UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

VITOR GREGORIO

Sistema automático para caracterização de RNAs não-codificantes

CORNÉLIO PROCÓPIO

2023

VITOR GREGORIO

Sistema automático para caracterização de RNAs não-codificantes

Automatic system for characterization of non-coding RNAs

Dissertação apresentada como requisito para obtenção do título de Mestre em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador: Dr. Alexandre Rossi Paschoal.
Coorientador: Dr. Douglas Silva Domingues.
Colaboradora: Liliane Santana Oliveira

CORNÉLIO PROCÓPIO

2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Cornélio Procópio

VITOR GREGORIO

SISTEMA AUTOMÁTICO PARA CARACTERIZAÇÃO DE RNAS NÃO-CODIFICANTES

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Bioinformática.

Data de aprovação: 09 de Outubro de 2023

Dr. Alexandre Rossi Paschoal, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Alan Mitchell Durham, Doutorado - Usp-Universidade de São Paulo

Dr. Flavia Lombardi Lopes, Doutorado - Universidade Estadual Paulista - Unesp

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 29/11/2023.

Dedico este trabalho à minha família e amigos, pelos momentos de ausência.

AGRADECIMENTOS

Certamente estes parágrafos não irão atender a todas as pessoas que fizeram parte dessa importante fase de minha vida. Portanto, desde já peço desculpas àquelas que não estão presentes entre essas palavras, mas elas podem estar certas que fazem parte do meu pensamento e de minha gratidão.

Agradeço ao meu orientador Prof. Dr. Alexandre Rossi Paschoal, pela sabedoria com que me guiou nesta trajetória, toda amizade e parceria.

A Fundação Araucária, pela bolsa técnica de recursos do NAPI de bioinformática. Junto agradeço ao meu orientador da bolsa Prof. Dr. Fabio Fernandes da Rocha Vicente, por todo auxílio e paciência para a realização das atividades.

Ao PPGBioinfo, pelo auxílio financeiro direcionados a participação e apresentação em eventos de bioinformática.

Ao DIRGE-CP e DIRPPG-CP, pelo auxílio destinado ao projeto, no intuito de aumentar a quantidade de pesquisas no campus.

Aos colaboradores Pedro, Flavia e Liliane, que me auxiliaram no desenvolvimento do meu trabalho desde o começo e me trouxeram muitos ensinamentos.

Enfim, a todos os que por algum motivo contribuíram para a realização desta pesquisa.

Eu denomino meu campo de Gestão do Conhecimento, mas você não pode gerenciar conhecimento. Ninguém pode. O que você pode fazer, o que a empresa pode fazer é gerenciar o ambiente que otimize o conhecimento. (DAVENPORT; PRUSAK, 2012).

LISTA DE ILUSTRAÇÕES

SUMÁRIO

RESUMO

Os RNAs não-codificantes (ncRNAs) são RNAs que podem ser transcritos, mas não traduzidos em proteínas. Embora suas funções não sejam totalmente conhecidas, os ncRNAs possuem diversas funções biológicas, geralmente relacionadas a processos regulatórios ou interacionais, como alterações da cromatina, regulação transcricional, organização nuclear, tradução, entre outros. Duas principais formas de identificar ncRNAs são a análise de similaridade de sequência (alinhamento), que pode ser realizada com a ferramenta BLAST, ou a busca estrutural, usando a ferramenta INFERNAL. No entanto, a análise de dados após a obtenção dos resultados de ambas as ferramentas ainda é um desafio. Nesse contexto, existem duas ferramentas principais (StructRNAfinder e FindNonCoding) que foram desenvolvidas para facilitar a anotação de ncRNAs. Porém, elas não abrangem todas as principais formas de identificação de ncRNAs. Para preencher essa lacuna, desenvolvemos um sistema automático e escalável para a análise de anotação de ncRNAs em larga escala, que utiliza tanto a estratégia de busca de sequência quanto a de busca estrutural para anotação de ncRNAs. Em nossa ferramenta, utilizamos a versão mais atualizada do INFERNAL em conjunto com os bancos de dados RFAM, e BLAST juntamente com os bancos de dados RNAcentral, para realizar a identificação de ncRNAs e fornecer os resultados em relatórios, arquivos e estatísticas de fácil compreensão para o usuário final. Para validar a ferramenta, apresentamos um teste comparativo com outras duas ferramentas que visam facilitar a anotação de ncRNAs (StructRNAfinder e FindNonCoding), e realizamos os testes em genomas públicos do RefSeq, Ensembl Plants e GENCODE. O conjunto de dados para o teste contém sete genomas nucleares disponíveis em bancos de dados públicos, que são *Chlamydia trachomatis, Drosophila melanogaster, Escherichia coli* e *Saccharomyces cerevisiae* do RefSeq; *Homo sapiens* do Gencode; *Arabidopsis thaliana, Oryza sativa* e *Zea mays* do Ensembl Plants. Nossa ferramenta apresenta maior sensibilidade e precisão em comparação com outras ferramentas, o que pode indicar que nosso método fornece melhores resultados para a anotação de ncRNAs.

Palavras-chave: anotação, ncRNAs, não-codificante, identificação, computacional.

ABSTRACT

Non-coding RNAs (ncRNA) are RNAs that can be transcribed, but not translated into proteins. Although their functions are not fully known, ncRNAs have many biological functions, generally focusing on regulatory or interactional processes, such as chromatin alterations, transcriptional regulation, nuclear organization, translation, etc. Two key ways to identify ncRNAs are by sequence similarity analysis (alignment), which can be done with the BLAST tool, or structural search, by using the INFERNAL tool. However, the post-results data analysis among both tools output is still a gap. In this context, there are two major tools (StructRNAfinder and FindNonCoding) that have been developed to facilitate the ncRNA annotation. However, they do not cover all the main strategies for ncRNA identification. To fill this gap, we developed an automatic and scalable system for large-scale data annotation analysis of ncRNAs which use both sequence and structural search strategy for ncRNA annotation. Our tool uses the most updated version of INFERNAL together with RFAM and BLAST along with RNAcentral databases to perform the ncRNA identification, and bring the output in user-friendly reports, files and statistics for the final user. To validate the tool, we present a benchmark with two other tools that aims to facilitate the annotation of ncRNAs (StructRNAfinder and FindNonCoding), and tested in public genomes from RefSeq, Ensembl Plants and GENCODE. The dataset for the test contained seven nuclear genomes available in public databases, which were *Chlamydia trachomatis, Drosophila melanogaster, Escherichia coli and Saccharomyces cerevisiae* from RefSeq; *Homo sapiens* from Gencode; *Arabidopsis thaliana, Oryza sativa* and *Zea mays* from Ensembl Plants. Our tool presents better sensitivity and accuracy when compared to other tools, which may indicate that our method presents better results for the annotation of ncRNAs.

Keywords: annotation, ncRNAs, non-coding, identification, computational;

1 INTRODUÇÃO

Os RNAs não-codificantes (ou do inglês *non-coding RNAs* ou apenas ncRNA) são RNAs que são transcritos, mas não traduzidos em proteínas (MATTICK; MAKUNIN, 2006; TANTRAY et al. 2023). Os ncRNAs desempenham funções como por exemplo, modificações da cromatina, regulação transcricional e traducional, controlando assim processos biológicos e patológicos (LEKKA; HALL, 2018).

Em geral, os ncRNAs são classificados por seu comprimento. Até 200 nucleotídeos (nt) são os RNAs pequenos, sendo o microRNA (miRNA) o mais estudado na literatura devido ao seu papel regulatório dos níveis de RNAs mensageiros (mRNAs) na célula (GRIFFITHS-JONES, 2004; PELLETIER et al. 2023). Já os longos RNAs não-codificantes (lncRNAs) são um tipo de ncRNA, que possuem pelo menos 200 nucleotídeos de comprimento e desempenham funções na organização celular, estrutura celular e expressão gênica, como por exemplo o controle de modificações da cromatina para todo o genoma (MATTICK et al. 2023).

Do ponto de vista computacional, para realizar a identificação (processo de anotação) de RNAs em dados biológicos, existem duas abordagens: (i) a busca por similaridade via sequência; e (ii) a busca estrutural (STEIN, 2001; FU et al. 2021).

Na busca por similaridade, são utilizadas informações disponíveis em bancos de dados para buscar via alinhamento de sequências, utilizando sequências que estão contidas no banco de dados, sendo comumente aplicada a ferramenta *Basic Local Alignment Search Tool* (BLAST) (ALTSCHUL et al., 1990).

A segunda forma usada para anotar ncRNAs é via busca estrutural, sendo analisadas as estruturas secundárias das sequências, através de estruturas conservadas já conhecidas (EDDY e DURBIN, 1994). Para realizar a busca estrutural, pode ser utilizada a ferramenta *INFERence of RNA Alignment* (INFERNAL) (NAWROCKI; EDDY, 2013).

Essas duas formas de busca e as ferramentas disponíveis auxiliam na anotação de dados biológicos, porém ainda existe a necessidade de analisar a saída individual de cada ferramenta, uma vez que ambas as ferramentas podem apresentar resultados duplicados, de baixa qualidade ou com sobreposições de coordenadas, tornando-se necessário uma ferramenta *user-friendly* que automatize esses processos de anotação.

De forma a facilitar a anotação de ncRNAs, através de interfaces web e com relatórios automatizados, programas como tRNAscan-SE (CHAN, 2019) ou sRNAtoolbox (APARICIO-PUERTA et al., 2022) foram desenvolvidos. Contudo, a ferramenta tRNAscan-SE foi elaborada para identificar apenas RNAs transportadores (tRNAs), e, da mesma forma, o sRNAtoobox foi projetado para encontrar exclusivamente pequenos RNAs (smallRNAs ou sRNAs).

A ferramenta StructRNAfinder (ARIAS-CARRASCO et al., 2018) foi desenvolvida de forma a facilitar a utilização do INFERNAL, uma vez que automatiza a instalação local e disponibiliza uma interface gráfica no formato web. Da mesma forma, Wright (2022) desenvolveu a ferramenta FindNonCoding (WRIGHT, 2022) com o objetivo de ser fácil de utilizar, analisando as estruturas dos ncRNAs. Porém, nenhuma das duas ferramentas utiliza os dois métodos de busca (similaridade e estrutural) juntos. Ainda, a ferramenta FindNonCoding foi desenvolvida apenas como um pacote da linguagem R, obrigando o usuário a ter um conhecimento prévio da linguagem de programação.

Mesmo existindo ferramentas que podem ser utilizadas para anotar RNAs não-codificantes, ainda tem a falta de um sistema que integre as duas formas de anotar ncRNAs (busca estrutural e por similaridade), além de ser um sistema que facilite a experiência do usuário com o auxílio de interfaces gráficas, e que apresente os resultados de forma clara e organizada, ampliando o número de usuários e, consequentemente, a velocidade de anotação.

Dado esse cenário, o objetivo do presente projeto é a criação de um sistema automático e fácil de utilizar para anotar RNAs não-codificantes, integrando as duas estratégias de identificação de ncRNAs. A ferramenta foi avaliada no servidor do Programa de Pós Graduação em Bioinformática (PPGBioinfo), com sete genomas disponíveis em bancos de dados públicos, além de ser avaliado em dois estudos de caso: o primeiro em um estudo do Canadá que trabalha com *Leishmania donovani* em vesículas extracelulares; e o segundo estudo de caso em quatro genomas de *Theobroma spp.* disponibilizado pelo Prof. Dr. Douglas Silva Domingues da Universidade de São Paulo (USP) e Prof. Dr. Alessandro de Mello Varani, da Universidade Estadual Paulista (UNESP).

O sistema desenvolvido facilitou a anotação de RNAs não-codificantes, uma vez que automatizou o processo de execução das ferramentas. Além disso, a análise dos resultados se tornou rápida, pois o sistema já realiza a remoção de resultados

duplicados ou de baixa qualidade. Outra contribuição foi com os estudos de caso, que permitiu que outras etapas fossem realizadas enquanto a anotação dos genomas estava sendo realizada automaticamente. Além disso, o sistema já está sendo utilizado em outra colaboração, onde serão anotados mais de 10 mil genomas de algas e fungos.

O sistema pode ser utilizado através de linha de comando no Linux (*stand-alone*) e é distribuído no formato container para o Docker. A utilização de container facilita a utilização uma vez que auxilia na instalação do programa, pois instala todas as dependências necessárias de uma vez. Esperamos disponibilizar este sistema para a comunidade científica, com uma interface WEB e depois adaptá-la para a plataforma Galaxy, de forma a incentivar os pesquisadores a realizarem anotações de genomas.

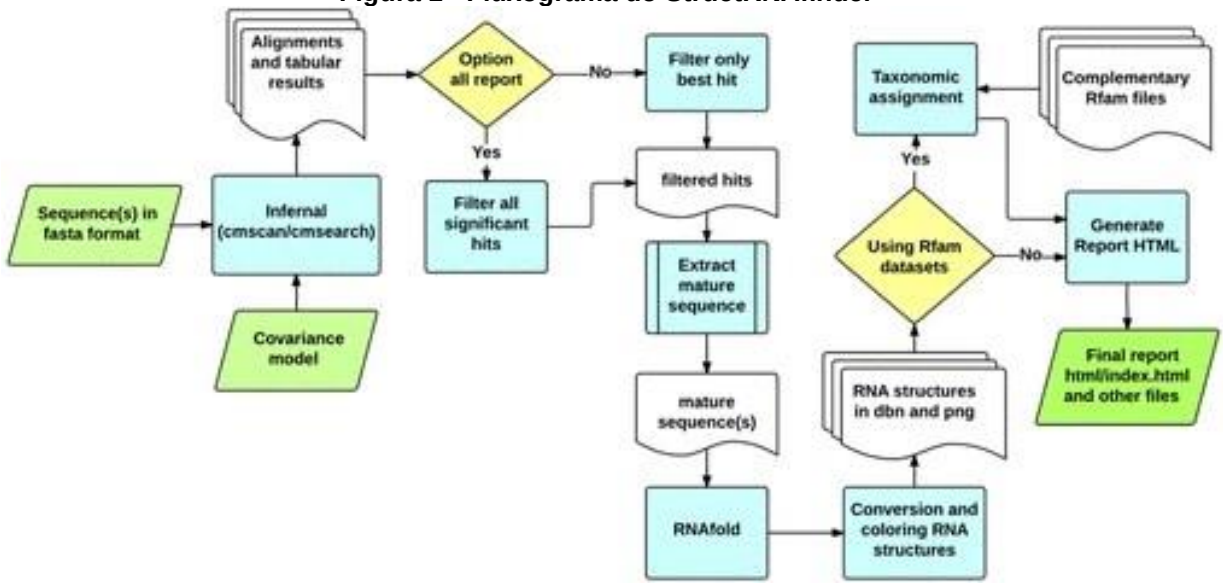
1.1 Revisão da literatura: Abordagens computacionais para caracterização de RNA não-codificante

1.1.1 StructRNAfinder

O *StructRNAfinder* (ARIAS-CARRASCO et al., 2018) é uma ferramenta que automatiza a execução de outras duas ferramentas (INFERNAL e RNAfold), com o objetivo de facilitar a identificação e anotação de ncRNA, utilizando como dado de entrada sequências genômicas ou transcriptômicas. A instalação da ferramenta também é automatizada, que através de um arquivo já é instalado o INFERNAL e RNAfold, e é realizado o *download* do banco de dados.

O primeiro processo da execução da ferramenta é a etapa de anotação, utilizando a ferramenta INFERNAL, junto com o banco de dados Rfam através de modelos de covariância. Com o resultado do INFERNAL, a ferramenta realiza uma filtragem dos resultados mais significantes de acordo com os valores do *e-value, bit score* ou filtro do modelo de covariância, para em seguida serem usadas como entrada na ferramenta RNAfold, que realiza a predição das estruturas secundárias. Por fim, é entregue uma tabela com os resultados do INFERNAL, RNAfold e informações taxonômicas. A Figura 1 é o fluxograma da ferramenta.

Figura 1 - Fluxograma do StructRNAfinder



Fonte: (ARIAS-CARRASCO et al., 2018)

A ferramenta foi desenvolvida em PERL para o sistema operacional Linux, realizando a integração dos programas, e criando relatórios com os resultados. Ela pode ser encontrada para download no próprio site (https://structrnafinder.integrativebioinformatics.me/index.html), e através do terminal é possível realizar sua instalação com apenas um comando. Ainda pelo terminal o usuário executa a ferramenta inserindo como parâmetro (i) o arquivo FASTA, (ii) o modelo de covariância; e opcionalmente (iii) o valor de corte para filtragem do *e-value* ou *bit score*, (iv) uma opção para o reportar todos os resultados mais significantes, (v) e a opção para realizar a busca nas duas fitas da sequência (negativa ou positiva) ou em apenas uma. Também foi desenvolvida uma página Web para a ferramenta, onde o usuário pode utilizar arquivos de sequência de até 10 megabytes.

1.1.2 FindNonCoding

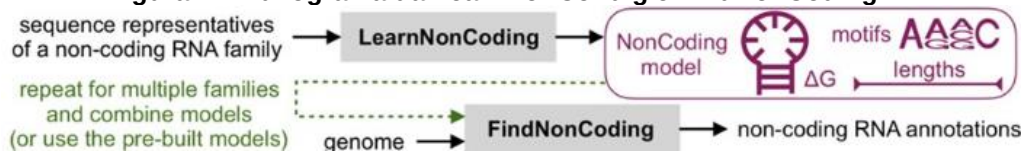
O FindNonCoding (WRIGHT, 2022) é uma ferramenta desenvolvida para anotar ncRNAs, feita com o principal objetivo de ser fácil de utilizar. De acordo com Wright (2022), o motivo é que a maioria das ferramentas de anotação são específicas para uma família de ncRNA, não são fáceis de utilizar.

A ferramenta foi desenvolvida na linguagem de programação R e distribuída no pacote DECIPHER (WRIGHT, 2016), que também pertence ao autor do FindNonCoding. Este pacote conta com diversas ferramentas para análises de DNA, RNA e proteínas. Para a execução da ferramenta, é necessário apenas um genoma

ou transcriptoma no formato FASTA, pois a própria ferramenta possibilita a utilização de modelos pré-treinados de *Archaea, Bacteria* ou *Eukarya*, para serem utilizados como banco de dados.

Caso o usuário deseje utilizar o próprio banco de dados, ele deverá antes utilizar uma outra ferramenta do pacote DECIPHER chamada LearnNonCoding, que criará um modelo com as sequências, para assim ser utilizado no FindNonCoding. Para isso, o usuário deverá selecionar um alinhamento múltiplo de sequências, de uma mesma família de ncRNA, e a LearnNonCoding extrairá quatro características desse arquivo: (i) uma matriz com os *motifs*, (ii) os *hairpins* e pseudo-nós que conservam a estrutura secundária, (iii) perfis de *k-mer,* e (iv) a distribuição dos comprimentos das sequências. Com essas características, a ferramenta cria um modelo para as sequências que representam aquela família do ncRNA. Esse processo está resumido no fluxograma da figura 2.

Figura 2 - Fluxograma da LearnNonCondig e FindNonCoding



Fonte: (WRIGHT, 2022)

O FindNonCoding necessita de cerca de sete comandos para analisar um genoma ou transcriptoma. Porém, se o usuário desejar utilizar um banco de dados diferente, ele precisará passar pelo treinamento do LearnNonCoding, o que aumentaria a quantidade de comandos necessários. O resultado do FindNonCoding é apenas uma tabela com as coordenadas das sequências e *scores* obtidos, e outra tabela com as respectivas famílias de ncRNAs encontradas. No manual é ensinado um comando para salvar as sequências encontradas, porém o manual não mostra nenhuma forma de salvar essas sequências, ou ainda a tabela com os resultados, em um arquivo texto, sendo necessário um conhecimento prévio do usuário na linguagem de programação R.

1.2 Objetivos do trabalho

1.2.1 Objetivo geral

Desenvolver um sistema automático de anotação de RNAs não-codificantes de forma que seja simples e fácil de utilizar.

1.2.2 Objetivos específicos

- Realizar a anotação de dados biológicos utilizando ferramentas de busca estrutural e busca por similaridade.
- Analisar os resultados de ambas as ferramentas, integrar os resultados e remover redundâncias.
- Automatizar o sistema de forma a facilitar sua utilização.
- Apresentar os resultados através de relatórios, gráficos e arquivos (FASTA, GFF e CSV) para uma posterior análise dos resultados.
- Validar o sistema com estudos de caso e em genomas retirados de bancos de dados públicos, de forma a colaborar com outros estudos envolvendo anotação de genomas.
- Desenvolver uma interface gráfica web de forma a tornar o acesso a ferramenta amigável e fácil ao usuário final.

2 CONCEITOS

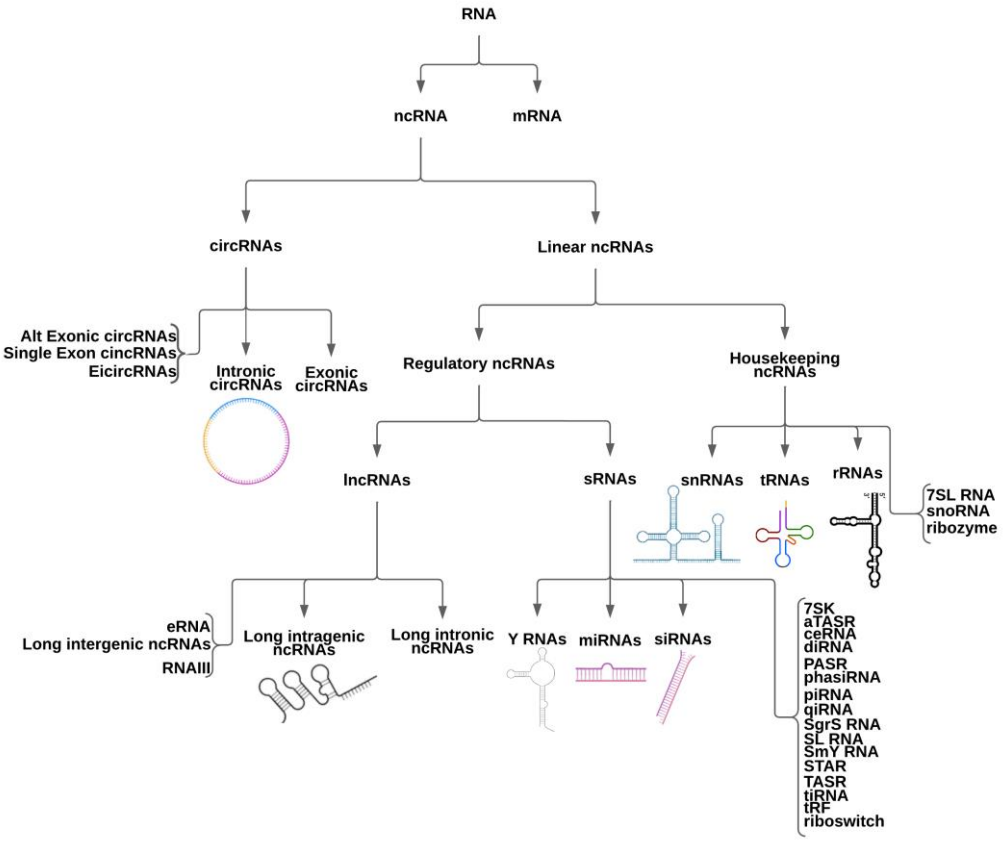
2.1 Conceitos biológicos

2.1.1 RNAs não-codificantes

Os ncRNAs, RNAs que são transcritos, mas que não codificam proteínas, desempenham funções importantes na regulação e na estrutura das células, além de representarem a maior parte do genoma dos mamíferos (MATTICK; MAKUNIN, 2006).

Os ncRNAs podem ser classificados de várias formas, como por exemplo: a natureza molecular, se é de fita simples ou dupla; pelo tipo, se é circular ou linear; pelo papel, se é catalítico, de armazenamento, ou regulador; pelo tamanho, pequenos ou longo (menores ou maiores que 200 nucleotídeos); processo biológico, se participa da tradução, duplicação do DNA, entre outros; pela estrutura secundária; ou sentido da fita (senso ou antisenso) (PEREIRA, 2017). Na Figura 3, podemos observar uma forma de classificar os RNAs, sendo que os ncRNAs lineares (do inglês *Linear ncRNAs*) podem ser divididos pelo seu papel regulatório (do inglês *Regulatory ncRNAs*) e de manutenção (do inglês *Housekeeping RNAs*). Os RNAs regulatórios incluem pequenos RNAs não codificantes (menores que 200 nucleotídeos) e longos RNAs não codificantes (maiores que 200 nucleotídeos), enquanto os RNAs de manutenção estão envolvidos em processos ribossomais (CHEN et al., 2019).

Figura 3 - Classificação de RNAs



Fonte: Adaptado de PEREIRA (2017) e Camilo Rebolledo et al. (2023)

Os RNAs de manutenção são altamente expressos e com importantes funções, como por exemplo o RNA ribossômico (rRNA) que participa da síntese proteica, ou o RNA transportador (tRNA) que funciona como um adaptador de RNA para ligar os mRNAs em suas cadeias peptídicas durante a etapa de tradução de mRNAs (PEREIRA, 2017).

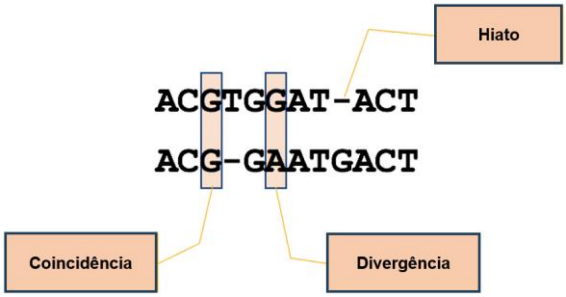
Já os RNAs regulatórios apresentam funções importantes na normalização de processos que envolvem outras moléculas ou organelas, como por exemplo os longos RNAs não-codificantes (lncRNAs), que em uma das suas funções ajuda na regulação da expressão gênica, se ligando a proteínas, DNA, enzimas modificadoras de cromatina ou outros RNAs (MATTICK et al. 2023; PEREIRA, 2017).

2.1.2 Alinhamento de sequências

O alinhamento de sequências é o processo de comparar sequências de DNA, RNA ou proteínas, com o objetivo de identificar regiões de similaridade entre as sequências (CHAO, 2022; HAQUE et al. 2009). Na Figura 4 está representado um

exemplo de alinhamento entre duas sequências, sendo coincidências (do inglês *matches*) os nucleotídeos que alinharam corretamente com o mesmo nucleotídeo, divergências (do inglês *mismatches*) são os nucleotídeos que alinharam com outro nucleotídeo e hiatos (do inglês *gaps*) são os espaços criados para melhorar o alinhamento (CHAO, 2022; HAQUE et al. 2009).

Figura 4 - Exemplo de alinhamento de duas sequências



Fonte: Traduzida de CJ Bioscience, Inc, (2017)

O alinhamento realizado na Figura 4, é um exemplo de alinhamento par a par, onde são comparadas apenas duas sequências, possuindo três formas de ser realizado: (i) global, (ii) semi-global e (iii) local (PIJARI e KANADAM, 2022).

O alinhamento global é quando as duas sequências são completamente alinhadas, levando em consideração todas as bases das duas sequências e o hiatos existentes; já o semi-global é igual ao global, com a diferença de não penalizar as bases que não são alinhadas no começo e no final; e o local identifica as regiões mais conservadas entre as sequências, identificando apenas as regiões que mais apresentaram coincidências de bases (PIJARI e KANADAM, 2022).

Para realizar esses alinhamentos de forma computacional, foram desenvolvidos algoritmos baseados em programação dinâmica pois é garantido um alinhamento otimizado, como por exemplo o algoritmo de Needleman-Wunsch (NEEDLEMAN e WUNSCH, 1970) para alinhamentos globais e semi-globais, e Smith-Waterman (SMITH e WATERMAN, 1981) para alinhamento local (DAILY, 2016; PIJARI e KANADAM, 2022).

Além do alinhamento par a par, outro tipo é alinhamento múltiplo de sequências, onde pelo menos três sequências de tamanhos aproximadamente iguais são alinhadas simultaneamente, e são utilizadas para inferir relações evolucionárias (EDGAR,2022). O principal desafio do alinhamento múltiplo é adaptar aos diferentes tamanhos das sequências, através de técnicas para adicionar novos nucleotídeos (SHEN, ZAHARIAS e WARNOW, 2022).

2.1.3 Anotação genômica

A anotação genômica é o processo utilizado para identificar e caracterizar os elementos funcionais presentes no genoma, existindo vários processos diferentes, mas que possuem a mesma essência, sendo dividida em duas etapas: computacional e curagem (YANDELL e ENCE, 2012).

Na etapa computacional é realizada primeiramente a identificação de repetições e mascaramentos, de forma a melhorar a qualidade da anotação, sendo possível identificar elementos transponíveis e repetitivos, e realizar o mascaramento dessas regiões (substituição das regiões repetitivas por bases “N”), utilizando por exemplo a ferramenta RepeatMasker (SMIT, HUBLEY e GREEN, 2010) (EJIGU e JUNG, 2023; YANDELL e ENCE, 2012). Em seguida, é realizado o alinhamento do genoma com informações conhecidas de proteínas, RNA-seq ou ncRNAs, de forma a identificar proteínas, transcritos ou ncRNAs do genoma (EJIGU e JUNG, 2023; YANDELL e ENCE, 2012). O alinhamento pode apresentar muitos resultados, por isso é importante filtrar de acordo com graus de similaridade ou identidade, de forma a remover resultados com baixa qualidade, e remover resultados com sobreposições no alinhamento, uma vez que representam uma mesma região do gene (YANDELL e ENCE, 2012). Além disso, muitos RNAs não-codificantes apresentam sua estrutura secundária conservada, podendo ser outra forma de identificar novos ncRNAs nos genomas, através da ferramenta INFERNAL (NAWROCKI; EDDY, 2013), por exemplo (NAWROCKI, 2013).

Na etapa de curagem é realizada a validação baseada em evidências com predição *ab initio*, sendo normalmente um procedimento manual, mas que devido ao alto custo vem sendo automatizada, através de ferramentas como EVidenceModeler (HAAS et al., 2008) (EJIGU e JUNG, 2023; YANDELL e ENCE, 2012). Dessa forma, é possível confirmar as anotações realizadas, pois serão identificadas regiões conservadas que tem suas funções associadas.

Com as anotações validadas, os resultados podem ser armazenados em quatro formatos diferentes: GFF3 (do inglês *General feature format* 3), GTF (do inglês *Gene transfer format*), GenBank e formatos do EMBL; de forma a facilitar o entendimento das anotações em mais de uma ferramenta, e permitir que outras

pessoas consigam utilizar essas anotações (EJIGU e JUNG, 2023; YANDELL e ENCE, 2012).

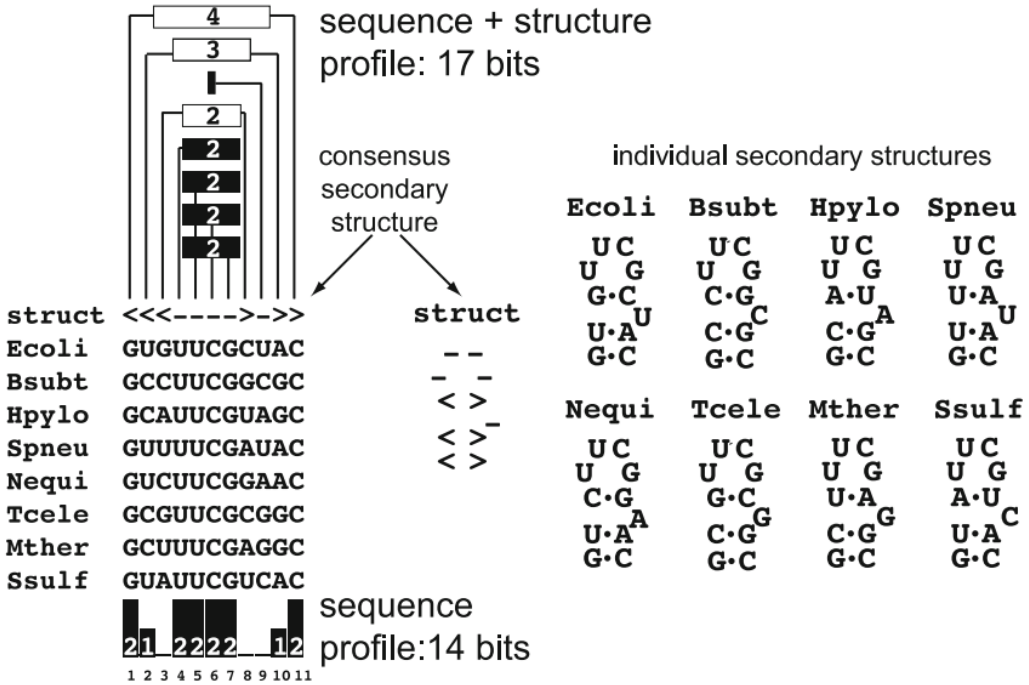
2.1.4 Modelo de covariância

Um modelo de covariância é definido como uma gramática probabilística livre de contexto estocástica para determinar uma relação entre duas variáveis, semelhantes a perfis de Hidden Markov Models (HMM) (NAWROCKI; EDDY, 2013).

Na bioinformática, um modelo de covariância representa a estrutura primária e secundária consenso de um RNA, que facilita a predição de estruturas secundarias e o alinhamento múltiplo de sequências (EDDY; DURBIN, 1994). Através desses modelos, é possível identificar estruturas similares em bancos de dados, para identificar se os RNAs pertencem a uma mesma família (NAWROCKI; EDDY, 2013).

Na Figura 5 temos um exemplo de oito sequências de RNAs de uma mesma família, que foram alinhados de forma a identificar uma estrutura secundária consenso. A representação da estrutura secundária é através dos símbolos “<” e “>” quando existe um pareamento de nucleotídeos, e pelo “-“ quando há uma ausência de pareamento. No exemplo, podemos observar que as oito espécies possuem diferentes sequências primárias, mas com uma estrutura secundária idêntica.

Figura 5 - Exemplo de alinhamento para criação de modelo de covariância



Fonte: NAWROCKI e EDDY (2013)

O *score* utilizado para avaliar o alinhamento de sequências com um modelo de covariância, é diferente dos *scores* utilizados por ferramentas como BLAST, pois os modelos de covariância levam em consideração probabilidades para alinhar cada nucleotídeo da sequência com cada posição do modelo (NAWROCKI; EDDY, 2013).

2.2 Ferramentas de anotação de ncRNAs

2.2.1 Basic Local Alignment Search Tool (BLAST)

O BLAST é um conjunto de ferramentas que realizam o alinhamento local de sequências de forma par-a-par, com o objetivo de encontrar sequências similares (ALTSCHUL et al., 1990).

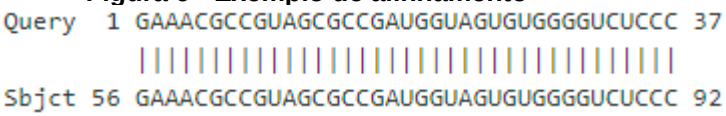
Neste projeto, foi utilizada a ferramenta BLASTn, que realiza o alinhamento local das sequências genômicas e transcriptomas de nucleotídeos contra um banco de dados de nucleotídeos, de forma a identificar outros RNAs. O programa pode ser utilizado na versão *stand-alone* ou na versão WEB (o que limita o tamanho dos arquivos).

O BLASTn realiza diversos cálculos para validar a significância dos resultados, dentre eles é importante realçar a cobertura, identidade e *e-value*. De acordo com Newell (2013), a cobertura é a porcentagem da sequência de entrada que foi alinhada com relação ao tamanho da sequência do banco, sendo assim, quanto maior a cobertura, mais confiável é o alinhamento. Já a identidade representa a quantidade de nucleotídeos idênticos entre a sequência de entrada e do banco (Newell et al., 2013). Além da cobertura e identidade, podemos também levar em consideração o *e-value*, que de acordo com Newell (2013) é um parâmetro de confiança do alinhamento, que descreve o número de acertos ao acaso que podem ser alinhados no banco de dados, sendo quanto mais próximo o *e-value* de zero, melhor o alinhamento.

No entanto, é crucial compreender que uma sequência 100% idêntica pode não ser representativa do RNA para aquela espécie, especialmente se for uma sequência de nucleotídeos pequena (por exemplo 20 nucleotídeos), que alinhou com uma sequência muito grande (por exemplo 200 nucleotídeos) do banco de dados, dessa forma ela vai possuir uma cobertura baixa. Por isso, é importante combinar os parâmetros cobertura e identidade, pois assim é maior a probabilidade de o RNA ser o esperado. Como por exemplo na Figura 6 temos um alinhamento de uma sequência

de entrada (*query*) e uma sequência de um banco de dados (*Sbjct* ou *subject*). Podemos observar na figura que todos os nucleotídeos da sequência de entrada são iguais aos nucleotídeos da sequência do banco de dados, ou seja, temos um alinhamento 100% idêntico. Porém, é notável a presença do número “56” no início da sequência do banco de dados, que indica que o alinhamento começou no nucleotídeo de número 56 da sequência, então a sequência de entrada não cobre toda a sequência do banco de dados, ou seja, não possui 100% de cobertura.

Figura 6 - Exemplo de alinhamento



Fonte: Captura de tela do RNACENTRAL CONSORTIUM (2021)

2.2.2 INFERence of RNA Alignment (INFERNAL)

O INFERNAL é um conjunto de programas que realiza busca estrutural contra um banco de dados contendo modelos de covariância. O modelo utilizado pelo INFERNAL é como um perfil das sequências, representando um alinhamento múltiplo de sequências consenso de famílias de ncRNAs, e estrutura secundária da família do ncRNA (Sean, 1994). Um banco de dados com modelos de covariância de ncRNA amplamente utilizado é o Rfam, que conta com 4108 famílias de ncRNAs (KALVARI et al., 2020).

Dentre as ferramentas do INFERNAL, será utilizada neste projeto a cmscan, que determina se uma sequência possui similaridade, através da estrutura secundária, com alguma família de ncRNA do modelo de covariância, gerando dois arquivos como resultado. O primeiro arquivo é uma lista dos acertos, apresentando por exemplo as coordenadas de início e fim, e a pontuação de acordo com o modelo de covariância. O segundo arquivo já apresenta todos os possíveis acertos para a sequência de entrada, além de apresentar o alinhamento das estruturas secundárias.

Dentre os parâmetros que o INFERNAL permite utilizar, vale ressaltar o –cut_ga que é um filtro para determinar quais acertos serão relatados, que leva em consideração os limites curados e confiáveis informados pelo modelo de covariância do Rfam e o *bit score* (NAWROCKI et al., 2015).

O *bit* score é um valor determinado pela ferramenta para cada sequência similar que encontra, junto com outras informações, como o início e fim da sequência. O *Bit score* é uma métrica que corresponde o quanto a sequência de entrada (*query*)

é parecida com o modelo, e quanto maior o *bit score,* mais a sequência se identifica ao modelo (KALVARI, et al, 2020). A equação 1 é a fórmula para determinar o *Bit score,* sendo P_{cm} a probabilidade da sequência (*query*) dada o modelo, e P_{null} a probabilidade da sequência (*query*) dado o modelo nulo, ou seja, probabilidade de a sequência não ser parecida com o modelo (KALVARI, et al, 2020).

$$bit\ score = log_2\left(\frac{P_{cm}}{P_{null}}\right) \qquad (1)$$

3 ARTIGO

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome analysis

VitorTool: Sistema automático para anotação de RNAs não-codificantes

Vitor Gregorio[1], Bruno Thiago de Lima Nichio[1], Flávia Cristina de Paula Freitas[3], Liliane Santana Oliveira[1], Pedro Gabriel Nachtigall[4] and Douglas Silva Domingues[1,2], Alexandre Rossi Paschoal [1*]

[1]Universidade Tecnológica Federal do Paraná, UTFPR, Cornélio Procópio. [2]University of São Paulo, ESALQ/USP, Piracicaba, SP, Brasil. [3]Universidade Federal de São Carlos, UFSCAR, São Carlos.

[4]CeTICS, Instituto Butantan, São Paulo.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation:

Results:

Availability:

Contact:

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introdução

A anotação genômica é um processo usado para identificar e caracterizar os elementos funcionais presentes no genoma (YANDELL; ENCE, 2012). Do ponto de vista computacional, é comumente utilizada a busca por similaridade via sequência para identificar genes em dados biológicos (STEIN, 2001; FU et al. 2021). A busca por similaridade é o processo para comparar duas sequências através do alinhamento, onde são observados os nucleotídeos correspondentes, substituídos e com hiatos, para assim definir uma pontuação para esse alinhamento (Altschul et. al. 1990).

No caso dos RNAs não-codificantes (ou do inglês non-coding RNAs ou apenas ncRNA), que são RNAs que são transcritos, mas não traduzidos em proteínas (MATTICK; MAKUNIN, 2006), eles apresentam a sua estrutura secundária mais conservada do que sua estrutura primária (EDDY e DURBIN, 1994). Nesse contexto, a busca estrutural emerge como uma abordagem relevante, classificando as sequências de acordo com suas estruturas secundárias, e atribuindo pontuações com base na semelhança das estruturas e *hairpins loops* formados (NAWROCKI; EDDY, 2013).

De forma a facilitar a anotação de ncRNAs, foram desenvolvidas algumas ferramentas que são específicas para um tipo de ncRNA, como tRNAscan-SE (CHAN, 2019) ou sRNAtoolbox (APARICIO-PUERTA et al., 2022). A ferramenta tRNAscan-SE é exclusiva para identificar tRNAs, e a ferramenta sRNAtoolbox é exclusiva para sRNA.

Além disso, existem duas ferramentas que buscam anotar ncRNAs em geral, que são StructRNAfinder (ARIAS-CARRASCO et al., 2018) e FindNonCoding (WRIGHT, 2022). Porém, nenhuma das duas ferramentas integra os formatos de busca estrutural e por similaridade simultaneamente. Além disso, a ferramenta FindNonCoding não apresenta relatórios com os resultados de forma simples e de fácil acesso do usuário, além de não apresentar nenhuma interface (Web ou Desktop), sendo necessário um conhecimento prévio na linguagem de programação R.

Sendo assim, existe a falta de um sistema que integre de forma simples e usual, através de interface e gráficos, essas formas de identificação de ncRNA e a grande quantidade de resultados heterogêneos que elas apresentam.

Dado esse cenário, o objetivo do presente projeto foi criar um sistema automático, amigável e escalável para anotação em larga escala de RNAs não-codificantes, utilizando as duas abordagens de busca de ncRNAs. Para validar o sistema, nós utilizamos genomas disponibilizados em bancos de dados públicos (RefSeq, Ensembl Plants e GENCODE), além de ser testado em estudos de casos de trabalhos de colaboradores. Por fim, este sistema será disponibilizado para a comunidade científica, com uma interface WEB e para a plataforma Galaxy, de forma a incentivar os pesquisadores a realizarem anotações de genomas.

2 Ferramentas comparativas para anotação de ncRNA

Nesta seção, iremos explorar duas ferramentas desenvolvidas para a anotação de RNAs não-codificantes.

StructRNAfinder

O StructRNAfinder é uma ferramenta que automatiza a execução de outras duas ferramentas (INFERNAL e RNAfold (LORENZ et al., 2011)), com o objetivo de identificar e anotar famílias de RNA em dados de entrada que podem ser sequências genômicas ou transcriptômicas (ARIAS-CARRASCO et al., 2018). Nesta abordagem, inicialmente realiza-se o alinhamento das sequências com perfis de famílias de RNA utilizando o INFERNAL, junto com o banco de dados Rfam. Com os resultados obtidos, a ferramenta realiza uma filtragem dos resultados mais significantes de acordo com o e-value ou score dos resultados que, posteriormente, são utilizados como entrada na ferramenta RNAfold, para a predição das estruturas secundárias. Por fim, são entregues algumas tabelas e figuras com os resultados do INFERNAL e do RNAfold. Além da versão *stand-alone*, foi desenvolvida uma página *Web* para a ferramenta, onde o usuário pode utilizar arquivos de sequência de até 10 megabytes.

FindNonCoding

De acordo com Wright (2021), o FindNonCoding é uma ferramenta desenvolvida com o principal objetivo de ser fácil de utilizar. Ela foi desenvolvida na linguagem de programação R e hospedada no pacote DECIPHER, que também pertence ao autor do FindNonCoding. Este pacote conta com diversas ferramentas para análises de DNA, RNA e proteínas. Para a execução da ferramenta, é necessário um genoma ou transcriptoma no formato FASTA, e, caso o usuário desejar, pode utilizar um outro arquivo FASTA como banco de dados, porém, a própria ferramenta disponibiliza modelos pré-treinados para Archaea, Bacteria ou Eukarya. Caso o usuário deseje utilizar um banco de dados próprio, ele deverá antes utilizar uma outra ferramenta do pacote DECIPHER chamada LearnNonCoding, que cria modelos a partir de um arquivo FASTA, para posteriormente ser utilizado no FindNonCoding. De acordo com Wright (2021), a ferramenta que ele desenvolveu é diferente das outras pois analisa as extremidades do RNA em vez de olhar por toda a sequência ou estrutura. Além disso, quando é encontrado mais de um resultado na mesma região, a ferramenta seleciona apenas o resultado que apresentar a maior pontuação. O resultado do FindNonCoding é uma tabela com as coordenadas das sequências e pontuações obtidas, e outra tabela com as respectivas famílias de ncRNAs encontradas. No manual é ensinado um comando para salvar as sequências de ncRNA encontrados, porém não há nenhuma forma de salvar essas sequências ou a tabela com os resultados em um arquivo texto, sendo necessário um conhecimento prévio do usuário na linguagem de programação R.

3 Metodologia

A Figura 1 representa o fluxograma resumido utilizada para a construção do sistema, começando pela realização das buscas por similaridade e estrutural, seguindo de uma pós-análise para filtrar os resultados dos alinhamentos. O fluxograma completo está ilustrado na figura S1, disponível no arquivo suplementar.

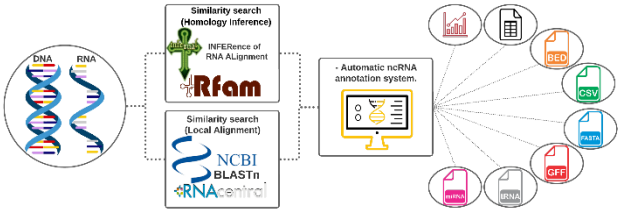


Figure 1- Fluxograma resumido da ferramenta

Estratégia de Busca por similaridade

Para a busca por similaridade, utilizamos a ferramenta BLASTn (Altschul et. al. 1990) utilizando como banco de dados o RNACentral (RNACENTRAL CONSORTIUM, 2021). Neste projeto, foi utilizado a versão stand-alone, que permite uma maior manipulação dos parâmetros de execução da ferramenta, como por exemplo, mudar a quantidade de núcleos do processador que a ferramenta irá utilizar, ou o formato e as colunas do arquivo com os resultados.

O BLASTn realiza cálculos para validar a significância dos resultados, dentre eles é importante realçar a cobertura, identidade e e-value. No entanto, é crucial compreender que quando um alinhamento possui 100% de identidade, pode não ser representativa para aquela espécie, especialmente se for uma sequência de nucleotídeos pequena, resultando em uma cobertura percentual reduzida. Por isso, é importante combinar os parâmetros de cobertura e identidade, pois assim é maior a probabilidade do ncRNA ser o esperado.

Por padrão, neste projeto, apenas os resultados que apresentarem cobertura e identidade superiores a 95% são salvos, podendo ser ajustados pelo usuário conforme desejado. Este valor foi determinado através de testes com diferentes valores de cobertura e identidade (50%, 60%, 70%, 80%, 90%, 95%, 98% e 100%), onde anotamos o genoma de *Chlamydia trachomatis* e comparamos os resultados entre si. No caso, nossa ferramenta apresentou a mesma anotação final para os oito testes, indicando que independente da porcentagem, o melhor resultados estava presente nos oito testes. Dessa forma, determinamos o filtro em 95%, por ser um valor alto de similaridade, mas não tão severo como 98% ou 100%.

Estratégia de busca estrutural

Para a identificação de ncRNAs por busca estrutural, foi utilizada o programa INFERNAL (Inference of RNA alignments) (NAWROCKI; EDDY, 2013) (versão 1.1.4), que são um conjunto de programas desenvolvidos para analisar sequências utilizando perfis de sequências de RNA. Esses perfis são denominados modelos de covariância, que representam sequências consenso na estrutura primária e secundária do RNA (Sean, 1994). Esses modelos de covariância são construídos a partir de sequências relacionadas e dessa forma é possível identificar padrões entre as sequências, onde a alteração de um nucleotídeo está relacionada a alteração de outro nucleotídeo em outra posição, e desses padrões, são criados modelos estatísticos que descrevem as probabilidades para variações de uma sequência (KALVARI et al., 2020).

Dentre as ferramentas do INFERNAL, será utilizada neste projeto a *cmscan*, que determina se uma sequência possui similaridade, através da estrutura secundária, com alguma família de ncRNA do modelo de covariância e gera uma lista classificando os acertos com pontuação e com os alinhamentos.

Da mesma forma que o BLAST, o INFERNAL também permite realizar alterações de parâmetros ao executar a ferramenta. Uma opção é o parâmetro –cut_ga que é um filtro para determinar quais acertos serão relatados, levando em consideração o limite informado pelo modelo de covariância do Rfam e o bit-score.

Os resultados são salvos em dois arquivos, onde no primeiro é uma tabela com a classificação de cada sequência, ordenada pela similaridade, e no segundo, estão as sequências ao lado de sua sequência alvo (sequência que apresentou o maior grau de similaridade).

Banco de dados

Para a execução do INFERNAL, foi utilizado o Rfam (KALVARI et al., 2020) na versão 14.9, e para o BLAST foi utilizado o RNAcentral (RNACENTRAL CONSORTIUM, 2021) na versão 21.

O Rfam é um banco de dados curado de modelos de covariância caracterizando famílias de ncRNAs, que estão representadas por alinhamentos múltiplos de sequências, estruturas secundárias consenso e modelos de covariância, para anotar ncRNAs em datasets de nucleotídeos. O Rfam é o principal banco de dados de anotação em sites como ENSEMBL, NCBI Procariotos e Eukaryotic Gene Anotation. Atualmente, o Rfam conta com 4108 famílias de ncRNAs.

Já o RNAcentral é um banco de RNAs não codificantes redundantes que conta com sequências oriundas de 49 bancos de dados diferentes, disponibilizando mais de 30 milhões de sequências. O principal objetivo deste banco é garantir acesso aberto a um conjunto abrangente de sequências de ncRNAs para uma grande variedade de espécies, permitindo ao usuário acessá-los e baixá-los, para, por exemplo, realizar o treinamento de ferramentas ou comparar com resultados de análise RNA-seq.

Desenvolvimento da ferramenta

Nossa ferramenta foi desenvolvida na linguagem de programação Python (versão 3.9), que inicialmente executa as ferramentas BLAST e INFERNAL, e utiliza os resultados de ambas para a pós-análise.

Os resultados são ordenados de acordo com alguns atributos, de forma a otimizar a execução da nossa ferramenta, sendo no caso do BLAST: (i) nome do cromossomo, (ii) maior identidade, (iii) maior cobertura, (iv) menor valor esperado (expect value – E-value), (v) coordenada de início e (vi) coordenada de fim. E no caso do INFERNAL: (i) nome do cromossomo, (ii) maior bit-score, (iii) menor valor esperado (expect value – E-value), (iv) coordenada de início e (v) coordenada de fim. Em seguida são removidos os resultados duplicados, ou seja, resultados que apresentarem os mesmos valores para esses atributos, uma vez que ambas as ferramentas podem apresentar resultados duplicados.

Para o BLAST, essa análise é desenvolvida levando em consideração o sentido da fita, dessa forma os resultados são separados em fitas positivas (*sense*) e negativas (*anti-sense*), uma vez que os ncRNAs podem apresentar diferentes funções dependendo do sentido da fita. Como o INFERNAL realiza a busca estrutural e leva em consideração apenas a estrutura formada, não é realizada a separação dos resultados.

Além disso, utilizamos os Ids do RNAcentral e Rfam para criar um dicionário dos ncRNAs, dessa forma podemos determinar qual é o RNA não-codificante de cada resultado obtido pelas ferramentas.

Em seguida, é realizado o tratamento de regiões sobrepostas. Nesse sentido, teremos casos de resultados sem sobreposição, que chamaremos de “únicos”. Já as sobreposições podem acontecer de cinco formas, como mostra a Figura 2. Utilizando como exemplo a sequência A da Figura 2, ela pode apresentar uma sobreposição completa quando as coordenadas de início e fim são iguais, como representado na sequência B. Já a sequência C é um exemplo de uma sobreposição quando as coordenadas de início são iguais, mas o de fim é diferente. A sequência D é o inverso de C, no caso as coordenadas de fim são iguais, mas o de início é diferente. Nas sequências E e F ocorre quando as coordenadas de início ou de fim estão dentro da sequência A, ocorrendo também uma sobreposição. Por fim, temos a sequência G que é uma sobreposição que está dentro da sequência A.

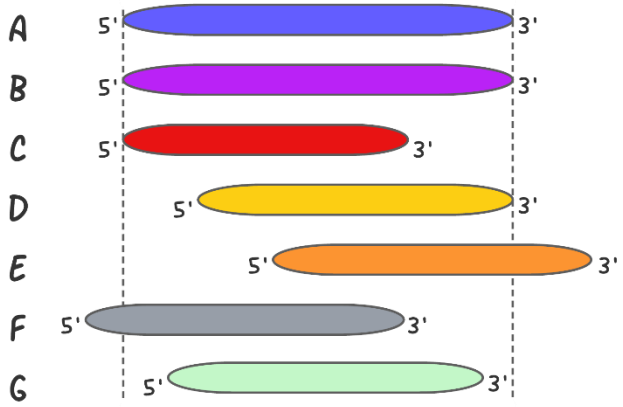


Figure 2 - Exemplos de sobreposição

Quando acontece alguma dessas sobreposições, nossa ferramenta realiza a união das coordenadas, mantendo apenas um resultado com a menor coordenada de início e a maior de fim, além de verificar se as sequências identificaram os mesmos ncRNAs e, caso sejam diferentes, será definido como “ncRNA”. O usuário poderá acessar o arquivo com IDs originais para verificar todas as coordenadas e ncRNAs de regiões que tiveram sobreposição.

O tratamento de regiões sobrepostas é executado três vezes nessa primeira etapa, sendo duas vezes para o BLAST (fita positiva e negativa) e uma para o INFERNAL.

Em seguida, é realizada outra vez o tratamento de sobreposições, mas nesse caso será entre os resultados do BLAST com o INFERNAL, de forma a tratar as regiões que se sobrepõem entre as duas ferramentas, onde não é mais levado em consideração o sentido da fita. Além disso, de forma a indicar ao usuário que ocorreu um caso de sobreposição naquele resultado, o resultado será classificado como *both*, indicando que o resultado foi identificado nas duas ferramentas. Ainda, esses dois resultados estarão presentes em um arquivo com os resultados originais das ferramentas, identificados por IDs, que será abordado adiante.

A seguir, é realizada uma última filtragem opcional, que é a opção "melhor resultado", onde é salvo apenas o melhor resultado entre duas fitas. Ou seja, quando existir dois resultados com as mesmas coordenadas de início e fim, mas em fitas diferentes, será mantido o resultado que apresentar: (i) menor e-value, (ii) maior score. Por padrão, essa filtragem está ativada, mas o usuário pode optar por desativá-la ao executar a ferramenta.

Por fim, com todas as análises finalizadas, além de arquivos no formato TSV (tab-separated values), CSV (tab-separated values) e GFF3 (General Feature Format), serão apresentadas tabelas e figuras de forma a proporcionar uma rápida visualização dos resultados obtidos. Ficarão disponíveis: (i) arquivo com resultados originais, (ii) arquivo FASTA com os ncRNAs obtidos, (iii) tabela com a quantidade de cada ncRNA obtidas, (iv) arquivo contendo apenas os resultados com tRNA (CSV, TSV, GFF e FASTA), (v) arquivo contendo apenas os resultados com miRNA (CSV, TSV, GFF e FASTA), (vi) diagrama de Venn com a quantidade de resultados de cada ferramenta, (vii) gráfico de barras com a quantidade de cada ncRNA.

Métricas de avaliação de desempenho

De forma a validar os resultados obtidos pela ferramenta, foram selecionados dois arquivos com transcritos de RNAs de *Drosophila melanogaster* e *Saccharomyces cerevisiae*, e dois genomas de *Chlamydia trachomatis* e *Escherichia coli*, que foram retirados do banco de dados do RefSeq (O'Leary et al, 2016). Além disso, foram selecionados os genomas de *Arabidopsis thaliana* e *Oryza sativa*, que foram retiradas do banco de dados do Ensembl Plants (CUNNINGHAM et al, 2022). Por fim, também foi selecionado o genoma de *Homo sapiens* que foi obtido do GENCODE (FRANKISH, 2021).

Para os dados de transcritos, foram selecionadas as anotações mais recentes dos genomas disponíveis no site do NCBI, juntamente com o arquivo FASTA contendo todas as sequências de RNA derivadas do genoma, englobando tanto RNAs mensageiros quanto ncRNAs. Em seguida, essas sequências foram separadas entre ncRNA e mRNA, gerando os conjuntos de sequências positivas e negativas. Dessa forma, adotou-se uma abordagem de amostragem aleatória simples, a qual envolve a seleção de uma quantidade predefinida de amostras da população (PAULA, 2019). O objetivo da amostragem foi obter uma quantidade equivalente de ncRNAs e mRNAs de cada espécie. Desse modo, foram criados cinco arquivos FASTA para cada espécie, os quais foram utilizados como dados de entrada na nossa ferramenta, na StructRNAfinder e FindNonCoding. Os resultados de cada ferramenta foram comparados e quantificados com a anotação original disponibilizado nos bancos de dados de cada espécie.

Para a execução da ferramenta StructRNAfinder, utilizamos o banco de dados do Rfam na versão 14.9, no modo cmscan e com parâmetros no modo padrão. No final, utilizamos o arquivo no formato BED para comparar com a anotação original. Já para FindNonCoding, seguimos o tutorial disponibilizado no R, mudando apenas o modelo do próprio FindNonCoding de acordo com o domínio do genoma que utilizamos. Assim, a anotação era salva em um CSV para ser comparada com a anotação original.

Foram comparadas as anotações das ferramentas de acordo com o tipo do ncRNA que elas anotaram, levando em consideração todas as sequências de entrada, onde os mRNAs correspondem ao banco negativo. Sendo assim, foram classificadas como Verdadeiro Positivo (TP, do inglês True Positive) as anotações que acertarem o tipo do ncRNA que foi informado na anotação original, Falso Negativo (FN, do inglês False Negative) quando não identificaram a sequência ou anotaram o tipo do ncRNA errado para uma sequência que é ncRNA, Falso Positivo (FP, do inglês False Positive) quando classificaram como não-codificante uma sequência que originalmente é classificada como mRNA, e Verdadeiro Negativo (TN, do inglês True Negative) quando não identificaram uma sequência que é mRNA. Na Figura 3 está ilustrado um exemplo de como foi realizada essa matriz de confusão.

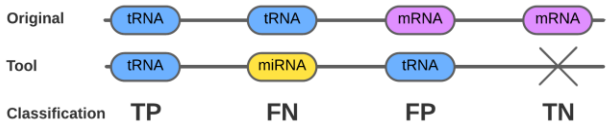


Figure 3 - Classificação utilizada para criar a matriz de confusão

Por fim, foi calculado os valores de sensibilidade (I), especificidade (II), acurácia (III), precisão (IV) e F1-score (V), pois dessa forma é possível identificar a proporção dos resultados negativos e positivos. Além disso, para os resultados dos transcritos de Drosophila melanogaster e Saccharomyces cerevisiae foi calculado a média e desvio padrão dos testes realizados.

$$Sensitivity = \frac{TP}{TP+FN} \quad \text{(I)}$$

$$Specificity = \frac{TN}{TN+FP} \quad \text{(II)}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad \text{(III)}$$

$$Precision = \frac{TP}{TN+FP} \quad \text{(IV)}$$

$$F1\ score = \frac{2TP}{2TP+FN+FP} \quad \text{(V)}$$

4 Experimentos realizados

Estudo de caso I: Identificação de ncRNAs de *L. donovani* em vesículas extracelulares

Por meio de uma colaboração internacional com aluna Camila dos Santos Meira, da Universidade de Calgary sob orientação do Prof. Lashitew Gedamu, nosso sistema foi testado no genoma de *L. donovani*. Este estudo, que foi desenvolvido no Canadá, teve como objetivo identificar ncRNAs de *Leishmania donovani* em Vesículas Extracelulares (do inglês *Extracellular Vesicles* - EVs). Para isso, o grupo do Canadá realizou um sequenciamento *single-end* (Illumina Hiseq 2500) de pequenos RNAs em EVs, e através de análises de RNA-seq realizaram a obtenção dos transcritos dessas amostras. Então, foram comparados perfis de elementos semelhantes a ncRNAs, em pequenos EVs de *L. donovani*, usando três tipos de vesículas de parasitos (tipo selvagem (LdCM), tipo nocaute para catepsina B (LdKO), e tipo nocaute para catepsina B com plasmídeo expressando catepsina B (LdWT)) (GERBABA e GEDAMU, 2018).

Sendo assim, o grupo disponibilizou os dados dos transcritos para serem avaliados em nosso projeto, com o intuito de caracterizar ncRNAs nesses dados de EVs em *L. donovani*. Deste modo, de forma a identificar

os tipos dos RNAs dos transcritos, foi primeiro realizada a análise do genoma da *Leishmania donovani* com a utilização da nossa ferramenta, para em seguida ser comparada com os resultados dos transcritos obtidos.

A re-anotação do genoma é importante para validar as anotações originais do genoma, uma vez que novas informações podem ser adicionadas com o tempo. Além disso, com a anotação do genoma, podemos comparar os ncRNAs encontrados em novos sequenciamentos com os antigos, e identificar possíveis mutações.

Estudo de caso II: Identificação de ncRNAs em *Theobroma*

O segundo estudo de caso é uma colaboração com o Prof. Dr. Douglas Silva Domingues da Universidade de São Paulo (USP), e Prof. Dr. Alessandro de Mello Varani, da Universidade Estadual Paulista (UNESP). Eles disponibilizaram dados com sequências de quatro genomas referentes ao cupuaçu (*Theobroma grandiflorum*). O objetivo do projeto foi realizar a identificação e anotação dos genomas recém sequenciados. Os quatro genomas são oriundos do clone 174 e 1074, que é resistente à vassoura de bruxa, sua principal doença. Sendo assim, elas apresentam 10 cromossomos cada, sendo um genoma haplóide e o outro diplóide de cada clone.

5 Resultados

Nesta seção serão apresentados os resultados obtidos pela nossa ferramenta, dos testes realizados para a validação da ferramenta, e os obtidos nos estudos de caso.

Comparação da nossa ferramenta contra ferramentas da área

Inicialmente, foram calculadas as métricas de desempenho na análise dos transcritos de RNA. Na tabela 1, é possível observar os valores médios da quantificação para as cinco análises de *Drosophila melanogaster*, que contam com 4579 ncRNAs, e *Saccharomyces cerevisiae*, que contam com 419 ncRNAs. Os resultados de cada análise estão disponíveis no arquivo suplementar, na tabela S1 para *D. melanogaster* e tabela S2 para *S. cerevisiae*.

Para *D. melanogaster*, nossa ferramenta apresentou 89,30% de sensibilidade, sendo a melhor entre as três ferramentas, enquanto a segunda melhor foi a StructRNAfinder com 18,34%. Além disso, nossa ferramenta apresentou os melhores resultados para acurácia, precisão e F1-score. No caso da especificidade, as três ferramentas apresentaram valores próximos, onde a FindNonCoding apresentou a melhor especificidade com 99,86%, mas não ficou muito longe da nossa, que apresentou 98,11%.

Já nos transcritos de *S. cerevisiae* a nossa ferramenta acertou todos os ncRNAs que existiam, apresentando uma sensibilidade de 100%. O FindNonCoding se destaca com uma melhor especificidade. Porém, a nossa ferramenta continua apresentando os melhores resultados nas outras métricas.

Em seguida, foram quantificados os resultados das anotações dos genomas. A tabela 2 foi construída de acordo com a quantidade de mRNAs e ncRNAs informada nas anotações de cada genoma. Na tabela 3, está representado a quantidade de ncRNA anotado pela nossa ferramenta, e na Figura S5 estão os gráficos de radar com as métricas de proporção de cada um dos genomas.

Tabela 1- Resultado das métricas de proporção para as análises dos transcritos de *Drosophila melanogaster* e *Saccharomyces cerevisiae*

Species	Tool	Sensitivity	Specificity %	Accuracy %	Precision %	F1-score %
D. melanogaster	VitorTool	**89,30**	98,11 ± 0,08	**93,70 ± 0,04**	**97,93 ± 0,09**	**93,41 ± 0,04**
	StructRNAfinder	18,34	93,47 ± 0,18	55,91 ± 0,09	73,74 ± 0,52	29,38 ± 0,04
	FindNonCoding	2,73	**99,86 ± 0,04**	51,30 ± 0,02	95,14 ±1,32	5,31 ± 0,00
S. cerevisiae	VitorTool	**100,00**	98,81 ± 0,45	**99,40 ± 0,22**	**98,82 ± 0,44**	**99,41 ± 0,22**
	StructRNAfinder	95,70	98,33 ± 0,56	97,02 ± 0,28	98,29 ± 0,56	96,98 ± 0,27
	FindNonCoding	3,58	**99,95 ± 0,11**	51,77 ± 0,05	98,75 ± 2,80	6,91 ± 0,01

Tabela 2- Quantidade de mRNAs e ncRNAs em cada genoma de acordo com a anotação

Species	mRNA	ncRNA
A. thaliana	27655	4871
C. trachomatis	888	45
E. coli	5155	126
H. sapiens	19393	25942
O. sativa	37960	1011

No genoma de *A. thaliana*, as ferramentas apresentaram um baixo valor de sensibilidade, principalmente o StructRNAfinder que encontrou apenas cinco ncRNAs corretos. Um motivo para apresentar esse resultado, é que na anotação apresenta vários lncRNAs, e o Rfam não é um

bom banco de dados para esse tipo de ncRNA, mostrando a importância de realizar a busca por similaridade além da busca estrutural.

Tabela 3- Matriz de confusão para os resultados da análise dos genomas nas três ferramentas

Species	Tool	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1-score
A. thaliana	VitorTool	3326	6137	21518	1545	**68,28%**	77,81%	76,38%	35,15%	**46,41%**
	StructRNAfinder	5	3	27652	5173	0,10%	**99,99%**	84,24%	62,50%	0,19%
	FindNonCoding	676	168	27487	4502	13,06%	99,39%	**85,78%**	**80,09%**	22,45%
C. trachomatis	VitorTool	45	1	887	0	**100,00%**	99,89%	**99,89%**	97,83%	**98,90%**
	StructRNAfinder	1	0	888	44	2,22%	**100,00%**	95,28%	**100,00%**	4,35%
	FindNonCoding	45	1	887	0	**100,00%**	99,89%	**99,89%**	97,83%	**98,90%**
E. coli	VitorTool	126	32	5123	0	**100,00%**	99,38%	99,39%	79,75%	88,73%
	StructRNAfinder	126	93	5062	0	**100,00%**	98,20%	98,24%	57,53%	73,04%
	FindNonCoding	126	1	5154	0	**100,00%**	**99,98%**	**99,98%**	**99,21%**	**99,60%**
H. sapiens	VitorTool	7847	6532	12861	18095	**30,25%**	66,32%	**45,68%**	**54,57%**	**38,92%**
	StructRNAfinder	96	563	18830	25846	0,37%	**97,10%**	41,75%	14,57%	0,72%
	FindNonCoding	-	-	-	-	-	-	-	-	-
O. sativa	VitorTool	593	1829	36131	418	**58,65%**	95,18%	94,23%	24,48%	**34,55%**
	StructRNAfinder	5	0	37960	1006	0,49%	**100,00%**	**97,42%**	**100,00%**	0,98%
	FindNonCoding	-	-	-	-	-	-	-	-	-

Para o genoma de *C. trachomatis*, nossa ferramenta apresentou exatamente o mesmo resultado que o FindNonCoding, onde ambas acertaram todos os ncRNAs. O StructRNAfinder apresentou apenas um resultado no final, e como nossa ferramenta também utiliza o INFERNAL como um dos programas para a anotação, achamos curioso esse resultado. Ao analisar o resultado do INFERNAL que o StructRNAfinder apresenta, podemos observar que existe mais de um resultado na anotação, porém, o resultado filtrado do StructRNAfinder apresenta apenas um resultado, sugerindo que, no momento da filtragem, o StructRNAfinder acaba removendo muitos resultados.

Já no caso do genoma de *E. coli*, todas as ferramentas acertaram todos os ncRNAs que existem na anotação original, a diferença dos resultados está nos erros, onde FindNonCoding errou menos e dentre as métricas apresentou o melhor resultado.

Para os genomas de *H. sapiens* e *O. sativa*, não foi possível executar o FindNonCoding, uma vez que apresentou um erro dizendo que as sequências eram muito grandes. Sendo assim, nossa ferramenta apresentou o melhor resultado quando comparado com o StructRNAfinder, apresentando uma sensibilidade muito superior.

De forma geral nossa ferramenta apresentou ótimos resultados de sensibilidade e não apresentou os piores resultados para especificidade. No caso de genomas grandes as três ferramentas não conseguiram alcançar altos valores de sensibilidade, podendo indicar a falta de anotações para essas espécies ou até mesmo anotações ruins. Além disso, durante os testes nossa ferramenta apresentou alguns resultados que não existiam na anotação original, ou seja, regiões que não tem nem mRNA e nem ncRNA anotado, podendo indicar um novo ncRNA naquela região.

Desempenho da nossa ferramenta em estudos de caso

I. Identificação de ncRNAs de *L. donovani* em vesículas extracelulares

No estudo de caso I, foi realizada a anotação do genoma de *Leishmania donovani* de forma a identificar os ncRNAs presentes neste genoma e em seguida a anotação de três amostras de transcritos de vesículas extracelulares de *L. donovani*.

Tabela 4- Comparação das anotações entre as amostras e o genoma de *L. donovani*

ncRNA	Genome	LdWT	LdKO	LdCM
misc_RNA	7955	54	56	53
snoRNA	128	90	84	78
tRNA	91	69	69	69
rRNA	26	28	30	27
sRNA	4	0	1	1
ncRNA	2	3	3	3
snRNA	1	4	4	4
ribozyme	1	0	0	0
SRP_RNA	1	1	1	1

Na figura S2, temos os diagramas de Venn disponibilizado pela nossa ferramenta, que apresenta a quantidade de ncRNAs em cada abordagem, ou seja, a quantidade de anotações que vieram apenas do BLAST, apenas do INFERNAL ou dos dois. Em seguida, temos os gráficos de barras, que também é gerado automaticamente pela nossa ferramenta, fornecendo ao usuário uma rápida visualização da quantidade de cada tipo de ncRNAs anotados. Na anotação do genoma é possível observar uma grande quantidade de misc-RNAs, que é um tipo ncRNA que não se

enquadra nos outros tipos de RNAs (CUNNINGHAM et al, 2022) (mRNA, tRNA, precursor_RNA, etc).

Em seguida foi realizada a anotação dos transcritos obtidos, no caso, a quantidade de ncRNAs encontrados nas três amostras é bem próxima, sendo a maior parte snoRNAs e tRNAs. Na tabela 4 estão presentes os resultados comparando as amostras com a anotação do genoma. É possível observar que a quantidade de ncRNAs tiveram alteração nas três amostras, que pode estar ligada aos fatores que os transcritos foram expostos, que seria o nocaute para catepsina B (LdKO) e o nocaute para catepsina B com plasmídeo expressando catepsina B (LdWT).

II. Identificação de ncRNAs em *Theobroma*

No estudo de caso II, foi realizado a anotação de quatro genomas recém sequenciados de *Theobroma grandiflorum* de forma a identificar os ncRNAs presentes nestes genomas.

Na figura S3 podemos observar os diagramas de Venn construído por nossa ferramenta com a quantidade de ncRNA identificado pelo BLAST e INFERNAL. Dessa forma podemos observar que a maioria dos ncRNAs anotados foram obtidos através da abordagem estrutural. Além disso, podemos observar os gráficos de barras com as quantidades de cada tipo de ncRNA anotado, em que os clones C174 e C174P apresentaram uma grande quantidade de snoRNAs, enquanto os clones C1074 e C1074P tiveram os rRNAs como principal ncRNA identificado.

Além disso, na figura S4, podemos observar as distribuições dos ncRNAs em cada cromossomo, de cada amostra, criado automaticamente pela nossa ferramenta. É interessante que as quatro amostras apresentaram uma grande quantidade de ncRNAs no cromossomo 2, principalmente de rRNAs.

6 Discussão

A ferramenta foi validada em dois arquivos com sequências de RNAs do genoma e em quatro genomas retirados de bancos de dados públicos, onde apresentando aproximadamente 89% de sensibilidade nos transcritos de *D. melanogaster*, enquanto o StructRNAfinder e FindNonCoding apresentaram 18% e 2%, respectivamente, de um total de 4579 ncRNAs. Além disso, em *S. cerevisiae*, nossa ferramenta identificou todos os 419 ncRNAs presentes, ou seja, 100% de sensibilidade. Dessa forma, a ferramenta apresentou excelentes resultados na validação da anotação de transcritos, apresentando as maiores sensibilidades, precisões, acurácias e F1-score.

Já na validação dos genomas, todas as ferramentas apresentaram 100% de sensibilidade no genoma de *E. coli*, variando apenas na quantidade de falsos positivos. Já para as plantas, os resultados pioraram em todas as ferramentas, mas nossa ferramenta continuou apresentando o melhor resultado. No genoma de *A. thaliana* nós acertamos 68% dos ncRNAs, enquanto o FindNonCoding acertou 13% e o StructRNAfinder acertou apenas 5 dos 5178 ncRNAs, representando 0,1% dos ncRNAs. A quantidade de acertos do StructRNAfinder se repetiu para o genoma de *O. sativa*, que conta com 1011 ncRNAs, representando então 0,5% dos ncRNAs, enquanto o FindNonCoding não conseguiu executar e nossa ferramenta identificou 58% dos ncRNAs. Dessa forma, é possível inferir que a busca estrutural pode não ser tão interessante para a anotação de plantas, uma vez que não apresentou uma quantidade significativa de ncRNAs anotados.

Através dos estudos de caso conseguimos testar nossa ferramenta em projetos mais práticos. O primeiro envolvendo re-anotação do genoma de *L. donovani* e comparando os resultados com amostras dos transcritos de vesículas extracelulares, e o segundo na anotação dos genomas de *Theobroma grandiflorum.* Os dois estudos de casos reforçam a necessidade da utilização das duas abordagens computacionais para anotar ncRNAs (busca estrutural e busca por similaridade). Pois, observado os diagramas da figura S2, podemos constatar que a maioria dos resultados do estudo de caso I foram obtidos através da ferramenta BLAST (busca por similaridade), enquanto, observando os diagramas da figura S3, podemos verificar que a maioria dos resultados do estudo de caso II foram obtidos através da ferramenta INFERNAL (busca estrutural).

7 Conclusão

Este artigo apresentou o desenvolvimento de uma ferramenta que automatiza a anotação de RNAs não-codificantes, através de duas abordagens computacionais: busca por similaridade e busca estrutural. Além disso, ela realiza uma pós-análise de forma a identificar as sobreposições e unilas, para assim apresentar o melhor resultado. Utilizando um banco de dados curado com modelos de covariância e outro banco com sequências de 49 outros bancos de ncRNAs, a ferramenta apresentou os melhores resultas de sensibilidade quando comparado com as ferramentas que apresentam objetivos similares ao nosso (StructRNAfinder e FindNonCoding).

Além da anotação apresentada pela ferramenta, são geradas tabelas com as quantidades de ncRNAs anotados, junto de figuras para apresentar uma rápida visualização dos resultados. Dessa forma, o usuário possui todos os resultados de uma forma simples e objetiva, e sem necessitar de aprendizado em uma linguagem de programação.

De forma a facilitar ainda mais a anotação de ncRNAs, está sendo desenvolvida uma versão Web da ferramenta, para o usuário conseguir realizar análises de forma online, sem precisar realizar o download da nossa ferramenta. Além disso, nossa ferramenta enviada para o Galaxy, de forma a ser outra opção online para o usuário realizar seus experimentos.

Funding

O trabalho contou com o apoio da NAPI Bioinformática via Fundação Araucária por meio do convênio 66/2021.

Referências

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

Aparicio-Puerta, E., et al. sRNAbench and sRNAtoolbox 2022 update: accurate miRNA and sncRNA profiling for model and non-model organisms, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W710–W717, https://doi.org/10.1093/nar/gkac363

Arias-Carrasco, R., Vásquez-Morán, Y., Nakaya, H. I., & Maracaja-Coutinho, V. (2018). StructRNAfinder: an automated pipeline and web server for RNA families prediction. BMC Bioinformatics, 19(1), 55.

Chan, P.P. and Lowe, T. M. (2019) tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. Methods Mol Biol. 1962:1-14.

Cunningham, F. and others, Ensembl 2022, Nucleic Acids Research, Volume 50, Issue D1, 7 January 2022, Pages D988–D995, https://doi.org/10.1093/nar/gkab1049

Esteller, M. (2011). Non-coding RNAs in human disease. Nature Reviews Genetics, 12(12), 861-874. doi: 10.1038/nrg3074

Frankish A, et al. GENCODE 2021. Nucleic Acids Res 2021 : 49 ; d1 ; D916-D923. PUBMED: 33270111; PMC: PMC7778937; DOI: 10.1093/nar/gkaa1087

Gerbaba T.K., Gedamu L.; (2013) Cathepsin B Gene Disruption Induced Leishmania donovani Proteome Remodeling Implies Cathepsin B Role in Secretome Regulation. PLoS ONE 8(11): e79951. https://doi.org/10.1371/journal.pone.0079951

Kalvari, I. et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Research, v. 49, n. D1, p. D192–D200, 11 2020. ISSN 0305-1048. Disponível em:<https://doi.org/10.1093/nar/gkaa1047>.

Lekka, E.; Hall, J. (2018), Noncoding RNAs in disease. FEBS Lett, 592: 2884-2900. https://doi.org/10.1002/1873-3468.13182

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. Algorithms for Molecular Biology, 6(1), 26. doi: 10.1186/1748-7188-6-26

Mattick, John S, and Igor V Makunin. "Non-coding RNA." Human molecular genetics vol. 15 Spec No 1 (2006): R17-29. doi:10.1093/hmg/ddl046

Nawrocki, E. P.; Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, v. 29, n. 22, p. 2933–2935, 09 2013. ISSN 1367-4803. Disponível em:<https://doi.org/10.1093/bioinformatics/btt509>.

Newell, P.D., Fricker, A.D., Roco, C.A., Chandrangsu, P., Merkel, S.M. A Small-Group Activity Introducing the Use and Interpretation of BLAST. J Microbiol Biol Educ. 2013 Dec 2;14(2):238-43. doi: 10.1128/jmbe.v14i2.637. PMID: 24358388; PMCID: PMC3867762.

PAULA, Tainah de. Técnicas de Amostragem. CAPCS / UERJ. 12 de Ago de 2019. Disponível em: http://www.capcs.uerj.br/tecnicas-de-amostragem/. Acesso em: 16 de Dez de 2022.

O'Leary, Nuala A et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research vol. 44,D1 (2016): D733-45. doi:10.1093/nar/gkv1189.

RNACENTRAL CONSORTIUM. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res. 2021 Jan 8;49(D1):D212-D220. doi: 10.1093/nar/gkaa921. PMID: 33106848; PMCID: PMC7779037.

Sean R. Eddy, Richard Durbin, RNA sequence analysis using covariance models, Nucleic Acids Research, Volume 22, Issue 11, 11 June 1994, Pages 2079–2088, https://doi.org/10.1093/nar/22.11.2079

Wright, E.S. FindNonCoding: rapid and simple detection of non-coding RNAs in genomes, Bioinformatics, Volume 38, Issue 3, February 2022, Pages 841–843, https://doi.org/10.1093/bioinformatics/btab708

MATERIAIS SUPLEMENTARES

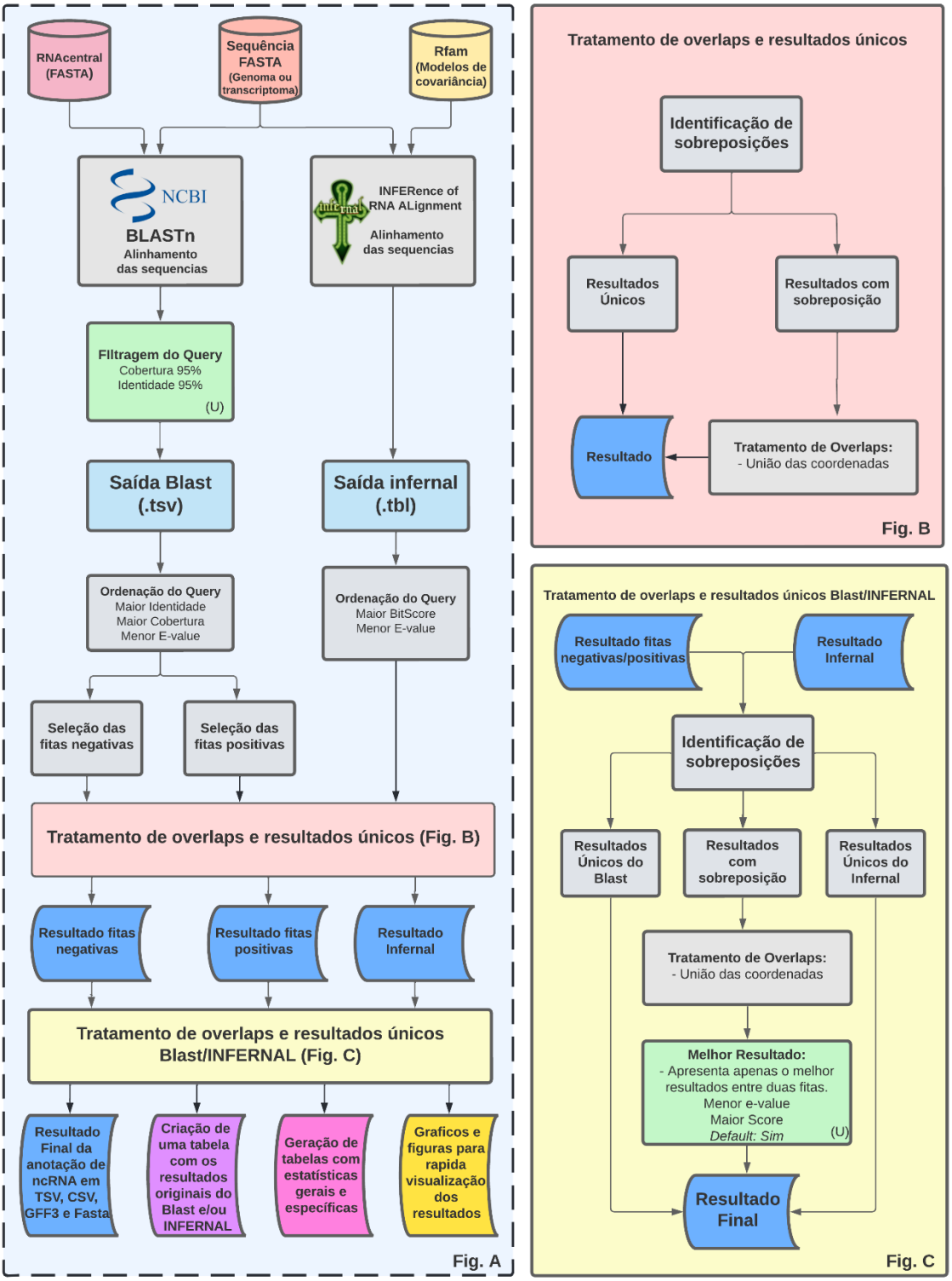


Figura S1 - Fluxogramas da ferramenta. Em A é apresentado o fluxograma da ferramenta, mostrando os bancos de dados e programas utilizados. Em B temos o fluxograma para primeira etapa de Tratamento de sobreposições e resultados únicos, onde é analisado individualmente cada resultado (Fita positiva e fita negativa que vieram do BLAST e INFERNAL). Em C temos a segunda etapa de Tratamento de overlaps e resultados únicos, que analisa em conjunto os resultados do BLAST e INFERNAL, e seleciona os melhores resultados. As caixas verdes com o simbolo (U) representam etapas que o usuário pode adaptar.

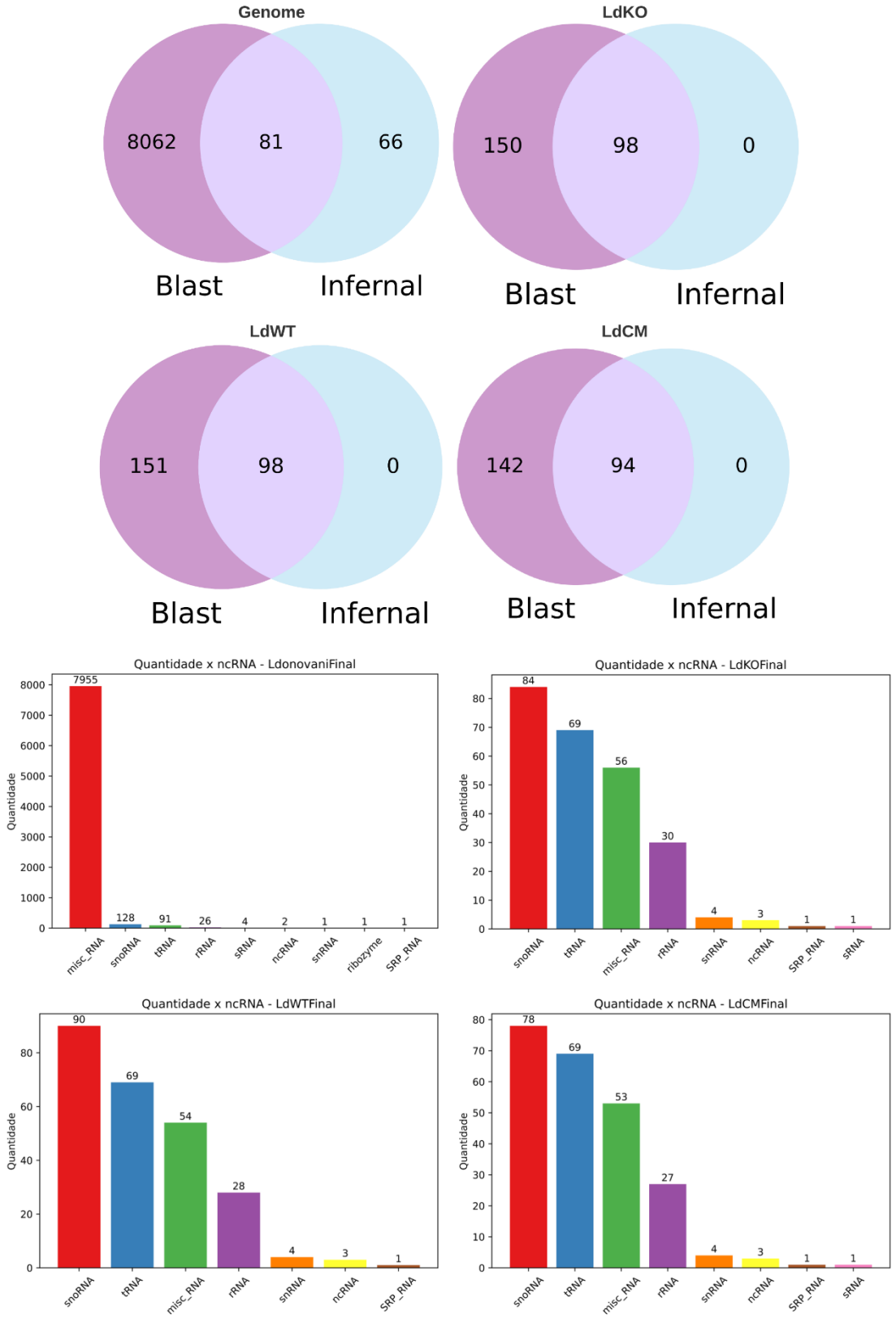


Figura S2 – Diagrama de Veen da quantidade de ncRNAs identificados por cada ferramenta para o genoma de *L. donovani* e três amostras do estudo de caso I. No caso LdCM representa tipo selvagem, LdKO tipo nocaute para catepsina B, e LdWT tipo nocaute para catepsina B com plasmídeo expressando catepsina B. Em seguida é apresentado quatro gráfico de barras representando a quantidade de ncRNAs anotados no genoma e nas três amostras.

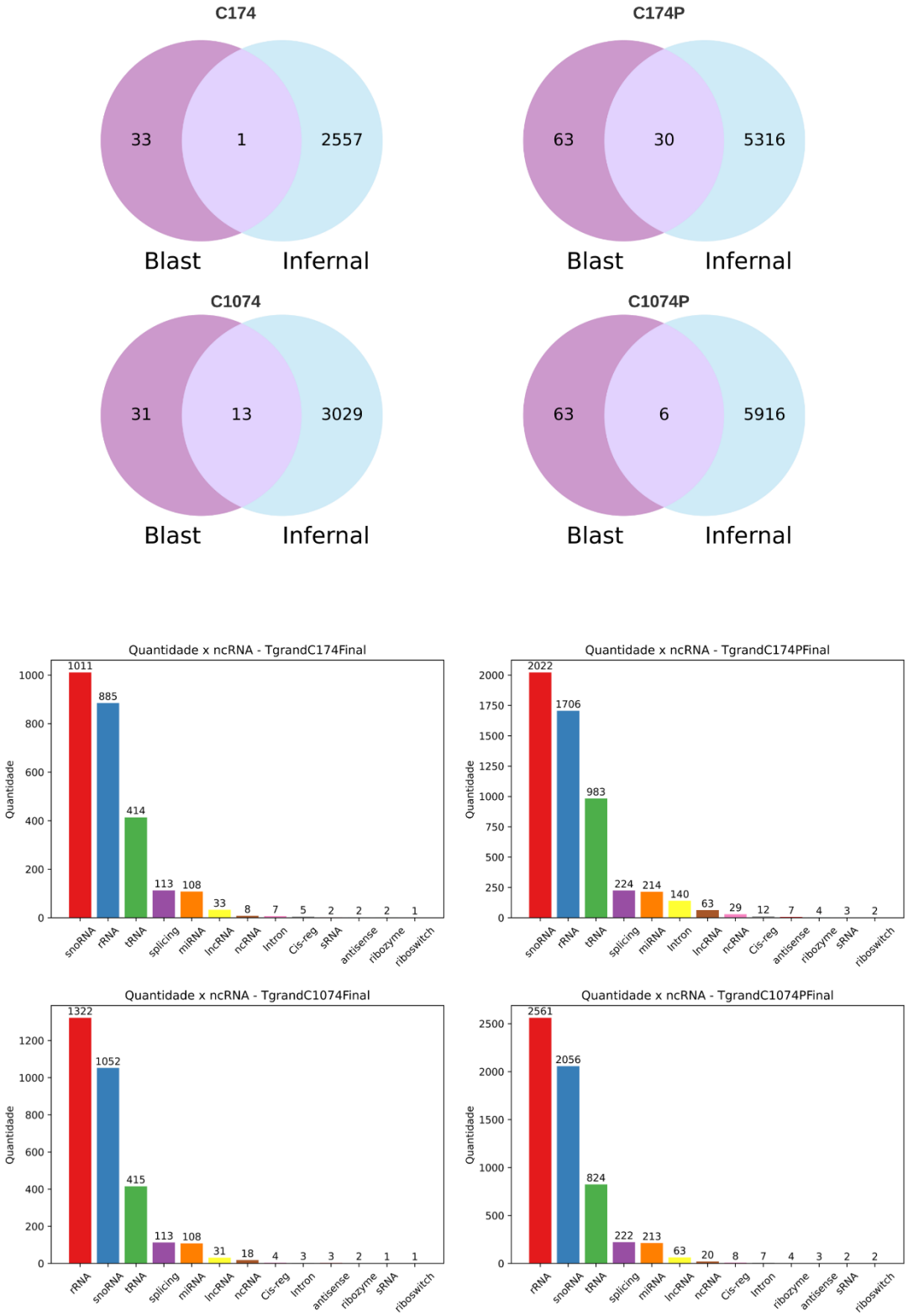


Figura S3 – Diagrama de Veen da quantidade de ncRNAs identificados por cada ferramenta para as quatro amostras do estudo de caso II. Em seguida é apresentado quatro gráfico de barras representando a quantidade de ncRNAs anotados nas quatro amostras. Neste caso, as quatro amostras são genomas recém sequenciados de *Theobroma grandiflorum*.

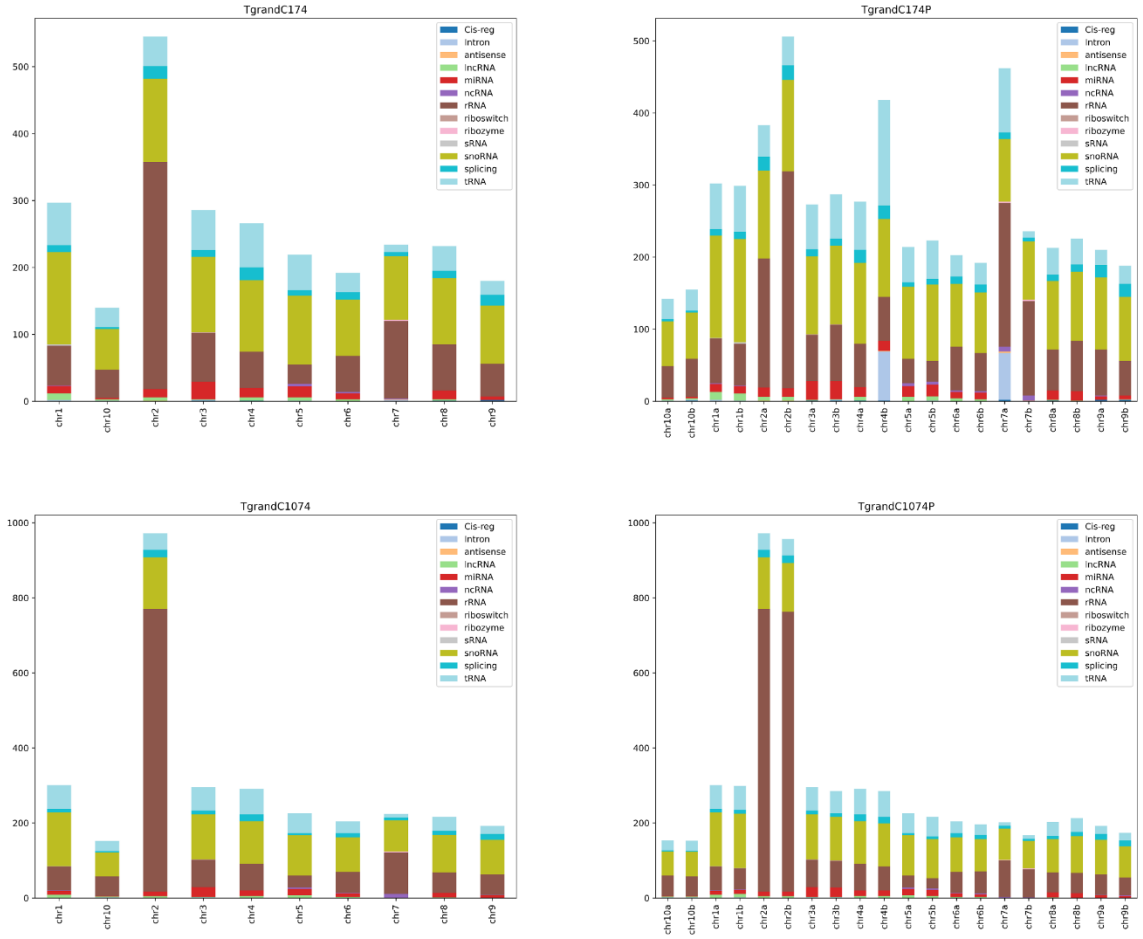


Figura S4 – Gráficos de barras empilhadas contendo a quantidade de ncRNAs anotados em cada cromossomo de cada amostra do estudo de caso II.

Tabela S1 - Matriz de confusão e métricas de proporção para o resultado da análise dos transcritos de *Drosophila melanogaster* nas três ferramentas

	ncRNAs	mRNAs	Total
D. melanogaster 1	4579	4579	9158

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	4089	91	4488	490	89,30%	98,01%	93,66%	97,82%	93,37%
StructRNAfinder	840	297	4282	3739	18,34%	93,51%	55,93%	73,88%	29,39%
FindNonCoding	125	8	4571	4454	2,73%	99,83%	51,28%	93,98%	5,31%

D. melanogaster 2

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	4089	85	4494	490	89,30%	98,14%	93,72%	97,96%	93,43%
StructRNAfinder	840	296	4283	3739	18,34%	93,54%	55,94%	73,94%	29,40%
FindNonCoding	125	7	4572	4454	2,73%	99,85%	51,29%	94,70%	5,31%

D. melanogaster 3

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	4089	89	4490	490	89,30%	98,06%	93,68%	97,87%	93,39%
StructRNAfinder	840	292	4287	3739	18,34%	93,62%	55,98%	74,20%	29,42%
FindNonCoding	125	4	4575	4454	2,73%	99,91%	51,32%	96,90%	5,31%

D. melanogaster 4

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	4089	81	4498	490	89,30%	98,23%	93,77%	98,06%	93,47%
StructRNAfinder	840	313	4266	3739	18,34%	93,16%	55,75%	72,85%	29,31%
FindNonCoding	125	5	4574	4454	2,73%	99,89%	51,31%	96,15%	5,31%

D. melanogaster 5

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	4089	87	4492	490	89,30%	98,10%	93,70%	97,92%	93,41%
StructRNAfinder	840	298	4281	3739	18,34%	93,49%	55,92%	73,81%	29,39%
FindNonCoding	125	8	4571	4454	2,73%	99,83%	51,28%	93,98%	5,31%

Tabela S2 - Matriz de confusão e métricas de proporção para o resultado da análise dos transcritos de *Saccharomyces cerevisiae* nas três ferramentas

	ncRNAs	mRNAs	Total
S. cerevisiae 1	419	419	838

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	419	3	416	0	100,00%	99,28%	99,64%	99,29%	99,64%
StructRNAfinder	401	9	410	18	95,70%	97,85%	96,78%	97,80%	96,74%
FindNonCoding	15	0	419	404	3,58%	100,00%	51,79%	100,00%	6,91%

S. cerevisiae 2

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	419	6	413	0	100,00%	98,57%	99,28%	98,59%	99,29%
StructRNAfinder	401	10	409	18	95,70%	97,61%	96,66%	97,57%	96,63%
FindNonCoding	15	0	419	404	3,58%	100,00%	51,79%	100,00%	6,91%

S. cerevisiae 3

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	419	3	416	0	100,00%	99,28%	99,64%	99,29%	99,64%
StructRNAfinder	401	5	414	18	95,70%	98,81%	97,26%	98,77%	97,21%
FindNonCoding	15	0	419	404	3,58%	100,00%	51,79%	100,00%	6,91%

S. cerevisiae 4

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	419	6	413	0	100,00%	98,57%	99,28%	98,59%	99,29%
StructRNAfinder	401	6	413	18	95,70%	98,57%	97,14%	98,53%	97,09%
FindNonCoding	15	0	419	404	3,58%	100,00%	51,79%	100,00%	6,91%

S. cerevisiae 5

	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	Precision	F1 score
VitorTool	419	7	412	0	100,00%	98,33%	99,16%	98,36%	99,17%
StructRNAfinder	401	5	414	18	95,70%	98,81%	97,26%	98,77%	97,21%
FindNonCoding	15	1	418	404	3,58%	99,76%	51,67%	93,75%	6,90%

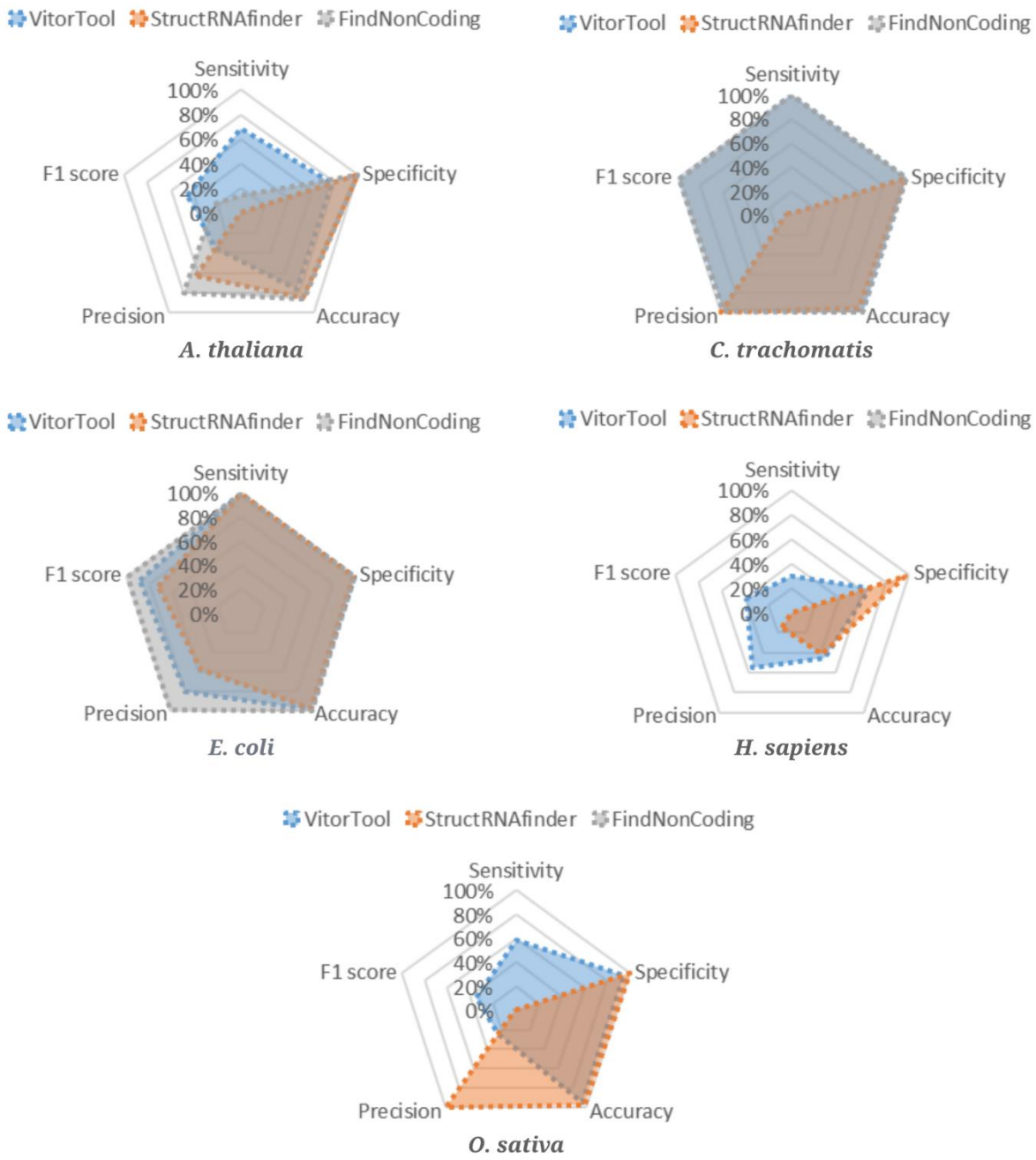


Figura S5 – Gráficos de radar com as métricas de proporção para os resultados das análises dos genomas.

4 RELATÓRIOS DA FERRAMENTA

O sistema desenvolvido apresenta relatórios, gráficos e tabelas com os resultados das anotações, de forma a facilitar para o usuário a utilização dos resultados em futuras análises. Neste capítulo, serão apresentados todos os relatórios, tabelas e figuras geradas.

4.1 Anotações

As anotações são disponibilizadas em dois formatos: *Comma-separated values* (CSV) e *Gene transfer format* (GFF). Na Figura 7 é uma captura de tela de parte de uma anotação no formato GFF. Na primeira coluna contém o nome da sequência do arquivo FASTA, seguido da ferramenta que identificou esse ncRNA (BLAST, INFERNAL ou as duas). Na terceira coluna temos o tipo do ncRNA encontrado, seguido da sua coordenada de início e fim, e na sexta e sétima coluna temos o score determinado por cada ferramenta e o sentido da fita (“+” para *sense* e “-“ para *anti-sense*) respectivamente. Por fim, na última coluna temos um identificador que criamos para cada resultado, que poderá ser utilizado para encontrar o resultado inicial das ferramentas.

Figura 7 - Anotação no formato GFF

```
chr6    Blast   lncRNA  41751421        41753348        3044.0  -       .       id2436
chr6    Infernal        snoRNA  41797538        41797630        37.1    +       .       id2437
chr6    Infernal        snoRNA  42181794        42181892        94.8    -       .       id2438
chr6    Infernal        rRNA    42237710        42237826        54.6    +       .       id2439
chr6    Infernal        rRNA    42287744        42287860        52.2    +       .       id2440
chr6    Infernal        rRNA    42297293        42297409        63.4    -       .       id2441
chr7    Infernal        rRNA    1       615     569.5   +       .       id2442
chr7    Infernal        rRNA    4890    6697    1793.5  +       .       id2443
chr7    Both    ncRNA   7325    10718   -       +       .       id2444
chr7    Infernal        rRNA    14987   16793   1793.6  +       .       id2445
chr7    Both    ncRNA   17423   20813   -       +       .       id2446
chr7    Infernal        rRNA    25088   26895   1793.5  +       .       id2447
```

Na Figura 8 tem uma captura de tela da anotação no formato CSV, que apresenta as mesmas informações que no GFF, mas com uma ordenação diferente. A primeira coluna continua representando o nome da sequência, seguida da coordenada de início e fim nas colunas 2 e 3 respectivamente. No caso dos resultados do BLAST, a coluna 4 e 5 apresentam a cobertura e identidade do alinhamento, enquanto no resultado do INFERNAL, pode apresentar apenas o Clan da família do

ncRNA. Já na coluna 6 teremos o ID do banco de dados para o ncRNA encontrado, e na coluna 7 o sentido da fita (“+” para *sense* e “-“ para *anti-sense*). Por fim, nas seguintes colunas teremos o score, e-value, tipo do ncRNA específico, tipo do ncRNA geral, ferramenta que identificou e o identificador do resultado, respectivamente.

Figura 8 - Anotação no formato CSV

```
chr6,41751421,41753348,95.243,100.156,URS00023A7000,-,3044.0,0.0,ncRNA,lncRNA,Blast,id2436
chr6,41797538,41797630,-,-,RF01284,+,37.1,5.0,snoR8a,snoRNA,Infernal,id2437
chr6,42181794,42181892,-,-,RF00149,-,94.8,1e-15,snoZ103,snoRNA,Infernal,id2438
chr6,42237710,42237826,CL00113,-,RF00001,+,54.6,1.6e-05,5S_rRNA,rRNA,Infernal,id2439
chr6,42287744,42287860,CL00113,-,RF00001,+,52.2,6.5e-05,5S_rRNA,rRNA,Infernal,id2440
chr6,42297293,42297409,CL00113,-,RF00001,-,63.4,8.1e-08,5S_rRNA,rRNA,Infernal,id2441
chr7,1,615,CL00112,-,RF02543,+,569.5,2.3e-149,LSU_rRNA_eukarya,rRNA,Infernal,id2442
chr7,4890,6697,CL00111,-,RF01960,+,1793.5,0.0,SSU_rRNA_eukarya,rRNA,Infernal,id2443
chr7,7325,10718,97.825,96.434,URS000214FAE6,+,5553.0,0.0,ncRNA,ncRNA,Both,id2444
chr7,14987,16793,CL00111,-,RF01960,+,1793.6,0.0,SSU_rRNA_eukarya,rRNA,Infernal,id2445
chr7,17423,20813,97.886,100.0,URS00000A350B,+,5808.0,0.0,ncRNA,ncRNA,Both,id2446
chr7,25088,26895,CL00111,-,RF01960,+,1793.5,0.0,SSU_rRNA_eukarya,rRNA,Infernal,id2447
```

4.2 Fragmentos das sequências

Outra contribuição é a criação de um arquivo FASTA com os fragmentos dos ncRNAs identificados. Na Figura 9, temos um exemplo de dois fragmentos, que apresentam o ncRNA da sequência e o identificador daquele resultado.

Figura 9 - Fragmentos de ncRNAs

```
>fragment_1171 lncRNA;ID=id1171
AGAATCAGGCGGTGGCGAATCCCAGTTCGGAATCGGACCCGAAAGTTCAGGCTGTGGATT
CATCAGCGCCACCTGCGCCTGCACCAACACCGGCACCGGCACCAGCATCAGCACCAGTGC
AAGTTTCAGCTTCAGCTCCAGCTCCAGCTCCGACTCCAGCTGCACCAACTCTGCCTCAGT
CAACTAGTGTGATATCCTCTTCAGTTTTGCCAATTCGAAGTTCTG
>fragment_1172 lncRNA;ID=id1172
GGAACTTGAGTAAAAAATCAAAATTTCATTTTCAAAAAAGAAAACAAAAACAATTGTGAA
GTTAATTTTCGTAGGGATTAAATGACTTCTCATTTGGAGTATTTTTCATTAAATTTGCAA
CAATTTTCCAGATTCATCTAGTGACATTTAGAATCTATCAGCAACGTAAACTAAAGAACA
AATCTTTTAGAGATGGGGTGACTGAAGATTTTTTATTCAGATGATTCTGGACACTGCTAG
GGAAGGTAGGCCCAACTTCATTAATAGTGTTTTCCATATGGTTGCTTCTTTTTTGCCCAT
GCTTTTTCCATGAATGCTCTTTTCATGTCGACCAGGACGTAATTAGGGTAGGGCGAGGCG
GCGGGCGGGGCCTTCGCGCCCCAAACCCCAAAGTTCCTAAAATTGTT
```

4.3 Resultados originais de cada ferramenta

De forma a apresentar os resultados originais das ferramentas BLAST e INFERNAL, criamos um arquivo com todos os resultados obtidos. Na Figura 10 é possível observar um exemplo do arquivo criado, sendo possível observar que o id2443, id2445 e id2447 apresentam apenas um resultado do INFERNAL, enquanto os id2444 e id2446 apresentam resultado do BLAST e INFERNAL. Os resultados do

BLAST apresentam mais de um resultado na mesma região, porém, com diferentes valores de cobertura e identidade.

Figura 10 - Arquivo com os resultados originais de cada ferramenta

```
id2443
Infernal:
chr7    4890    6697    CL00111 -       RF01960 +       1793.5  0.0     SSU_rRNA_eukarya

id2444
Blast:
chr7    URS000214FAE6   97.825  96.434  3219    67      3       7325    10542   1       3217    +       0.0     5553.0  40140279        3337
chr7    URS0000EC65BE   96.798  99.794  3404    92      16      7326    10718   3       3400    +       0.0     5668.0  40140279        3400
chr7    URS0000000B37   96.596  99.851  3349    98      15      7354    10693   1       3342    +       0.0     5550.0  40140279        3345
chr7    URS000032B947   96.587  99.851  3369    104     10      7330    10690   1       3366    +       0.0     5574.0  40140279        3366
chr7    URS00019A29C8   96.528  100.118 3399    100     17      7327    10718   3388    1       -       0.0     5607.0  40140279        3388
Infernal:
chr7    7327    10718   CL00112 -       RF02543 +       3265.0  0.0     rRNA    LSU_rRNA_eukarya        Infernal

id2445
Infernal:
chr7    14987   16793   CL00111 -       RF01960 +       1793.6  0.0     SSU_rRNA_eukarya

id2446
Blast:
chr7    URS00000A350B   97.886  100.0   3359    70      1       17430   20788   1       3358    +       0.0     5808.0  40140279        3359
chr7    URS00003ECC43   96.845  99.91   3328    98      6       17446   20768   1       3326    +       0.0     5570.0  40140279        3326
chr7    URS0000A2DE2B   96.815  100.089 3360    104     2       17426   20785   1       3357    +       0.0     5609.0  40140279        3357
Infernal:
chr7    17423   20813   CL00112 -       RF02543 +       3264.5  0.0     rRNA    LSU_rRNA_eukarya        Infernal

id2447
Infernal:
chr7    25088   26895   CL00111 -       RF01960 +       1793.5  0.0     SSU_rRNA_eukarya
```

4.4 Resultados de miRNA e tRNA

Outra contribuição da ferramenta, é a criação de anotações exclusiva para miRNAs ou tRNAs, ou seja, é criado um arquivo CSV e GFF apenas com miRNAs ou tRNAs encontrados na anotação original. Além disso, é gerado um arquivo FASTA com apenas os fragmentos de miRNAs ou tRNAs. Por fim é gerada uma tabela com a quantidade de cada miRNA ou tRNA específico que foi anotado, como pode ser observado na Figura 11, que representa a quantidade de miRNAs encontrados, seguido do ID no banco de dados.

Figura 11 - Tabela de miRNAs

```
,ncRNA,Qtd,IDdb
0,MIR169_2,12,RF00645
1,mir-166,10,RF00075
2,MIR159,9,RF00638
3,MIR171_1,9,RF00643
4,MIR169_5,7,RF00865
5,mir-399,7,RF00445
6,mir-172,5,RF00452
7,mir-156,4,RF00073
8,MIR396,4,RF00648
9,mir-395,3,RF00451
10,MIR167_1,3,RF00640
11,mir-160,3,RF00247
12,MIR164,3,RF00647
13,MIR397,3,RF00704
14,MIR2275,2,RF03896
15,MIR473,2,RF00778
16,MIR403,2,RF00842
17,MIR394,2,RF00688
18,MIR390,2,RF00689
19,mir-393,2,RF02516
20,MIR398,1,RF00695
21,MIR828,1,RF01026
```

4.5 Tabelas

De forma a apresentar uma rápida quantificação dos resultados para o usuário, a ferramenta cria uma tabela com a quantidade de ncRNAs anotados, como pode ser observada na Figura 12.

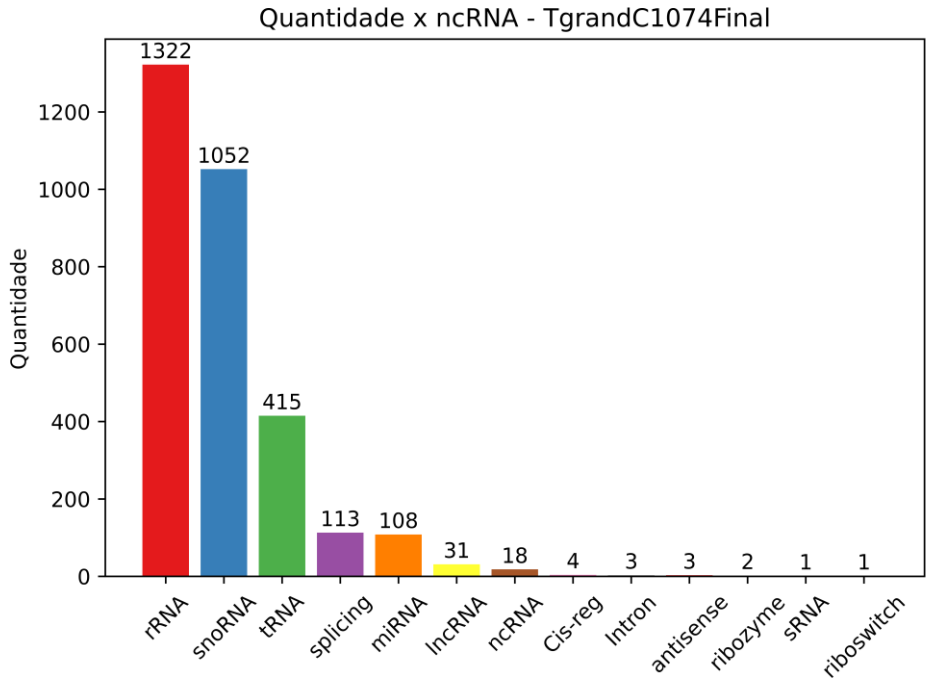
Figura 12 - Tabela com a quantidade de ncRNAs

```
,ncRNA,Qtd
0,rRNA,1322
1,snoRNA,1052
2,tRNA,415
3,splicing,113
4,miRNA,108
5,lncRNA,31
6,ncRNA,18
7,Cis-reg,4
8,Intron,3
9,antisense,3
10,ribozyme,2
11,sRNA,1
12,riboswitch,1
```

4.6 Gráficos

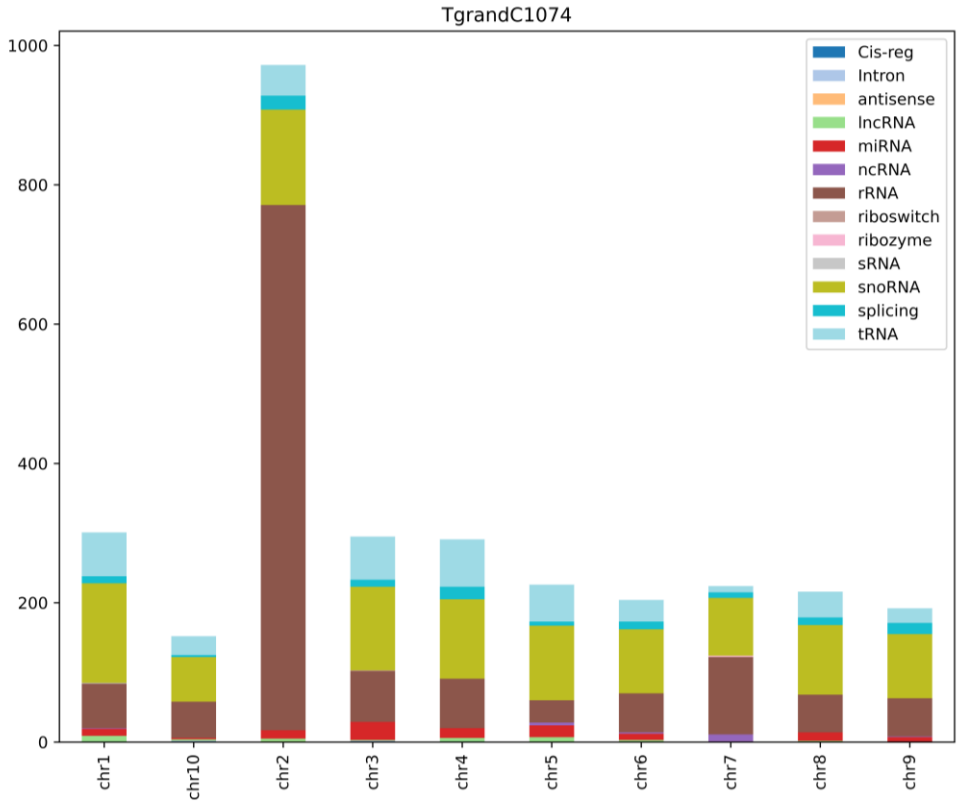
Utilizando a mesma tabela gerada na Figura 12, a ferramenta utiliza os mesmos dados para criar um gráfico de barras, como pode ser observado na Figura 13.

Figura 13 - Gráfico de barras com a quantidade de ncRNAs



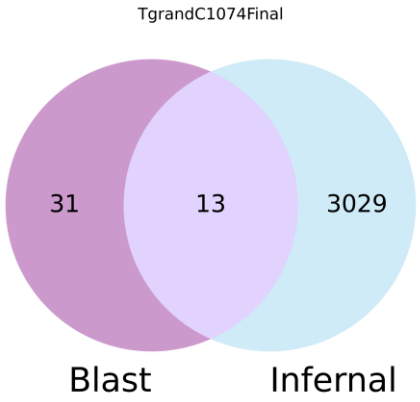
Além deste gráfico, a ferramenta cria um gráfico de barras empilhadas, separando a quantidade de ncRNAs de acordo com cada sequência do arquivo FASTA. Esse tipo de gráfico se apresenta mais interessante quando estamos analisando um genoma com as sequências separadas em cromossomos, como na Figura 14.

Figura 14 - Gráfico de barras empilhadas com a quantidade de ncRNAs em cada cromossomo



Por fim, um último gráfico que a ferramenta desenvolve é um diagrama de Venn, como o da Figura 15, com a quantidade de anotações realizadas pelo BLAST, INFERNAL ou pelas duas. Dessa forma é possível compreender qual método de busca foi mais utilizado nessa anotação.

Figura 15 - Diagrama de Venn com a quantidade de anotações de cada ferramenta

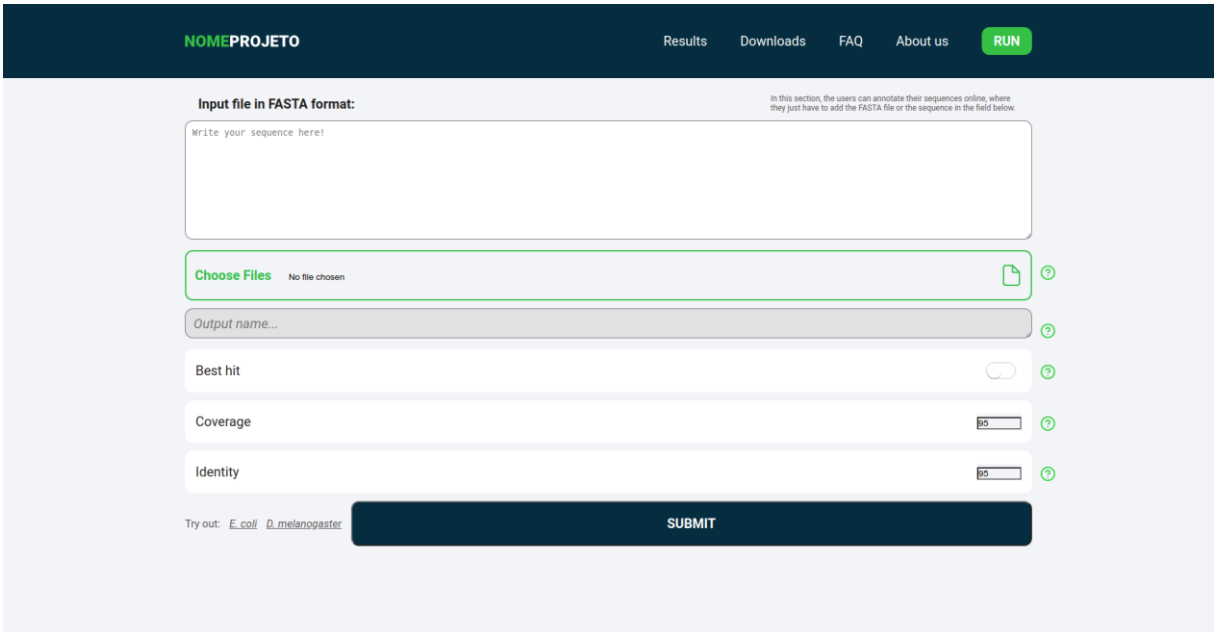


5 INTERFACE GRÁFICA

De forma a facilitar a experiência do usuário, está sendo desenvolvida uma interface gráfica baseada em tecnologias web para a nossa ferramenta.

Na Figura 16 é possível observar uma captura de tela do protótipo da interface gráfica da ferramenta. Ela foi desenvolvida pelo designer Pedro Henrique Pelisson, onde foi levado em consideração a estrutura, cor e acessibilidade de forma a apresentar uma melhor experiência ao usuário.

Figura 16 - Página Web inicial da ferramenta



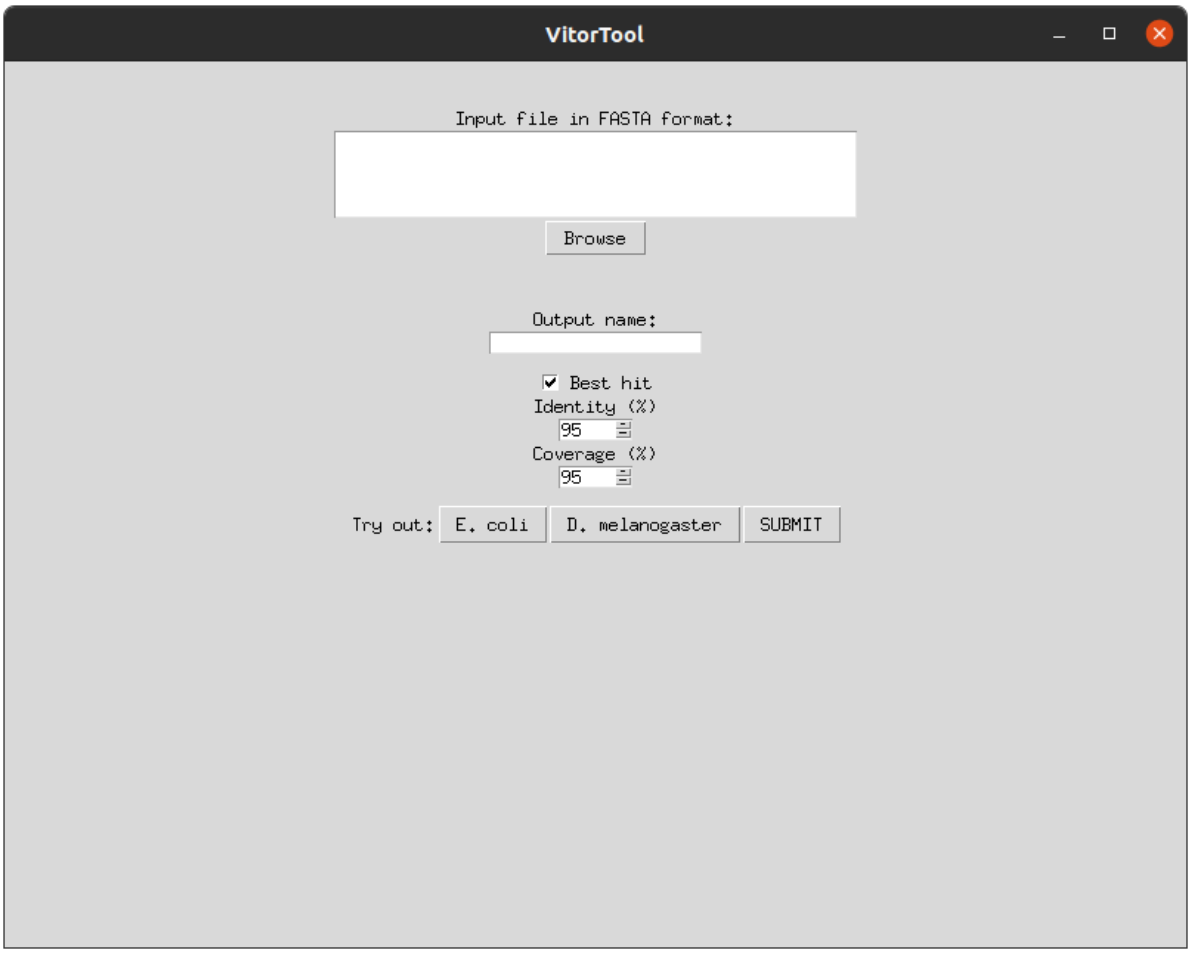
Fonte: Pedro Henrique Pelisson

O desenvolvimento da interface web é orientada pela premissa de simplicidade, com o intuito de minimizar a necessidade de aprendizado e familiaridade com as funcionalidades por parte do usuário. No caso, o usuário pode colocar apenas a sequência ou um arquivo com sequências FASTA, que a ferramenta já poderá ser utilizada. Adicionalmente, existem algumas opções customizáveis, como o nome do arquivo de saída, o valor da porcentagem de cobertura e identidade, e a opção melhor resultado (*best hit*).

Para as próximas versões, espera-se desenvolver a página de resultados, onde o usuário poderia acessar todos os resultados de suas análises através de um identificador fornecido pela nossa ferramenta quando executada.

Além disso, desenvolvemos uma interface *desktop* para ferramenta, para os usuários que desejarem utilizar a ferramenta com uma interface de forma local. Na Figura 17 podemos observar uma captura de tela da primeira interface *desktop* da ferramenta.

Figura 17 - Interface *desktop* da ferramenta



Fonte: Autoria própria

No caso a versão *desktop* foi inspirada na versão *Web*, porém de uma forma mais simples, para que funcione apenas utilizando a linguagem Python. Essa versão da ferramenta já está funcionando, possibilitando ao usuário utilizá-la, ou então, utilizar a versão *stand-alone* que não possui interface.

6 DIVERGÊNCIAS EM ANOTAÇÕES

Durante a etapa de validação da ferramenta, em que anotamos vários genomas diferentes, observamos que nossa ferramenta apresentou muitos resultados ainda não anotados para determinadas espécies.

A primeira situação foi no genoma de *Arabidopsis thaliana* disponível no Ensembl Plants na sua versão TAIR10. Neste caso, o Ensembl Plants apresenta uma anotação que conta com 5178 ncRNAs, sendo que 3292 deles foram identificados pela nossa ferramenta. Porém, nossa ferramenta também identificou 15254 novos ncRNAs, ou seja, regiões que não possuem nenhuma anotação de ncRNA ou mRNA no Ensembl Plants.

Nós acessamos o site do RNAcentral para verificação e foi possível identificar que eles vieram do repositório do ENA (European Nucleotide Archive). Nesse repositório, algumas informações como ncRNA e seu tamanho são apresentadas, porém, o próprio site não possui um arquivo com todas as anotações da espécie.

O genoma de *Bos taurus* também apresentou a mesma divergência de anotações. Neste caso, o genoma também foi retirado do Ensembl, e apresentava na anotação original 5954 ncRNAs, enquanto a nossa anotação identificou 43629 ncRNAs, sendo 37171 tRNAs, como pode ser observado na Figura 18.

Figura 18 - Tabela com os ncRNAs anotados de *B. taurus*

ncRNA	VitorTool	Referência
tRNA	37171	0
lncRNA	2563	2199
splicing	1193	0
miRNA	962	951
snoRNA	716	770
rRNA	453	393
Cis-reg	234	0
ncRNA	172	0
frameshift_element	44	0
sRNA	40	3
scaRNA	30	33
IRES	25	0
snRNA	14	1201
ribozyme	11	8
antisense_RNA	1	0
misc-RNA	0	372
mt_tRNA	0	22
pre-miRNA	0	0
Total	43629	5952

No caso da anotação original, não existia nenhum tRNA anotado, o que nos fez optar pela utilização da ferramenta tRNAscan-SE, uma vez que é uma ferramenta focada em identificar tRNAs, que identificou 266225 tRNAs no genoma de *Bos taurus*.

Considerando que ambos os organismos são modelos, estamos falando mesmo de novos ncRNAs? Por que os ncRNAs de *A. thaliana* não estão mais na anotação?

7 IDENTIFICATION OF NOVEL GENES AND PROTEOFORMS IN *ANGIOSTRONGYLUS COSTARICENSIS* THROUGH A PROTEOGENOMIC APPROACH

No início de 2022, nós recebemos a proposta de uma colaboração para anotar o genoma de *Angiostrongylus costaricensis*, que se trata de um parasita que causa uma inflamação intestinal denominada Angiostrongilíase Abdominal (AA) (DA SILVA et al., 2022).

Neste estudo, foi realizada integração da análise RNA-seq e da espectrometria de massa de proteína (MS/MS) do genoma, uma vez que as expressões de genes codificantes e de splicing alternativo podem ser encontradas tanto no nível de proteínas quanto transcricional (DA SILVA et al., 2022).

No capítulo 2.8 *Annotation of Non-Coding RNA Genes,* é descrita a metodologia utilizada para anotar os ncRNAs. Na época, tratava-se da versão inicial da nossa ferramenta, onde era realizada a filtragem, tratamento de sobreposições e união dos resultados. Em seguida, comparamos os resultados obtidos com as anotações existentes do banco de dados WormBase, de forma a identificar novos ncRNAs anotados, onde foram encontrados 426 novos ncRNAs, sendo que 30 deles possuíam o suporte da análise RNA-seq (DA SILVA et al., 2022).

A anotação de ncRNAs não era o foco principal do projeto, mas ajudou a identificar contaminações presentes nos dados. Além disso, através desta colaboração foi possível aperfeiçoar nossa ferramenta, pois o autor nos indicou pontos para serem melhorados, como a automatização da ferramenta e criação de relatórios.

Figura 19 - Primeira página do artigo

pathogens

MDPI

Article

Identification of Novel Genes and Proteoforms in *Angiostrongylus costaricensis* through a Proteogenomic Approach

Esdras Matheus Gomes da Silva [1,2], Karina Mastropasqua Rebello [2,3], Young-Jun Choi [4], Vitor Gregorio [5], Alexandre Rossi Paschoal [5], Makedonka Mitreva [4], James H. McKerrow [6], Ana Gisele da Costa Neves-Ferreira [2,*] and Fabio Passetti [1,*]

1 Instituto Carlos Chagas, Fiocruz, Curitiba 81350-010, PR, Brazil
2 Laboratory of Toxinology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro 21040-900, RJ, Brazil
3 Laboratory of Integrated Studies in Protozoology, Oswaldo Cruz Institute, Fiocruz, Rio de Janeiro 21040-360, RJ, Brazil
4 Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA
5 Bioinformatics and Pattern Recognition Group (Bioinfo-CP), Department of Computer Science (DACOM), Federal University of Technology-Parana (UTFPR), Cornélio Procópio 86300-000, PR, Brazil
6 Center for Discovery and Innovation in Parasitic Diseases, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, CA 92093, USA
* Correspondence: anag@ioc.fiocruz.br (A.G.d.C.N.-F.); fabio.passetti@fiocruz.br (F.P.)

Abstract: RNA sequencing (RNA-Seq) and mass-spectrometry-based proteomics data are often integrated in proteogenomic studies to assist in the prediction of eukaryote genome features, such as genes, splicing, single-nucleotide (SNVs), and single-amino-acid variants (SAAVs). Most genomes of parasite nematodes are draft versions that lack transcript- and protein-level information and whose gene annotations rely only on computational predictions. *Angiostrongylus costaricensis* is a roundworm species that causes an intestinal inflammatory disease, known as abdominal angiostrongyliasis (AA). Currently, there is no drug available that acts directly on this parasite, mostly due to the sparse understanding of its molecular characteristics. The available genome of *A. costaricensis*, specific to the Costa Rica strain, is a draft version that is not supported by transcript- or protein-level evidence. This study used RNA-Seq and MS/MS data to perform an in-depth annotation of the *A. costaricensis* genome. Our prediction improved the reference annotation with (a) novel coding and non-coding genes; (b) pieces of evidence of alternative splicing generating new proteoforms; and (c) a list of SNVs between the Brazilian (Crissiumal) and the Costa Rica strain. To the best of our knowledge, this is the first time that a multi-omics approach has been used to improve the genome annotation of *A. costaricensis*. We hope this improved genome annotation can assist in the future development of drugs, kits, and vaccines to treat, diagnose, and prevent AA caused by either the Brazil strain (Crissiumal) or the Costa Rica strain.

Keywords: RNA-Seq; mass spectrometry; nematode; genome annotation; ncRNAs

Citation: da Silva, E.M.G.; Rebello, K.M.; Choi, Y.-J.; Gregorio, V.; Paschoal, A.R.; Mitreva, M.; McKerrow, J.H.; Neves-Ferreira, A.G.d.C.; Passetti, F. Identification of Novel Genes and Proteoforms in *Angiostrongylus costaricensis* through a Proteogenomic Approach. *Pathogens* **2022**, *11*, 1273. https://doi.org/10.3390/pathogens11111273

Academic Editor: Lawrence S. Young

Received: 21 September 2022
Accepted: 20 October 2022
Published: 31 October 2022

1. Introduction

One of the critical steps after genome sequencing and assembling is annotating the genomic features [1]. Many computational tools perform ab initio genome annotation using only genome sequence motifs [2–7]. However, predicting genes and splice variants within a genome sequence using this approach is particularly challenging for eukaryotic genomes because genes are usually distant from each other and are interrupted by introns [8]. Therefore, data from multiple high-throughput technologies such as RNA sequencing (RNA-Seq) and protein mass spectrometry (MS/MS) are often used to assist in predicting these genomic features [9]. Usually, RNA-Seq reads are aligned to a reference genome, followed by gene prediction and the assembly of transcript sequences [10,11]. Transcript sequences are computationally translated, generating a customized protein-sequence database for protein

8 CONCLUSÃO

Nossa ferramenta foi desenvolvida com a intenção de ser um sistema automático, de fácil utilização e escalável para anotar ncRNAs, de forma a melhorar e facilitar a anotação nas diferentes espécies. Para isso, utilizamos duas abordagens computacionais: (i) busca por similaridade, em que foi utilizado o BLAST com o banco de dados RNAcentral, e (ii) busca estrutural, em que foi utilizado o INFERNAL com o banco de dados Rfam. A automatização e análise dos resultados foi realizada em um *script* em Python, podendo facilitar a instalação e execução da ferramenta.

Para a validação da ferramenta, comparamos os resultados contra outras duas ferramentas do estado da arte (StructRNAfinder e FindNonCoding). Para isso, selecionamos dois genomas e quatro transcriptomas, e re-anotamos nas três ferramentas. Nas sete espécies, a nossa ferramenta apresentou a melhor sensibilidade, conseguindo em três espécies acertar todos os ncRNAs que existiam na anotação original.

Além disso, avaliamos a ferramenta em dois estudos de caso: (i) Identificação de ncRNAs de *L. donovani* em vesículas extracelulares, e (ii) Identificação de ncRNAs em *Theobroma spp*. Este teste foi importante para a validação das duas abordagens computacionais (busca por similaridade e busca estrutural), uma vez que a maioria dos resultados do primeiro estudo de caso vieram do BLAST (busca por similaridade), enquanto no segundo estudo de caso vieram do INFERNAL (busca estrutural), ou seja, ferramentas como StructRNAfinder e FindNonCoding não apresentariam muitos resultados no estudo de caso (i), pois utilizam apenas a busca estrutural. Além disso, nossa ferramenta gerou relatórios, gráficos e tabelas automaticamente, apresentando ao usuário diversas formas de visualização dos resultados.

Espera-se para próximas etapas que seja desenvolvida a interface *desktop* e *Web*, facilitando a utilização da ferramenta, e que seja submetida na plataforma Galaxy. Assim, acredita-se que o usuário consiga utilizar o sistema para realizar uma grande quantidade de anotações e re-anotações, para melhorar a quantidade e qualidade das anotações nos mais diversos genomas.

REFERÊNCIAS

ALTSCHUL, S. F. et al. “Basic local alignment search tool.” Journal of molecular biology vol. 215,3 (1990): 403-10. doi:10.1016/S0022-2836(05)80360-2.

ARIAS-CARRASCO, R., VÁSQUEZ-MORÁN, Y., NAKAYA, H. I., & MARACAJA-COUTINHO, V. (2018). StructRNAfinder: an automated pipeline and web server for RNA families prediction. BMC Bioinformatics, 19(1), 55

AUSTIN, C. P.; Open Reading Frame. 2020. Disponível em: https://www.genome.gov/genetics-glossary/Open-Reading-Frame. Acesso em: 03 fev 2022.

BAJCZYK, M.; JARMOLOWSKI, A.; JOZWIAK, M.; PACAK, A.; PIETRYKOWSKA, H.; SIEROCKA, I.; SWIDA-BARTECZKA, A.; SZEWC, L.; SZWEYKOWSKA-KULINSKA, Z. Recent Insights into Plant miRNA Biogenesis: Multiple Layers of miRNA Level Regulation. Plants 2023, 12, 342. https://doi.org/10.3390/plants12020342

BAO, W.; KOJIMA, K. K.; KOHANY, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, v. 6, n. 11, 2015. doi:10.1186/s13100-015-0041-9 Disponível em: https://www.girinst.org/repbase. Acesso em: 02 fev 2022.

BRENT, M. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet 9, 62–73 (2008). https://doi.org/10.1038/nrg2220

CASIMIRO-SORIGUER, C. S.; RIGUAL, M. M.; BROKATE-LLANOS, A. M.; MUÑOZ, M. J.; GARZÓN, A.; PÉREZ-PULIDO, A. J.; JIMENEZ, J.; Using AnABlast for intergenic sORF prediction in the Caenorhabditis elegans genome, Bioinformatics, Volume 36, Issue 19, 1 October 2020, Pages 4827–4832.

CHAO, J.; TANG, F.; XU, L. Developments in Algorithms for Sequence Alignment: A Review. Biomolecules 2022, 12, 546. https://doi.org/10.3390/biom12040546

CHEN, W.; LIU, D.; LI, Q.; ZHU, H. The function of ncRNAs in rheumatic diseases. Epigenomics 2019 11:7, 821-833. https://doi.org/10.2217/epi-2018-0135

CJ Bioscience, Inc. Pairwise Nucleotide Sequence Alignment For Taxonomy (2017). Disponível em: https://help.ezbiocloud.net/pairwise-nucleotide-sequence-alignment/

CORREIA, J. D.; CORREIA, A. D. Funcionalidades dos rna não codificantes (ncrna) e pequenos rna reguladores, nos mamíferos. REDVET. Revista Electrónica de Veterinaria, Veterinaria Organización, v. 8, n. 10, p. 1–22, 2007.

COUSO, JP., PATRAQUIM, P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol 18, 575–589 (2017). https://doi.org/10.1038/nrm.2017.58

DA SILVA, E.M.G.; REBELLO, K.M.; CHOI, Y.-J.; GREGORIO, V.; PASCHOAL, A.R.; MITREVA, M.; MCKERROW, J.H.; NEVES-FERREIRA, A.G.D.C.; PASSETTI, F. Identification of Novel Genes and Proteoforms in Angiostrongylus costaricensis

through a Proteogenomic Approach. Pathogens 2022, 11, 1273. https://doi.org/10.3390/pathogens11111273

DAILY, J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. BMC Bioinformatics 17, 81 (2016). https://doi.org/10.1186/s12859-016-0930-z

DALE, RK; PEDERSEN, BS; QUINLAN, AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics (2011). doi:10.1093/bioinformatics/btr539

EDDY, S.R.; DURBIN, R. RNA sequence analysis using covariance models. Nucleic Acids Res. 1994 Jun 11;22(11):2079-88. doi: 10.1093/nar/22.11.2079. PMID: 8029015; PMCID: PMC308124.

EJIGU, G.F.; JUNG, J. Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. Biology 2020, 9, 295. https://doi.org/10.3390/biology9090295

ELSIK, C.G., WORLEY, K.C., BENNETT, A.K. et al. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics 15, 86 (2014). https://doi.org/10.1186/1471-2164-15-86

FONSECA, B. H. R. d. Modelagem, integração e análise exploratória de dados públicos de mirtrons. 2018. Disponível em:<http://repositorio.utfpr.edu.br/jspui/handle/1/4507>.

FRANKISH, A. et al; GENCODE 2021. Nucleic Acids Res 2021 : 49 ; d1 ; D916-D923. PUBMED: 33270111; PMC: PMC7778937; DOI: 10.1093/nar/gkaa1087.

FU, L.; e outros. UFold: fast and accurate RNA secondary structure prediction with deep learning, Nucleic Acids Research, Volume 50, Issue 3, 22 February 2022, Page e14, https://doi.org/10.1093/nar/gkab1074

GERBABA, T.K., GEDAMU L.; (2013) Cathepsin B Gene Disruption Induced Leishmania donovani Proteome Remodeling Implies Cathepsin B Role in Secretome Regulation. PLoS ONE 8(11): e79951. https://doi.org/10.1371/journal.pone.0079951

GREGO, M. Conteúdo digital dobra a cada dois anos no mundo. 2014. Disponível em:<https://exame.com/tecnologia/conteudo-digital-dobra-a-cada-dois-anos-no-mundo/>.

GRIFFITHS-JONES, S. The microRNA Registry. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D109-11. doi: 10.1093/nar/gkh023. PMID: 14681370; PMCID: PMC308757.

HAO, Yajing et al. “SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci.” Briefings in bioinformatics vol. 19,4 (2018): 636-643. doi:10.1093/bib/bbx005

HAQUE, W., ARAVIND, A., & REDDY, B. (2009, March). Pairwise sequence alignment algorithms: a survey. In Proceedings of the 2009 conference on Information Science, Technology and Applications (pp. 96-103).

HARIDAS, S.; SALAMOV, A.; GRIGORIEV, I.V. Fungal Genome Annotation. Fungal Genomics: Methods in Molecular Biology 1775, c. 15, p. 171-184, 2018. doi:10.1007/978-1-4939-7804-5_15.

HAAS, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9, R7 (2008).

JERÓNIMO, S. M. d. Q. P. C. Análise visual de dados: uma ferramenta também para startups! 2016. Disponível em:<https://jornaleconomico.sapo.pt/noticias/analise-visual-de-dados-uma-ferramenta-tambem-para-startups-2923>.

KALVARI, I. et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. (2020). Nucleic Acids Research, v. 49, n. D1, p. D192–D200, 11 2020. ISSN 0305-1048. Disponível em:<https://doi.org/10.1093/nar/gkaa1047>.

KIM, D., LANGMEAD, B., Salzberg, S. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12,** 357–360 (2015). DOI: https://doi.org/10.1038/nmeth.3317

KOVAKA, S., ZIMIN, A.V., PERTEA, G.M. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol 20, 278 (2019). https://doi.org/10.1186/s13059-019-1910-1

KRUEGER F. Trim Galore. 2019. Disponível em: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; Acesso em: 08 fev 2022.

LEKKA, E.; HALL, J. (2018), Noncoding RNAs in disease. FEBS Lett, 592: 2884-2900. https://doi.org/10.1002/1873-3468.13182

LENZ, A. R. Anotação Genômica. In: DALL'ALBA, G.; NOTARI, D. L.; SILVA, S. A. Bioinformática: Contexto Computacional e Aplicações. Caxias do Sul, RS: Educs, 2020. p. 259-285. Disponível em: <https://www.ucs.br/educs/arquivo/ebook/bioinformatica-contexto-computacional-e-aplicacoes/>. Acesso em: 4 fev. 2022.

MATTICK, J.S., AMARAL, P.P., CARNINCI, P. et al. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol 24, 430–447 (2023). https://doi.org/10.1038/s41580-022-00566-8

MATTICK, J. S., AND MAKUNIN, I. V. "Non-coding RNA." Human molecular genetics vol. 15 Spec No 1 (2006): R17-29. doi:10.1093/hmg/ddl046

MEDRI, D. W. Análise exploratória de dados. 2011. Disponível em:<http://www.uel.br/pos/estatisticaeducacao/textosdidaticos/especializacaoestatistica.pdf>.

NAWROCKI, E. P., BURGE, S. W., BATEMAN, A., DAUB, J., EBERHARDT, R. Y., EDDY, S. R., FLODEN, E. W., GARDNER, P. P., JONES, T. A., TATE, J., AND FINN, R. D. (2015). Rfam 12.0: updates to the RNA families database. Nucl. Acids Res., 43:D130–D137

NAWROCKI, E. P.; EDDY, S. R. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, v. 29, n. 22, p. 2933–2935, 09 2013. ISSN 1367-4803. Disponível em:<https://doi.org/10.1093/bioinformatics/btt509>.

NCBI. National Center for Biotechnology Information. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 1988. Disponível em: <https://www.ncbi.nlm.nih.gov/>. Acesso em: 8 jul. 2021.

NEEDLEMAN, S.B.; WUNSCH, C.D. “A general method applicable to the search for similarities in the amino acid sequence of two proteins.” Journal of molecular biology vol. 48,3 (1970): 443-53. doi:10.1016/0022-2836(70)90057-4

NEWELL, P.D., FRICKER, A.D., ROCO, C.A., CHANDRANGSU, P., MERKEL, S.M. A Small-Group Activity Introducing the Use and Interpretation of BLAST. J Mi-crobiol Biol Educ. 2013 Dec 2;14(2):238-43. doi: 10.1128/jmbe.v14i2.637. PMID: 24358388; PMCID: PMC3867762

PASCHOAL, A. R.; COUTINHO, V.M.; SETUBAL, J. C.; SIMÕES, Z. L. P.; ALMEIDA, S. A.; DURHAM, A. M.; (2012) Non-coding transcription characterization and annotation, RNA Biology, 9:3, 274-282, DOI: 10.4161/rna.19352

PELLETIER, D.; RIVERA, B.; FABIAN, M.R.; FOULKES, W.D. (2023) miRNA biogenesis and inherited disorders: clinico-molecular insights. Trends in Genetics. VOLUME 39, ISSUE 5, P401-414 https://doi.org/10.1016/j.tig.2023.01.009

PEREIRA, T. C. Introdução ao universo dos non-coding RNAs. Ribeirão Preto: Sociedade Brasileira de Genética, 2017.

PETER, Swathik Clarancia et al. Encyclopedia of bioinformatics and computational biology. Ranganathan, S., Grib-skov, M., Nakai, K., Schönbach, C., Eds, p. 661-676, 2018.

PUJARI, J.J., KANADAM, K.P. (2022). Semi global pairwise sequence alignment using new chromosome structure genetic algorithm. Ingénierie des Systèmes d’Information, Vol. 27, No. 1, pp. 67-74. https://doi.org/10.18280/isi.270108

PUERTA, E. A. et al; sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression, Nucleic Acids Research, Volume 47, Issue W1, 02 July 2019, Pages W530–W535, https://doi.org/10.1093/nar/gkz415

QUINLAN, A. R.; HALL, I. M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.

REBOLLEDO, C.; SILVA, J. P.; SAAVEDRA, N.; COUTINHO, V.M. Computational approaches for circRNAs prediction and in silico characterization, Briefings in Bioinformatics, Volume 24, Issue 3, May 2023, bbad154, https://doi.org/10.1093/bib/bbad154

REDDY, K.B. MicroRNA (miRNA) in cancer. Cancer Cell Int 15, 38 (2015). https://doi.org/10.1186/s12935-015-0185-1

SHEN, C.; ZAHARIAS, P.; WARNOW, T. MAGUS+eHMMs: improved multiple sequence alignment accuracy for fragmentary sequences, Bioinformatics, Volume

38, Issue 4, February 2022, Pages 918–924, https://doi.org/10.1093/bioinformatics/btab788

SMIT, A. F., HUBLEY, R., GREEN, P. RepeatMasker 3.0 repeatmasker.org [online], (1996–2010).

SMITH, T.F; WATERMAN, M.S. “Identification of common molecular subsequences.” Journal of molecular biology vol. 147,1 (1981): 195-7. doi:10.1016/0022-2836(81)90087-5

STEIN, L. Genome annotation: from sequence to biology. Nat Rev Genet 2, 493–503 (2001). https://doi.org/10.1038/35080529

TANTRAY I.; OJHA R.; SHARMA A.P. (2023) Non-coding RNA and autophagy: Finding novel ways to improve the diagnostic management of bladder cancer. Frontiers in Genetics. 10.3389/fgene.2022.1051762

TEO, Y. Y. Exploratory data analysis in large-scale genetic studies.Biostatistics, v. 11,n. 1, p. 70–81, out. 2009. ISSN 1465-4644. Disponível em: <https://doi.org/10.1093/biostatistics/kxp038>. Acesso em: 18 jan. 2022.

WALSH A.T., TRIANT D.A., Le Tourneau J.J., SHAMIMUZZAMAN M., ELSIK C.G. Hymenoptera Genome Database: new genomes and annotation datasets for improved go enrichment and orthologue analyses. Nucleic Acids Res. 2021 Nov 8:gkab1018. doi: 10.1093/nar/gkab1018. Epub ahead of print. PMID: 34747465.

WATERMAN M.S., EGGERT M. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. 1987. J. Mol. Biol. 197(4):723-8. PubMed: 2448477. DOI: 10.1016/0022-2836(87)90478-5

WETHMAR, K.; BARBOSA-SILVA, A.; ANDRADE-NAVARRO, M. A.; LEUTZ, A. uORFdb—a comprehensive literature database on eukaryotic uORF biology, Nucleic Acids Research, Volume 42, Issue D1, 1 January 2014, Pages D60–D67, https://doi.org/10.1093/nar/gkt952.

WRIGHT, E.S. FindNonCoding: rapid and simple detection of non-coding RNAs in genomes, Bioinformatics, Volume 38, Issue 3, February 2022, Pages 841–843, https://doi.org/10.1093/bioinformatics/btab708

WRIGHT, E.S. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. 2016. The R Journal, 8(1), 352-359.

YANDELL, M., ENCE, D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13, 329–342 (2012). https://doi.org/10.1038/nrg3174

YATES, A. D. et al. Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Research 2022. https://doi.org/10.1093/nar/gkab1007

ZHAO Q, ZHAO Z, FAN X, YUAN Z, MAO Q, YAO Y (2021) Review of machine learning methods for RNA secondary structure prediction. PLoS Comput Biol 17(8): e1009291. https://doi.org/10.1371/journal.pcbi.1009291