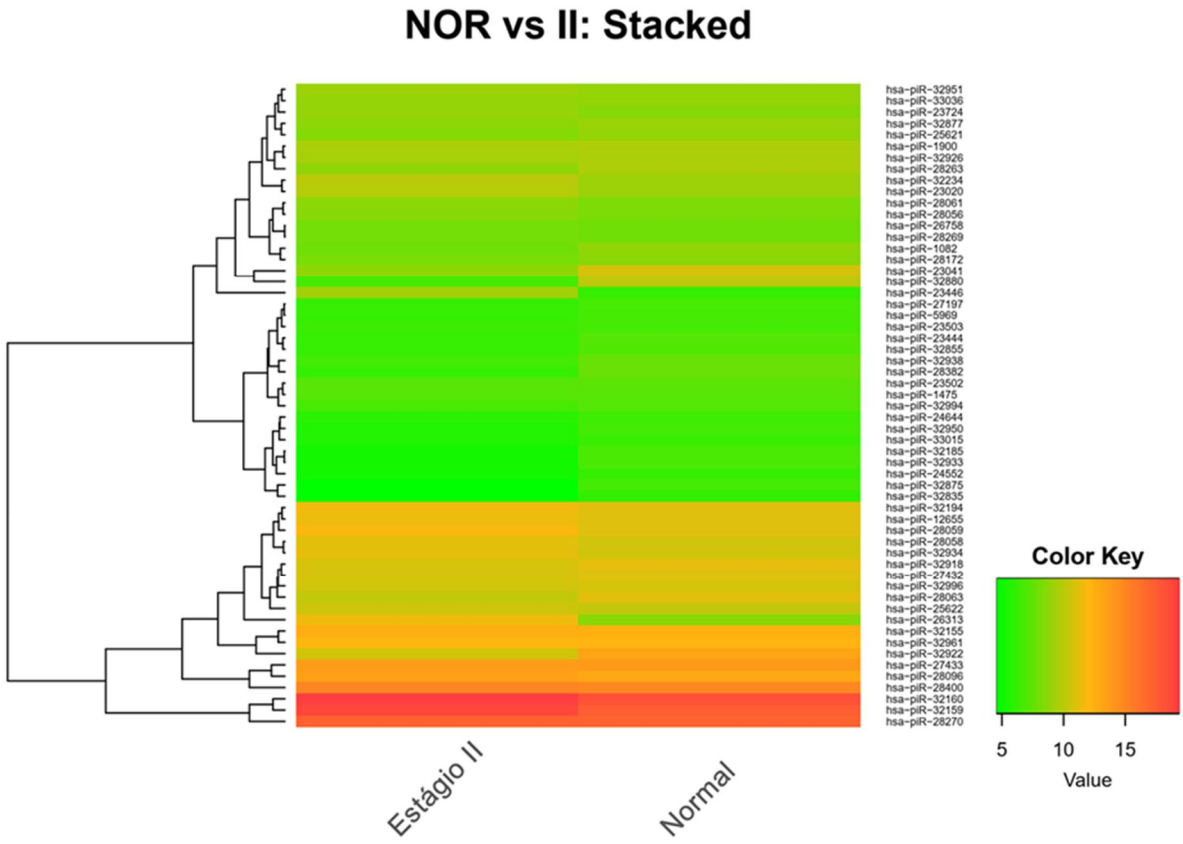

Figura 4 - Heatmap construído utilizando a comparação entre amostras normais versus estágio II do tecido BRCA. Esquema montado utilizando apenas os piRNAs que resultaram sendo diferencialmente expressos na comparação entre amostras normais versus estágio II. Quanto mais vermelha a barra correspondente ao piRNA em cada situação maior a quantidade de reads do mesmo, enquanto que a cor verde mais clara indica menor quantidade identificada de *reads* do piRNA. Fonte: Elaborada pelo autor

Dando seguimento às análises pretendidas para os dados de expressão de piRNAs, podemos notar no elemento gráfico chamado *heatmap*, que muitos piRNAs aparecem em expressão diferencial porém com pouca distinção em relação a sua contra-parte. Contudo, em diversos pontos podemos notar que piRNAs que possuem quantidades de reads sequenciados parecidos dentro das amostras comparadas, no

caso o tecido BRCA comparando normal versus estágio II. Esta proximidade em relação à quantidade ocorre principalmente em três regiões bem definidas do gráfico. Porém, existem piRNAs que estão em quantidades muito superiores ou inferiores em relação, como por exemplo o piRNA hsa−piR−26313 que ao visualizarmos no gráfico com a ajuda da legenda, possui menos reads nas amostras normais do que a quantidade encontrada nas amostras de estágio II. Situação inversa à encontrada com o piRNA hsa−piR−23041 por exemplo.

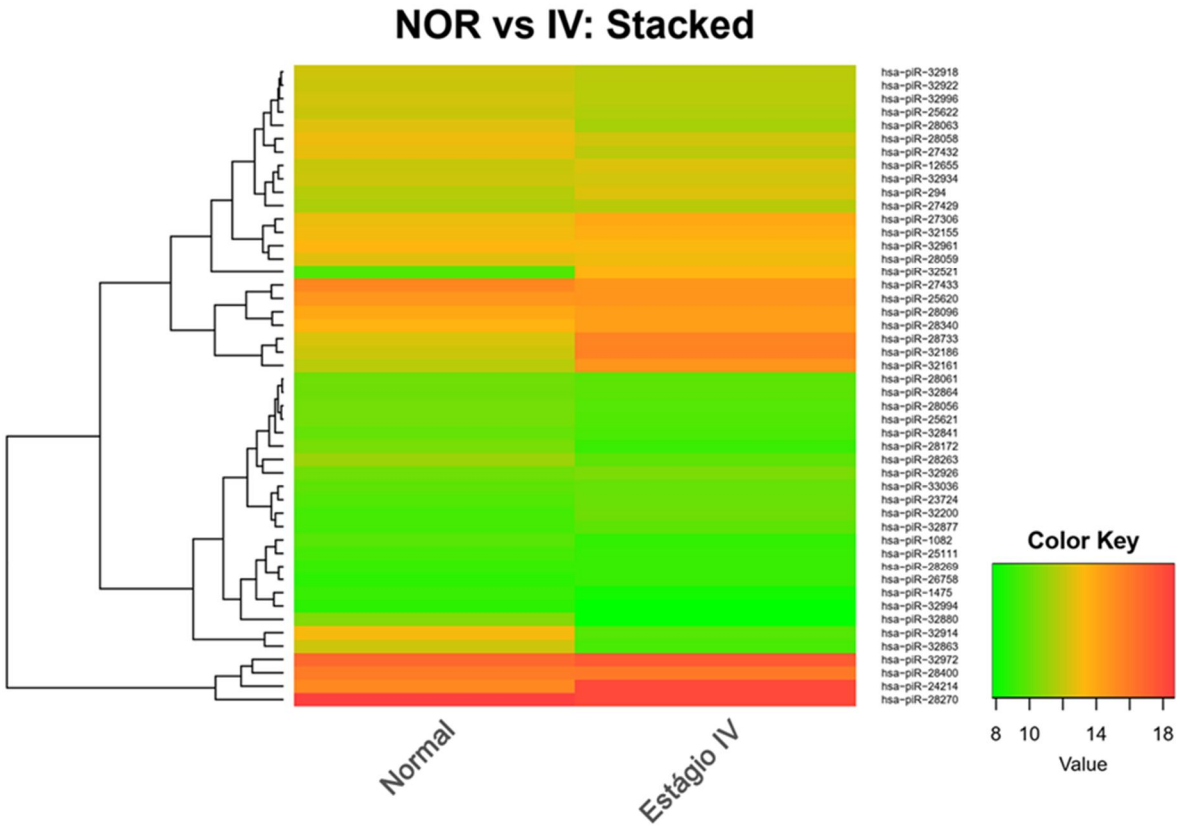

Figura 5 - Heatmap construído utilizando a comparação entre amostras normais versus estágio IV do tecido BRCA. Esquema montado utilizando apenas os piRNAs que resultaram sendo diferencialmente expressos na comparação entre amostras normais versus estágio IV. Quanto mais vermelha a barra correspondente ao piRNA em cada situação maior a quantidade de reads do mesmo, enquanto que a cor verde mais clara indica menor quantidade identificada de reads do piRNA. Fonte: Elaborada pelo autor

Ao avaliar este heatmap do mesmo tecido, BRCA, podemos notar que existe menos uniformidade em relação à separação dos piRNAs em grupos com expressão similar. Porém, ainda podemos identificar 3 grandes grupos de piRNAs que são similares em relação a sua expressão dentro deste tecido. Em relação a cada unidade de piRNAs, podemos notar uma maior disparidade de expressão, onde por exemplo o piRNA hsa−piR−32521 pode ser encontrado muito mais expresso nas amostras de tecido do estágio IV em comparação ao tecido normal de BRCA.

A fim de agregar informação às análises realizadas em relação à expressão de piRNAs e suas funções, realizamos a construção de gráficos do tipo *boxplot* da abundância de sequências de mRNA dos genes alvos dos piRNAs mais importantes para cada amostra.

A indicação de existir uma maior quantidade de piRNAs hsa-piR-32933 nas amostras normais nos mostra neste tipo de gráfico que o mRNA deste gene alvo está sendo encontrado em quantidade menor em relação ao estágio comparado. Aliando as informações relativas aos processos biogênicos e funcionais dos piRNAs encontramos um possível mRNA alvo do piRNA que pode estar sendo degradado. Ao contrário do visto na quantificação do mesmo mRNA alvo nas amostras de estágio II onde percebe-se a maior quantidade desta sequência. Diante destas informações podemos construir uma ideia do que está ocorrendo no interior da célula, junto com informações relativas a função do gene e ao final traçar uma possível relação do progresso ou aparecimento de uma doença tumoral neste indivíduo.

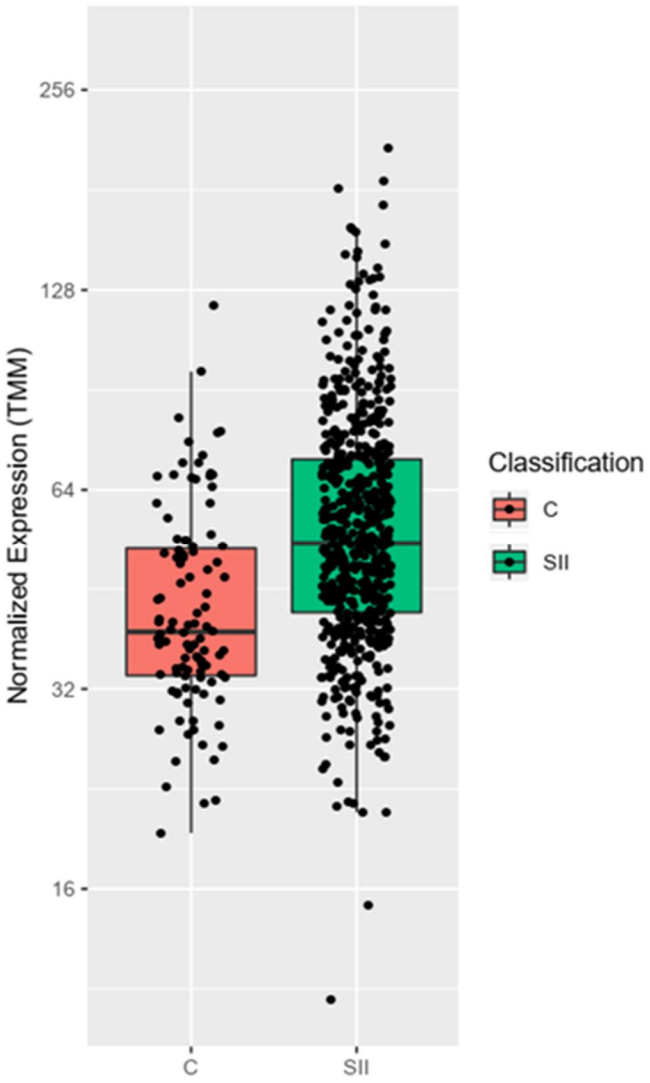

Figura 6 - Quantificação das sequências de mRNA do gene alvo MAST2 para o piRNA hsa-piR-32933 em BRCA plotado em formato de gráfico boxplot. Fonte: Elaborada pelo autor

5 Discussão

O crescente interesse no assunto e quantidade de informações disponíveis sobre os piRNAs indica uma oportunidade de se desenvolver novas ferramentas que forneçam informações relevantes, precisas e completas, assim como a necessidade de se analisar de uma forma mais abrangente esses RNAs curtos. Nesta tese realizamos uma análise sistemática dos piRNAs e construímos um banco de dados para armazenar e disponibilizar parte dessas informações (piRNAdb: https://www.pirnadb.org/).

O piRNAdb, uma plataforma on-line poderosa, integrativa e de uso facilitado, visa auxiliar o desenvolvimento de projetos associados à piRNAs. O uso de linguagens de programação modernas e arquitetura de código torna o piRNAdb capaz de fornecer rapidamente uma enorme quantidade de informações úteis ao usuário, garantindo também a estabilidade e escalabilidade para receber novos dados piRNA que serão processados. Também nos preocupamos em fornecer acesso à portabilidade por meio de dispositivos móveis ou de *desktop* para garantir uma interface amigável e eficiente. Atualmente, o piRNAdb possui maior diversidade de espécies do que outras bases de dados, como o piRNABank e o piRNAQuest, 4 e 3 espécies, respectivamente. Nós temos menos diversidade e quantidade de sequências em comparação com piRBase, mas este tem critérios duvidosos para seleção de piRNA. Esta base de dados contabilizou sequências curtas de RNA a partir de pequenas experiências de *smallRNA-seq* como piRNAs e depois adicionadas à base de dados. Este tipo de metodologia de sequenciamento poderia ser usado, porque há piRNAs incluídos, no entanto, a quantidade de outras sequências curtas de RNA é muito maior. Mesmo o tratamento das amostras com periodato não garante a remoção de todas as outras

sequências curtas de RNA (OHARA et al., 2007; BALARATNAM et al., 2018; WILLIAMS et al., 2015). Portanto, desenvolvendo piRNAdb focamos na seleção de sequências já classificadas como piRNAs ou aquelas que derivam de *smallRNA-seq* que utilizam filtros restritivos ou processos de predição de piRNA. Essas etapas restritivas de seleção tem o objetivo de manter a confiabilidade dos dados piRNAdb disponíveis para nossos usuários. Também coletamos a maior quantidade de conjuntos de dados publicados para o *H. sapiens* em comparação com o banco de dados mencionado acima. Este aspecto é relevante devido ao aumento da variabilidade de projetos e tecidos onde as sequências do piRNA foram encontradas.

Nós tivemos a preocupação de além de concatenar os piRNAs de *datasets* diferentes para a mesma espécie e remover os duplicados, manter as referências para os autores. Pois, acreditamos que caso um piRNA tenha sido encontrado em análises de sequenciamento por diferentes grupos de pesquisa se torna uma característica excelente, fornecendo maior confiabilidade de que seja uma sequência de piRNA e não apenas um artefato. Também fornecendo aos pesquisadores seu merecido reconhecimento por ter sequenciado e disponibilizado para o público seus dados das suas pesquisas.

Um passo importante para realizar todas as outras análises *downstream* do piRNAdb foi o alinhamento dos piRNAs ao genoma de referência dos organismos. Assim, escolher um software específico para alinhar sequências curtas, como BWA, e restringir parâmetros foi crucial para evitar a propagação de erros. Não permitir *mismatches* nem *gaps* nos forneçam uma visão mais confiável sobre a origem genômica do piRNA. Como notamos a ocorrência de piRNAs não alinhados ao genoma, decidimos retirá-los para evitar confundir o usuário. O piRNABank também

fez esse tipo de remoção, mas não o piRBase. Tomando nota que das 236183 sequências de piRNA coletadas nas fontes citadas, cerca de 36000 não alinham ao genoma, então fica claro que essa informação precisa ser melhor avaliada, evitando confusão ao usuário.

Foi surpreendente a quantidade de alinhamentos do piRNA ao genoma apesar dos parâmetros restritivos utilizados, principalmente quando levada em consideração a quantidade de sequências de piRNA armazenadas inicialmente. A maioria dos piRNAs se alinha apenas uma vez ao genoma, enquanto algumas sequências mostram muitos alinhamentos, esses piRNAs são os principais contribuintes para a quantidade de alinhamentos em quase todos os organismos analisados, exceto *D. melanogaster*. Em *R. norvegicus* e *M. musculus* exploramos ainda mais essa informação e a explicação vem após a avaliação da sobreposição dos alinhamentos nas anotações genômicas. Várias partes são sobrepostas à anotação de elementos transponíveis, especialmente elementos repetitivos, o que é plausível se levada em consideração a característica de “armadilha” dos clusters de piRNA mencionados. Ainda encontramos muitas sobreposições nas regiões do *lncRNA*, onde há especulações de que elas estão associadas à biogênese do piRNA, atuando como RNAs precursores (HA et al., 2014). Em seguida, analisar informações relacionadas a sobreposições genômicas de piRNAs sobre elementos genômicos anotados fornece informações relevantes relacionadas à provável origem da sequência. Ainda além, ao avaliar piRNAs que possuem alinhamentos que ocorrem em posições de genes e elementos transponíveis pode fornecer uma pista mais forte para filtrar de todos os mais de 27000 piRNAs em humanos, aqueles poucos 2000 que podem estar associados a movimentação de elementos e desenvolvimento de pseudogene, que

pode, como citado, ser incorporado a um *cluster* de piRNA e iniciar o trabalho de regulação deste.

Diversos softwares foram desenvolvidos para agrupar piRNAs por densidade em regiões genômicas. Os mais conhecidos proTRAC (ROSENKRANZ e ZISCHLER, 2012) e PILFER (RAY e PANDEY, 2018) usam basicamente uma metodologia de janela deslizante com algumas modificações. No entanto, os dados e seleção de parâmetros fornecidos para esses *softwares* produzem resultados divergentes. Então, decidimos agrupar piRNAs pelo método mais simples de janela deslizante e restringir aos piRNAs com apenas um alinhamento ao genoma. Fornecemos aos usuários esta informação, que auxilia na identificação de origem genômica e futuras avaliações relacionadas à conservação do piRNA. Além disso, mostramos a existência de *clusters* uni- e bi-direcional que é uma característica bem estabelecida dos piRNAs, mas indisponível na maioria dos bancos de dados.

A fim de fornecer informações atualizadas e aumentar o contexto biológico associado, buscamos predizer os transcritos de RNA que poderiam ser alvos pelos piRNAs. O piRNAQuest tem este tipo de informação, no entanto, considera apenas genes alvos na fita oposta de um alinhamento de piRNA. Na metodologia piRNAdb, seguimos as regras de complementaridade já conhecidas, e também novas introduzidas pelo sistema *web* piRScan (WU et al., 2018). Como em outras etapas, foram aplicados filtros restritivos que reduzem a cobertura de genes alvos, no entanto, fornecem resultados mais concisos e confiáveis. Por exemplo, existe um limite de 6 *mismatches* na região fora da *seed* no piRscan, mas este foi desenvolvido para ser usado para o organismo *C. elegans* de 21 nucleotídeos. No caso de sequências piRNA de outras espécies, esse número de bases pode variar, como em *H. sapiens*, onde

temos piRNAs de 16 a 36 nucleotídeos, essa característica poderia levar a um viés devido à maior probabilidade de sequências maiores terem *mismatches* em comparação para sequências mais curtas. Então, decidimos usar um filtro para selecionar apenas os pares *target*-piRNA que mostraram pelo menos 75% de complementaridade para atenuar o efeito do tamanho da sequência. Já descrito e reproduzido por nossas análises de dados, foi o maior número de alvos sobre RNAs longos não-codificantes, de fato alguns autores sugerem que este biótipo de transcrito poderia ser usado como plataforma para a biogênese dos piRNAs (HA et al., 2014). De acordo com nosso conhecimento, o piRNAdb é o único banco de dados de piRNA que exibe informações sobre os pares exclusivos de piRNA-alvo e termos de geneOntology associados aos genes. Por fim, o fornecimento de alvos preditos, principalmente aquelas exclusivas, e informações sobre o contexto biológico apoiam o pesquisador e cumprem nosso objetivo de fornecer uma plataforma útil e integradora para apoiar a pesquisa do piRNA.

Para facilitar e reduzir o tempo de processamento necessário para avaliar um conjunto de piRNA, fornecemos a abundância de piRNAs em uma coleção selecionada de dados de sequenciamentos de RNA curtos. Existem outros bancos de dados que exibem esse tipo de informação, mas problemas na visualização e comparação de amostras ou projetos dificultam sua usabilidade. No piRNAdb nós fornecemos a contagem de sequências normalizadas e então podemos comparar com outros tecidos utilizando o *boxplot* fornecido.

Além disso, para facilitar a migração para nossa plataforma, fornecemos uma ferramenta chamada "*CrossCodes*". Ele traduz os códigos de acesso de outros bancos de dados de piRNA e artigos publicados para aqueles usados no piRNAdb, onde o usuário pode encontrar mais informações, como possíveis alvos, expressão

de piRNA e características genômicas associadas. Isso ajudará esses pesquisadores a comparar 2 piRNAs em diferentes fontes ou publicações, permitindo saber se os diferentes códigos de acesso pertencem ao mesmo piRNA.

A primeira versão do piRNAdb (*RELEASE* 1), fornece informações sobre *H. sapiens, M. musculus, R. norvegicus, C. griseus, D. melanogaster* e *C. elegans*. Nós antecipamos que no próximo lançamento, piRNAdb incluirá piRNAs e suas informações relacionadas para outros primatas, bovinos e mosquitos. Além disso, mais recursos e sequências de piRNAs para os organismos já presentes.

Nosso estudo de avaliação de expressão dos piRNAs conhecidos utilizou 2026 amostras vindas de 4 tecidos diferentes, sequenciados por *miRNA-seq* e fornecidos pelo TCGA. Apesar da falta de provisionamento de classificações em estágio para o tecido da próstata, ganhamos muitas informações que poderemos refinar e oferecer aos usuários para um melhor entendimento do surgimento e desenvolvimento tumoral.

Encontramos, no total, quantidade de piRNAs anotados próximos aos encontrados por outros pesquisadores (MARTINEZ et al., 2015). Apesar da quantidade grande de piRNA, 27700 em humanos, pouco mais de 170, em média, foram encontrados. Aqui temos duas possibilidades, a primeira considerando o fato de que muitos piRNAs estejam nas amostras analisadas, porém ainda não catalogados, ou segundo essa seria a expressão normal de um piRNA em um tecido somático, tendo então atividade bem menor do que a que ocorre em tecidos germinativos.

De fato, ao avaliar os tecidos, encontramos a maior quantidade de piRNAs em expressão diferencial nos casos onde se compara tumores ou estágios tumorais

contra os tecidos normais. Isso indica um bom prognóstico para a pesquisa com piRNAs, visto que aumentam as chances de encontrar piRNAs que podem ser utilizados como marcadores de primeiros estágios da doença, alvo terapêutico ou até mesmo, neste momento, facilitando o entendimento da biologia tumoral. Porém, poucos foram encontrados ao avaliar os piRNAs entre estágios tumorais, exceto para alguns casos. Isso pode indicar que poucos ou nenhum piRNA estejam muito expressos na condição citada e, por consequência, alterando o ambiente tumoral para facilitação de invasão ou crescimento. Porém novos filtros, amostras e reavaliação do agrupamento de amostras pode ajudar a solucionar este impasse e trazer resultados mais conclusivos.

Os genes identificados como sendo *targets* dos nossos piRNAs diferencialmente expressos nos indicam que realmente ocorre a regulação de células somáticas e principalmente tumorigênese relacionados aos piRNAs. Por exemplo, NEFL, um alvo do piRNA hsa-piR-28400, gene este que já foi estudado em associação com microRNAs e proliferação celular e invasão em glioblastoma e tumores de mama (WANG et al., 2016). Ainda o gene alvo ASGR1 do hsa-piR-2980, que foi associado a supressão de metástases em carcinomas hepatocelulares (GU et al., 2016). Por fim, o gene SEZ6L, alvo do hsa-piR-32186, onde a baixa expressão do mesmo em linhagens celulares levou a maior incidência de tumores pulmonares (GORLOV et al., 2007). Ou seja, observar os piRNAs diferencialmente expressos podem dar informações relativas à progressão dos estágios tumorais, mas olhar também para aqueles que mantém suas expressões constantes podem dar pistas sobre a regulação de diversos genes envolvidos em processos metabólicos que são *hallmarks* de câncer.

Porém em alguns casos, as análises podem ter sido prejudicadas pela pouca quantidade de amostras dispostas para avaliação, por exemplo, possuímos apenas 3 amostras normais, de estágio III e estágio IV do tecido pancreático. Como a avaliação de expressão leva em consideração a quantidade de amostras para indicar a confiabilidade do resultado, esta pouca quantidade de amostras leva a valores de p-valor muito alto, sendo removido pelos filtros utilizados pelo nosso processamento. Uma solução seria a utilização de dados de outros sequenciamentos não contidos no TCGA para esse tecido, ou caso persista o problema, abandonar a avaliação de expressão com as categorias incluídas que possuem poucas amostras, e continuar com aqueles que retornaram resultados, como estágio I *versus* estágio II. Outro problema registrado foi no caso de amostras de próstata, onde o TCGA não informa a classificação de estágios para as amostras para as amostras, onde não pudemos fazer essa avaliação de forma certa, ficando apenas com a avaliação de tecido normal contra tumoral. Uma forma de contornar o problema já está sendo realizado onde serão utilizados marcadores moleculares já disponíveis para separar as amostras em novas categorias e poder realizar a análise de expressão diferencial.

6 Conclusão

piRNAdb fornece informações confiáveis sobre piRNAs e várias análises downstream para facilitar estudos de piRNA em seis espécies. Nossas descobertas sobre a quantidade de alinhamentos e alvos gênicos associados fornecem informações úteis sobre os piRNAs. Comprometemo-nos a fornecer vários tipos de informação, mas principalmente isso deve ser útil para pesquisadores e entusiastas no campo piRNA. Além disso, fornecemos nossos resultados em elementos gráficos para facilitar a visualização e interpretação de nossos usuários. Acreditamos que a comunidade científica pode se beneficiar com a informação relacionada ao contexto biológico fornecido pelo piRNAdb, como novos estudos validando as funções previstas. Por fim, quando utilizamos tecnologias modernas de programação e arquitetura, permitindo um ambiente multiplataforma aumentar a linha de base para o campo de desenvolvimento de bancos de dados, onde o maior beneficiário é o usuário. Ao final, ao avaliar todas as análises encontradas, desde informações e possibilidade de evolução do piRNAdb acreditamos que esta base de dados auxiliará muito no progresso da pesquisa do piRNA.

Ao avaliar a expressão de piRNAs entre amostras normais e de tumor, identificamos que muitos piRNAs estão alterados. Estes, por sua vez, podem ser ótimos candidatos iniciais para desenvolvimento de análises de biópsias ou kits de diagnósticos. Principalmente em relação ao fato de que muitos desses possuem como *targets*, genes relacionados a início e progressão tumoral já documentados.

Apesar de ter identificado muitas informações promissoras em relação às características genômicas e funcionais dos piRNAs, acreditamos que estudos em maior escala e mais aprofundados precisam ser realizados para filtrar os dados e

resultados, além de entender melhor as informações obtidas nestas primeiras análises de tecido tumoral. Acreditamos ainda, que muito precisa ser explorado, principalmente porque muito dos projetos públicos de sequenciamento de RNA curtos ainda pode ser minerado para buscar piRNAs.

Referências

AMIN MB, GREENE FL, EDGE SB, COMPTON CC, GERSHENWALD JE, BROOKLAND RK, MEYER L, GRESS DM, BYRD DR, WINCHESTER DP. **The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging**. CA Cancer J Clin. 2017 Mar;67(2):93-99.

ARAVIN A, GAIDATZIS D, PFEFFER S, LAGOS-QUINTANA M, LANDGRAF P, IOVINO N, MORRIS P, BROWNSTEIN MJ, KURAMOCHI-MIYAGAWA S, NAKANO T, CHIEN M, RUSSO JJ, JU J, SHERIDAN R, SANDER C, ZAVOLAN M, TUSCHL T. **A novel class of small RNAs bind to MILI protein in mouse testes.** Nature. 2006; 442:203–207.

ARAVIN A, HANNON G, BRENNECKE J. **The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race**. Science. 2007; 318(5851):761-764.

ASSUMPÇÃO CB, CALCAGNO DQ, ARAÚJO TM, SANTOS SE, SANTOS ÂK, RIGGINS GJ, BURBANO RR, ASSUMPÇÃO PP. **The role of piRNA and its potential clinical implications in cancer**. Epigenomics. 2015; 7(6):975-84.

BALARATNAM S, WEST N, BASU S. **A piRNA utilizes HILI and HIWI2 mediated pathway to down-regulate ferritin heavy chain 1 mRNA in human somatic cells.** Nucleic acids research. 2018, vol. 46, nº 20, pp. 10635-10648

BARTEL DP. **MicroRNAs: genomics, biogenesis, mechanism, and function**. Cell. 2004 Jan 23;116(2):281-97.

BORTVIN A. **PIWI-Interacting RNAs (piRNAs) - a Mouse Testis Perspective. Biochemistry (Mosc).** 2013; 78(6):592-602

CASPER J, ZWEIG AS, VILLARREAL C, TYNER C, SPEIR ML, ROSENBLOOM KR, RANEY BJ, LEE CM, LEE BT, KAROLCHIK D, HINRICHS AS, HAEUSSLER M, GURUVADOO L, NAVARRO J GONZALEZ, GIBSON D, FIDDES IT, EISENHART C, DIEKHANS M, CLAWSON H, BARBER GP, ARMSTRONG J, HAUSSLER D, KUHN M, KENT WJ. **The UCSC Genome Browser database: 2018 update**. Nucleic Acids Research. 2018, vol. 46, nº D1, pp. D762-D769

CHENG K, MAHATO RI. **Biological and therapeutic applications of small RNAs**. Pharm Res. 2011 Dec;28(12):2961-5.

CHENG J, DENG H, XIAO B, ZHOU H, ZHOU F, SHEN Z, GUO J. **piR-823, a novel non-coding small RNA, demonstrates in vitro and in vivo tumor suppressive activity in human gastric cancer cells**. Cancer Lett. 2012; 315:12-17

CLARK JP, LAU NC. **Piwi proteins and piRNAs step onto the systems biology stage.** Adv Exp Med Biol. 2014; 825: 159–197.

EDGE SB, COMPTON CC. **The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM**. Ann Surg Oncol. 2010 Jun;17(6):1471-4.

ESTELLER M. **Non-coding RNAs in human disease**. Nat Rev Genet. 2011; 12:861–874.

GANGARAJU VK, LIN H. **MicroRNAs: Key regulators of stem cells**. Nat Rev Mol Cell Biol. 2009; 10:116–125.

GEBERT D, KETTING RF, ZISCHLER H, ROSENKRANZ D. **piRNAs from Pig Testis Provide Evidence for a Conserved Role of the Piwi Pathway in Post-Transcriptional Gene Regulation in Mammals**. PLoS One. 2015; 10(5):e0124860.

GHILDIYAL M, ZAMORE PD. **Small silencing RNAs: An expanding universe.** Nat Rev Genet. 2009; 10:94–108.

GORLOV IP, MEYER P, LILOGLOU T, MYLES J, BOETTGER MB, CASSIDY A, GIRARD L, MINNA JD, FISCHER R, DUFFY S, SPITZ MR, HAEUSSINGER K, KAMMERER S, CANTOR C, DIERKESMANN R, FIELD JK, AMOS CI. **Seizure 6-like (SEZ6L) gene and risk for lung cancer**. Cancer Res. 2007 Sep 1;67(17):8406-11.

GU D, JIN H, JIN G, WANG C, WANG N, HU F, LUO Q, CHU W, YAO M, QIN W. **The asialoglycoprotein receptor suppresses the metastasis of hepatocellular carcinoma via LASS2-mediated inhibition of V-ATPase activity**. Cancer Lett. 2016 Aug 28;379(1):107-16.

HA H, SONG J, WANG S, KAPUSTA A, FESCHOTTE C, CHEN KC, XING J. **A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile element**s. BMC Genomics. 2014; 15:545.

HAMILTON AJ, BAULCOMBE DC. **A species of small antisense RNA in posttranscriptional gene silencing in plants**. Science. 1999;286:950–952.

HAMMOND SM, BERNSTEIN E, BEACH D, HANNON GJ. **An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells**. Nature. 2000;404:293–296.

HAYES J, PERUZZI PP, LAWLER S. MicroRNAs in cancer: Biomarkers, functions and therapy. Trends Mol Med.; 2014;20(8):460–9.

HIRANO T, IWASAKI Y, LIN Z, IMAMURA M, SEKI N, SASAKI E, SAITO K, OKANO H, SIOMI M, SIOMI H. **Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate**. RNA. 2014; 20(8):1223-37

ISHIZU H, SIOMI H, SIOMI1 MC**. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines**. Genes Dev. 2012; 26(21):2361-73

KAWAOKA S, IZUMI N, KATSUMA S, TOMARI Y. **3' End formation of PIWI-interacting RNAs in vitro.** Mol Cell. 2011; 43:1015-1022

KOZOMARA A AND GRIFFITHS-JONES S. MiRBase: **Annotating high confidence microRNAs using deep sequencing data**. Nucleic Acids Research. 2014, vol. 42, nº D1

KUFEL J, GRZECHNIK P. **Small Nucleolar RNAs Tell a Different Tale**. Trends Genet. 2019 Feb;35(2):104-117.

LAKSHMI SS, AGRAWAL S. **piRNABank: a web resource on classified and clustered Piwi-interacting RNAs**. Nucl. Acids Res. 2008; 36(suppl 1):D173-D177.

LE THOMAS A, STUWE E, LI S, DU J, MARINOV G, ROZHKOV N, CHEN YA, LUO Y, SACHIDANANDAM R, TOTH KF, PATEL D, ARAVIN AA. **Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing.** Genes & Development. 2014; 28:1667–1680.

LI H. AND DURBIN R. **Fast and accurate short read alignment with Burrows-Wheeler Transform.** Bioinformatics. 2010; 25:1754-60.

LUDVÍKOVÁ M, KALFEŘT D, KHOLOVÁ I. **Pathobiology of MicroRNAs and Their Emerging Role in Thyroid Fine-Needle Aspiration**. Acta Cytol. 2015;59(6):435-44.

MALONE CD, BRENNECKE J, DUS M, STARK A, MCCOMBIE WR. **Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary**. Cell. 2009; 137: 522–535.

MARTINEZ VD, VUCIC EA, THU KL, HUBAUX R, ENFIELD KSS, PIKOR LA, BECKER-SANTOS DD, BROWN CJ, LAM S & LAM WL. **Unique somatic and malignant expression patterns implicate PIWI-interacting RNAs in cancer-type specific biology**. Scientific Reports. 2015; 5:10423

MEI Y, CLARK D, MAO L. **Novel dimensions of piRNAs in cancer**. Cancer Lett. 2013; 336(1): 46–52

MOAZED D. **Small RNAs in transcriptional gene silencing and genome defence**. Nature. 2009; 457:413-420

NCBI RESOURCE COORDINATORS. **Database Resources of the National Center for Biotechnology Information**. Nucleic Acids Research. 2017, vol. 45, nº D1, pp. D12-D17, 4 1

NISHIDA KM, OKADA TN, KAWAMURA T, MITUYAMA T, KAWAMURA Y, INAGAKI S, HUANG H, CHEN D, KODAMA T, SIOMI H, SIOMI MC. **Functional involvement of Tudor and dPRMT5 in the piRNA processing pathway in Drosophila germlines**. EMBO J. 2009; 28:3820-31

O'LEARY NA, WRIGHT MW, BRISTER JR, CIUFO C, HADDAD D, MCVEIGH R, RAJPUT B, ROBBERTSE B, SMITH-WHITE B, AKO-ADJEI D, ASTASHYN A, BADRETDIN A, BAO Y, BLINKOVA O, BROVER V, CHETVERNIN V, CHOI J, COX E, ERMOLAEVA O, FARRELL CM, GOLDFARB T, GUPTA T, HAFT D, HATCHER E, HLAVINA W, JOARDAR VS, KODALI VK, LI W, MAGLOTT D, MASTERSON P, MCGARVEY KM, MURPHY MR, O'NEILL K, PUJAR S, RANGWALA SH, RAUSCH D, RIDDICK LD, SCHOCH C, SHKEDA A, STORZ SS, SUN H, THIBAUD-NISSEN F, TOLSTOY I, TULLY RE, VATSAN AR, WALLIN C, WEBB D, WU W, LANDRUM MJ, KIMCHI A, TATUSOVA T, DICUCCIO M, KITTS P, MURPHY TD, PRUITT KD. **Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation.** Nucleic Acids Research. 2016, vol. 44, nº D1, pp. D733-D745

OHARA T, SAKAGUCHI Y, SUZUKI T, UEDA H, MIYAUCHI K, SUZUKI T. **The 3′ termini of mouse Piwi-interacting RNAs are 2′-O-methylated**. Nature Structural and Molecular Biology. 2007, vol. 14, nº 4, pp. 349-350

OZATA D, GAINETDINOV J, ZOCH A, O'CARROLL D, ZAMORE P. **PIWI-interacting RNAs: small RNAs with big functions**. Nature Reviews Genetics, 2018

PANG KC, STEPHEN S, ENGSTRÖM PG, TAJUL-ARIFIN K, CHEN W, WAHLESTEDT C, LENHARD B, HAYASHIZAKI Y, MATTICK JS. **RNAdb - A comprehensive mammalian noncoding RNA database**. Nucleic Acids Research. 2005, vol. 33, nº DATABASE ISS

PRUITT K, TATUSOVA T, MAGLOTT DR. **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. Nucleic Acids Res. 2005; 33(Database issue):D501-4.

QUINLAN AR & HALL IM. **BEDTools: a flexible suite of utilities for comparing genomic features.** Bioinformatics, 2010; 26(6):841–842

QU F, ET AL. **RDR6 has a broad-spectrum but temperature-dependent antiviral defense role in Nicotiana benthamiana**. J Virol. 2005;79:15209–15217.

RAY R AND PANDEY P. **piRNA analysis framework from small RNA-Seq data by a novel cluster prediction tool – PILFER**. Genomics. 2018, vol. 110, nº 6, pp. 355-365

ROBINSON MD, MCCARTHY DJ AND SMYTH GK. **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** Bioinformatics. 2010; 26:pp.-1.

ROOVERS EF, ROSENKRANZ D, MAHDIPOUR M, HAN CT, HE N, CHUVA DE SOUSA LOPES SM, VAN DER WESTERLAKEN LA, ZISCHLER H, BUTTER F, ROELEN BA, KETTING RF. **Piwi proteins and piRNAs in mammalian oocytes and early embryos**. Cell Rep. 2015; 10(12):2069-82.

ROSENKRANZ D AND ZISCHLER H. **proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis**. BMC Bioinformatics. 2012, vol. 13, nº 1

ROSENKRANZ D, HAN CT, ROOVERS EF, ZISCHLER H, KETTING RF. **Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence**. Genomics Data, 2015, vol. 5, pp. 309-313

ROSENKRANZ D. **piRNA cluster database: a web resource for piRNA producing loci**. Nucleic Acids Res. 2016 Jan 4; 44(Database issue): D223–D230.

SANA J, FALTEJSKOVA P, SVOBODA M, SLABY O. **Novel classes of non-coding RNAs and cancer.** Journal of Translational Medicine. 2012; 10:103.

SARKAR A, MAJI RK, SAHA S, GHOSH Z. **piRNAQuest: searching the piRNAome for silencers**. BMC Genomics. 2014; 15:555

SAXE JP, LIN H. **Small noncoding RNAs in the germline**. Cold Spring Harb Perspect Biol. 2011; 3:a002717.

SHEN EZ, CHEN H, OZTURK AR, TU S, SHIRAYAMA M, TANG W, DING YH, DAI SY, WENG Z, MELLO CC. **Identification of piRNA Binding Sites Reveals the Argonaute Regulatory Landscape of the C. elegans Germline**. Cell. 2018, vol. 172, nº 5, pp. 937-940.e18

SIOMI MC, SATO K, PEZIC D, ARAVIN AA. **PIWI-interacting small RNAs: The vanguard of genome defence.** Nat Rev Mol Cell Biol. 2011; 12:246–258.

SUZUKI R, HONDA S, KIRINO Y. **PIWI expression and function in cancer.** Front Genet. 2012; 3:204.

TAN Y, LIU L, LIAO M, ZHANG C, HU S, ZOU M, GU M, LI X. **Emerging roles for PIWI proteins in cancer.** Acta Biochim Biophys Sin. 2015; 47(5), 315–324.

THE RNACENTRAL CONSORTIUM. **RNAcentral: A comprehensive database of non-coding RNA sequences**. Nucleic Acids Research. 2017, vol. 45, nº D1, pp. D128-D134

VILELA APP, AGUIAR ERGR, FERREIRA F V, RIBEIRO LS, OLMO RP, MARQUES JT. **Small Non-Coding RNAs as Biomarkers**. Recent Pat Biomark. 2012;2(2):119–30.

WANG ZY, XIONG J, ZHANG SS, WANG JJ, GONG ZJ, DAI MH. **Up-Regulation of microRNA-183 Promotes Cell Proliferation and Invasion in Glioma By Directly Targeting NEFL**. Cell Mol Neurobiol. 2016 Nov;36(8):1303-1310.

WATANABE T, LIN H. **Posttranscriptional regulation of gene expression by Piwi proteins and piRNAs**. Mol Cell. 2014; 56:18– 27.

WATANABE T, CHENG E, ZHONG M, LIN H. **Retrotransposons and pseudogenes regulate mRNAs and lncRNAs via the piRNA pathway in the germline**. Genome Res. 2015; 25:368-380

WILLIAMS Z, MOROZOV P, MIHAILOVIC A, LIN C, PUVVULA PK, JURANEK S, ROSENWAKS Z, TUSCHL T. **Discovery and Characterization of piRNAs in the Human Fetal Ovary**. Cell Reports. 2015, vol. 13, nº 4, pp. 854-863

WU WS, HUANG WC, BROWN JS, ZHANG D, SONG X, CHEN H, TU S, WENG Z, LEE HC. **PirScan: A webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans**. Nucleic Acids Research. 2018, vol. 46, nº W1, pp. W43-W48

YAMANAKA S, SIOMI MC, SIOMI H. **piRNA clusters and open chromatin structure**. Mobile DNA. 2014; 5:22.

ZERBINO DR, ACHUTHAN P, AKANNI W, AMODE MR, BARRELL D, BHAI J, BILLIS K, CUMMINS C, GALL A, GIRÓN CG, GIL L, GORDON L, HAGGERTY L, HASKELL E, HOURLIER T, IZUOGU OG, JANACEK SH, JUETTEMANN T, TO JK, LAIRD MR, LAVIDAS I, LIU Z, LOVELAND JE, MAUREL J, MCLAREN W, MOORE B, MUDGE J, MURPHY DN, NEWMAN V, NUHN M, OGEH D, ONG CK, PARKER A, PATRICIO M, RIAT HS, SCHUILENBURG H, SHEPPARD D, SPARROW H, TAYLOR K, THORMANN H, VULLO A, WALTS B, ZADISSA A, FRANKISH A, HUNT SE, KOSTADIMA M, LANGRIDGE N, MARTIN FJ, MUFFATO M, PERRY E, RUFFIER M, STAINES DM, TREVANION SJ, AKEN BL,

CUNNINGHAM F, YATES A, FLICEK P. **Ensembl 2018**. Nucleic Acids Research. 2018, vol. 46, nº D1, pp. D754-D761

ZHANG H, REN Y, XU H, PANG D, DUAN C, LIU C. **The expression of stem cell protein Piwil2 and piR-932 in breast cancer**. Su rg Oncol. 2013; 22(4):217-23

ZHANG D, TU S, STUBNA M, WU WS, HUANG WC, WENG Z, LEE HC. **The piRNA targeting rules and the resistance to piRNA silencing in endogenous genes**. Science. 2018, vol. 359, nº 6375, pp. 587-592

ZUO L, WANG Z, TAN Y, CHEN X, LUO X. **piRNAs and Their Functions in the Brain.** International Journal of Human Genetics. 2016; 16(1-2):53-60.