

Universidade Federal do Rio Grande do Sul  
Centro de Biotecnologia  
Programa de Pós-Graduação em Biologia Celular e Molecular

**Avaliação de pequenos RNAs regulatórios em leveduras patogênicas  
e seu potencial na interação patógeno-hospedeiro**

Tese de Doutorado

Rodrigo Silva Araujo Streit

Porto Alegre, Março de 2022

Universidade Federal do Rio Grande do Sul  
Centro de Biotecnologia  
Programa de Pós-Graduação em Biologia Celular e Molecular

**Avaliação de pequenos RNAs regulatórios em leveduras patogênicas  
e seu potencial na interação patógeno-hospedeiro**

Tese submetida ao Programa de  
Pós-Graduação em Biologia Celular e  
Molecular do Centro de Biotecnologia da  
UFRGS como requisito parcial para a  
obtenção do grau de Doutor em Ciências

**Rodrigo Silva Araujo Streit**

Charley Christian Staats - Orientador

Porto Alegre, Maio de 2022

Este trabalho foi desenvolvido na Unidade de Biologia Teórica e Computacional e no Laboratório de Microbiologia Molecular e Celular, no Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul (CBiot/UFRGS), sob orientação do professor Charley Christian Staats. Apoio financeiro: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS)

## **Banca Examinadora**

---

**Prof. Dr. Fabio Nogueira**

Universidade de São Paulo

---

**Prof. Rodrigo Ligabue Braun**

Universidade Federal de Ciências da Saúde de Porto Alegre

---

**Prof. Diego Bonatto**

Universidade Federal do Rio Grande do Sul

---

**Profa. Dra. Livia Kmetzsch**

Universidade Federal do Rio Grande do Sul

(Suplente)

## AGRADECIMENTOS

Primeiramente, e com maior importância, gostaria de agradecer aos meus pais, Jorge e Maria Lúcia. Vocês sempre lutaram, mesmo em tempos difíceis, para me dar as melhores oportunidades possíveis. Sempre me incentivaram, me permitiram sonhar, me permitiram tentar e, acima de tudo, não me permitiram desistir. Essa última parte é a que faz toda a diferença, pois sem a ajuda de vocês, eu definitivamente não teria chegado nesse ponto. Quando me faltou força, quando eu estava no chão, quando eu não acreditava nem por um segundo em mim mesmo, vocês não apenas acreditaram em mim, mas vocês lutaram para que eu também o fizesse. Vocês nunca desistiram de mim, e eu espero que essa conquista, assim como as que ainda virão, permitam que eu possa retribuir tudo o que vocês fazem por mim, que vocês não mais se preocupem mas sim se alegrem em serem meus pais.

Após meus pais, meu maior agradecimento vai para o meu orientador, Charley. Apesar de nomeá-lo como orientador, creio que esse termo falta em profundidade para exemplificar o papel do Charley para mim. Durante todos os anos que fui orientado pelo Charley, desde a minha graduação, ele me propiciou o melhor ambiente possível para que eu pudesse me desenvolver não apenas como bioinformata, mas como pesquisador e pessoa. Apesar de todos os erros que cometi durante esse aprendizado, o Charley nunca duvidou do meu potencial, nunca restringiu o que eu podia ou não fazer. Pelo contrário, o que recebi do Charley não foi apenas conhecimento, mas a liberdade e o incentivo para me aventurar pelas mais diversas áreas da bioinformática, para poder construir uma base sólida para a minha carreira. Eu posso afirmar com a mais absoluta e total certeza, que eu não poderia encontrar um orientador melhor que o Charley. Na realidade, mais que um orientador, um amigo que pretendo levar para a vida.

Gostaria também de agradecer aos meus amigos da UBTEC e do LMMC. Felizmente, fui agraciado por fazer parte não apenas de um, mas de dois laboratórios repletos de pessoas excepcionais, tanto na sua qualidade de pessoas como de pesquisadores. Vocês tiveram um papel fundamental nesse doutorado, indo desde a discussão e partilha de conhecimento, até as brincadeiras, risadas e, acima de tudo, aturar as minhas reclamações. Vocês tornaram tudo mais leve, mais divertido. Apesar de que agradeço a todos os que ao longo desses anos me deram o prazer de trabalhar comigo, mesmo que de forma breve, creio que devo agradecer particularmente ao Fábio, Ane e Patrícia. Vocês três não apenas são um presente que esse PPG me deu, excelentes amigos que fiz questão de manter mesmo após as respectivas defesas, vocês também são alguns dos mais incríveis pesquisadores que eu já tive o prazer de conviver, e espero que eu ainda tenha a alegria de voltar a trabalhar com vocês de alguma forma.

Por fim, gostaria de agradecer aos demais professores que tiveram participação durante a minha formação, como os professores Augusto Schrank, Marilene Vainstein e Lívia Kmetzsch, assim como aos meus colegas por eles orientados, aos professores que colaboraram diretamente para a execução do presente trabalho como o professor Márcio Dorn e Mendeli Vainstein, e à professora Lysangela Alves do Instituto Carlos Chagas que nos cedeu dados para a execução do presente trabalho.

# SUMÁRIO

<b>SUMÁRIO</b>	<b>6</b>
<b>RESUMO</b>	<b>8</b>
<b>ABSTRACT</b>	<b>9</b>
<b>Lista de Abreviaturas e Símbolos</b>	<b>12</b>
<b>INTRODUÇÃO</b>	<b>13</b>
Fungos Patógenos Humanos	13
Pequenos RNAs Regulatórios	17
Pequenos RNAs e a Interação Patógeno-Hospedeiro	22
Fragmentos de RNA derivados de tRNA	24
<b>OBJETIVOS</b>	<b>30</b>
<b>RESULTADOS</b>	<b>31</b>
Capítulo 1: Identificação de tRFs em <i>Cryptococcus</i> spp	33
Artigo: “Analysis of tRNA-derived RNA fragments (tRFs) in <i>Cryptococcus</i> spp.: RNAi-independent generation and possible compensatory effects in a RNAi-deficient genotype”	34
Discussão geral do artigo	45
Capítulo 2: Avaliação de sRNAs regulatórios em vesículas extracelulares de isolados <i>C. auris</i>	50
Materiais e métodos	51
Dados utilizados	51
Predição de tRFs em vesículas extracelulares	51
Predição de miRNAs em vesículas extracelulares	52
Predição de alvos para sRNAs em hospedeiros mamíferos	53
Resultados	54
Identificação de tRFs em vesículas de <i>C. auris</i>	54
Identificação de miRNAs em vesículas de <i>C. auris</i>	57
Identificação de alvos de sRNAs em hospedeiros mamíferos	58
Discussão	63
Capítulo 3: Desenvolvimento de pipeline para identificação de tRFs a partir de dados de sRNA-seq	70
Materiais e métodos	71
Dados genômicos e de sequenciamento de sRNAs	71
Ambiente de desenvolvimento e teste do detectRF	71
Comparação de acurácia	72
Resultados	73
Fluxo de funcionamento do pipeline detectRF	73

Estatísticas de avaliação de contagens aleatórias e algoritmos de aceleração e acurácia	
76	
Avaliação de acurácia e tempo de execução	82
Discussão	84
<b>CONCLUSÕES</b>	<b>89</b>
<b>PERSPECTIVAS</b>	<b>91</b>
<b>BIBLIOGRAFIA</b>	<b>92</b>

## RESUMO

Fungos patógenos humanos vêm ganhando notoriedade epidemiológica com o atual aumento do número de indivíduos imunocomprometidos, de maneira a se tornar um problema de saúde pública especialmente em países de baixa renda. Dentre os principais patógenos fúngicos cujas infecções apresentam risco de óbito ao hospedeiro encontram-se as leveduras dos gêneros *Cryptococcus* e *Candida*, que apresentam características epidemiologicamente relevantes, como a existência de linhagens hiper-virulentas no gênero *Cryptococcus*, capazes de infectar indivíduos imunocompetentes, assim como de linhagens resistentes aos antifúngicos disponíveis para comercialização no gênero *Candida*. Dada a necessidade de novos e mais eficazes tratamentos para infecções causadas por esses patógenos, a avaliação do processo de infecção é o passo inicial para a descobertas de novos mecanismos-alvo para fármacos. Enquanto que pequenos RNAs são reconhecidamente associados com a regulação de diversos processos celulares, incluindo a auto-regulação durante processos infecciosos tanto por parte do patógeno quanto do hospedeiro, pouca informação se encontra disponível quanto ao papel que essas moléculas exercem na comunicação entre patógeno e hospedeiro no contexto da infecção. Além disso, classes recentemente descobertas de pequenos RNAs, ainda pouco elucidadas, podem apresentar atuação em vias relacionadas à patogênese. Assim, um melhor entendimento dos mecanismos associados a essas moléculas é necessário, podendo ser facilitado pelo desenvolvimento de novas metodologias para identificação e estudo das mesmas. A presente tese tem por objetivo avaliar a presença de diferentes tipos de pequenos RNAs em leveduras dos gêneros *Cryptococcus* e *Candida* a fim de elucidar a associação de pequenos RNAs com mecanismos de auto-regulação celular e de interação com o hospedeiro durante infecções, além de produzir novas metodologias para a análise de pequenos RNAs. Através da metodologia *in silico* desenvolvida para o tratamento de dados de sRNA-seq, foram identificados fragmentos de RNA derivados de tRNAs (tRFs) em leveduras do gênero *Cryptococcus* em linhagens deficientes e proficientes na via de RNAi canônica, sugerindo a existência de múltiplos mecanismos associados a esses sRNAs no gênero. Além disso, foram identificados tRFs e miRNAs

em vesículas extracelulares de *Candida auris*, assim como os possíveis alvos para esses sRNAs no conjunto de mRNAs de mamíferos. Por fim, foi desenvolvida uma nova ferramenta para a predição de tRFs *in silico*, com o objetivo de prover uma alternativa que una uma análise organismo-independente e com suporte estatístico à uma execução simples. Assim, a presente tese constitui base para futuras pesquisas em múltiplas linhas associadas a sRNAs regulatórios, evidenciando a existência de mecanismos celulares ainda não totalmente explorados, assim como provendo novas ferramentas para o seu estudo.

## ABSTRACT

Human fungal pathogens are gaining an epidemiologic notoriety with the increased number of immunocompromised individuals, becoming a public health issue in low-income countries. Among the main fungal pathogens that produce fatal infections are yeasts from the *Cryptococcus* and *Candida* genera, which present characteristics of epidemiologic relevance, such as the existence of hypervirulent *Cryptococcus* lineages that are capable of infecting immunocompetent hosts, as well as lineages resistant to available antifungic in the *Candida* genus. Given the need for new and more effective treatments for infections with these pathogens, the evaluation of the infection process is the first step for the discovery of new target-mechanisms for drugs. While small RNAs are widely associated with the regulation of cellular processes, including self-regulation during infections in both pathogen and host, little is known about their role in the communication between pathogen and host in the context of infection. Also, recently discovered small RNA classes that are still poorly understood may play a role in the pathogenesis. Therefore, a better understanding of the mechanisms with which these molecules are associated is necessary, which might be aided by the development of new methods to identify and study them. This thesis has as its goal to evaluate the presence of different types of small RNAs in yeasts from *Cryptococcus* and *Candida* genera in order to elucidate the association of these small RNAs with cell self-regulation and infection-related host-interaction mechanisms, while also producing new methods to analyze small RNAs. Using a developed *in silico* methodology for sRNA-seq data filtering, tRFs were identified in both RNAi-deficient and proficient *Cryptococcus* lineages, suggesting that multiple mechanisms are associated with such sRNAs in this genera. Also, both tRFs and miRNAs were identified in *Candida auris* extracellular vesicles, as well as possible targets for these sRNAs within the mRNA set of mammals. At last, a new tool for *in silico* tRF prediction was developed, aiming to provide an alternative that combines an organism-independent and statistically supported analysis with a simple execution. Hence, the present thesis constitute basis for future research in multiple lines associated with regulatory sRNAs, highlighting the

existence of molecular mechanisms that are not fully understood, as well as providing new tools to study them.

## Lista de Abreviaturas e Símbolos

5'-RACE	<i>5'-Rapid Amplification of cDNA Ends</i> (5'-Amplificação Rápida das Extremidades do cDNA)
Ago	Proteína Argonauta
miRNA	MicroRNA
mRNA	<i>Messenger RNA</i> (RNA mensageiro)
GO	<i>Gene Ontology</i> (Ontologia gênica)
piRNA	<i>Piwi-interacting RNA</i> (RNAs interatores de Piwi)
RISC	<i>RNA-induced silencing complex</i> (Complexo de silenciamento induzido por RNA)
RNAi	<i>RNA interference</i> (RNA de interferência)
RNA-seq	<i>RNA sequencing</i> (Sequenciamento de RNA)
RT-qPCR	<i>Real Time quantitative Polymerase Chain Reaction</i> (Reação em Cadeia da Polimerase quantitativa em tempo real)
siRNA	<i>Small interfering RNA</i> (Pequeno RNA de interferência)
snoRNA	<i>Small nucleolar RNA</i> (Pequeno RNA nucleolar)
snRNA	<i>Small nuclear RNA</i> (Pequeno RNA nuclear)
sRNA	<i>Small RNA</i> (Pequeno RNA)
sRNA-seq	<i>Small RNA sequencing</i> (Sequenciamento de pequeno RNA)
tiRNA	<i>tRNA-derived stress-induced RNA</i> (RNA derivado de tRNA induzido por estresse)
tRF	<i>tRNA-derived RNA fragment</i> (Fragmento de RNA derivado de tRNA)
tRNA	<i>Transfer RNA</i> (RNA de transferência)
tsRNA	<i>tRNA-derived small RNA</i> (Pequeno RNA derivado de tRNA)
UTR	<i>Untranslated Region</i> (Região não-traduzida)

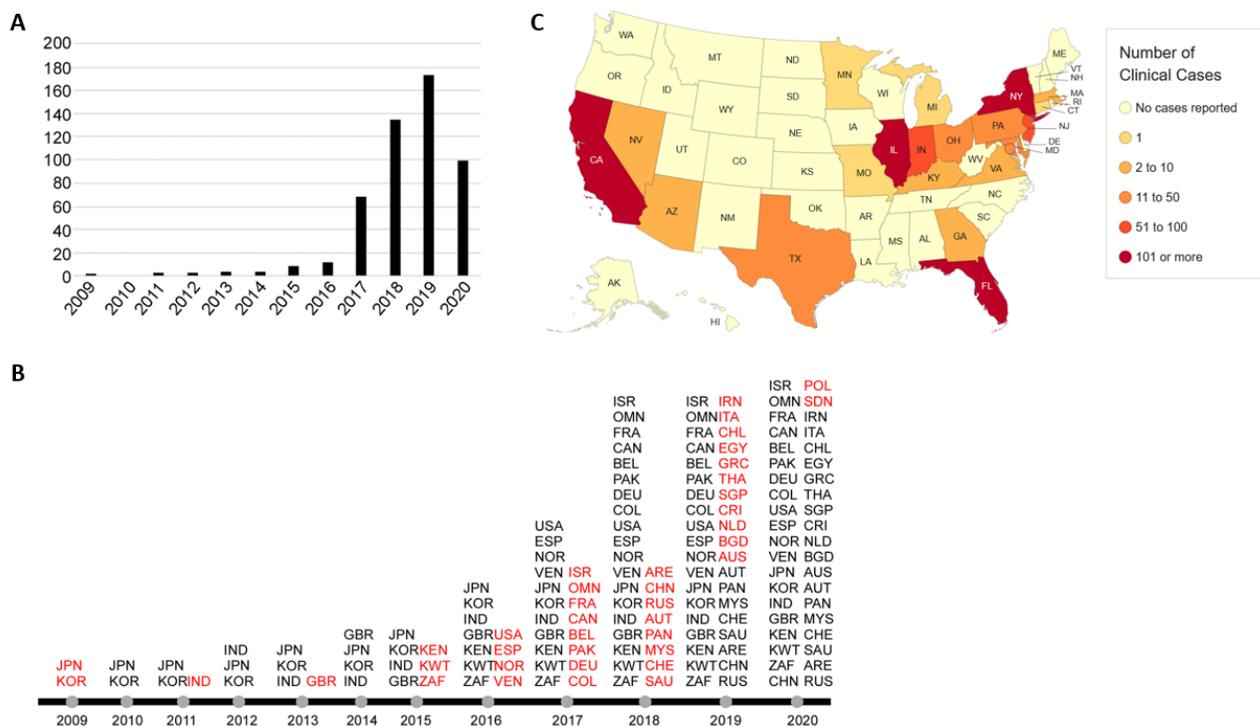
# 1. INTRODUÇÃO

## 1.1. Fungos Patógenos Humanos

Fungos patógenos humanos têm se tornado uma grande fonte de preocupação para a saúde pública nas últimas décadas, sendo estimados como responsáveis aproximadamente por 1.5 milhões de mortes anualmente, número equivalente à mortalidade por tuberculose e mais de 3 vezes maior que o número de mortes por malária (BROWN et al., 2012; FIRACATIVE, 2020). O aumento do número de infecções fúngicas se dá principalmente pelo também aumento do número de indivíduos imunocomprometidos, sendo esse último associado com uso de terapias imunossupressoras, como nos casos de pacientes com órgãos transplantados, de quimioterápicos para o tratamento de câncer, que podem gerar quadros de neutropenia, e de portadores de síndromes de imunodeficiência, como em casos de HIV (BAJPAI et al., 2019; FIRACATIVE, 2020). Frente às constantes mudanças na epidemiologia de doenças fúngicas de risco à vida, fungos patogênicos dos gêneros *Candida* e *Cryptococcus* encontram-se consistentemente entre os principais causadores das mesmas (BONGOMIN et al., 2017; BROWN et al., 2012; FIRACATIVE, 2020).

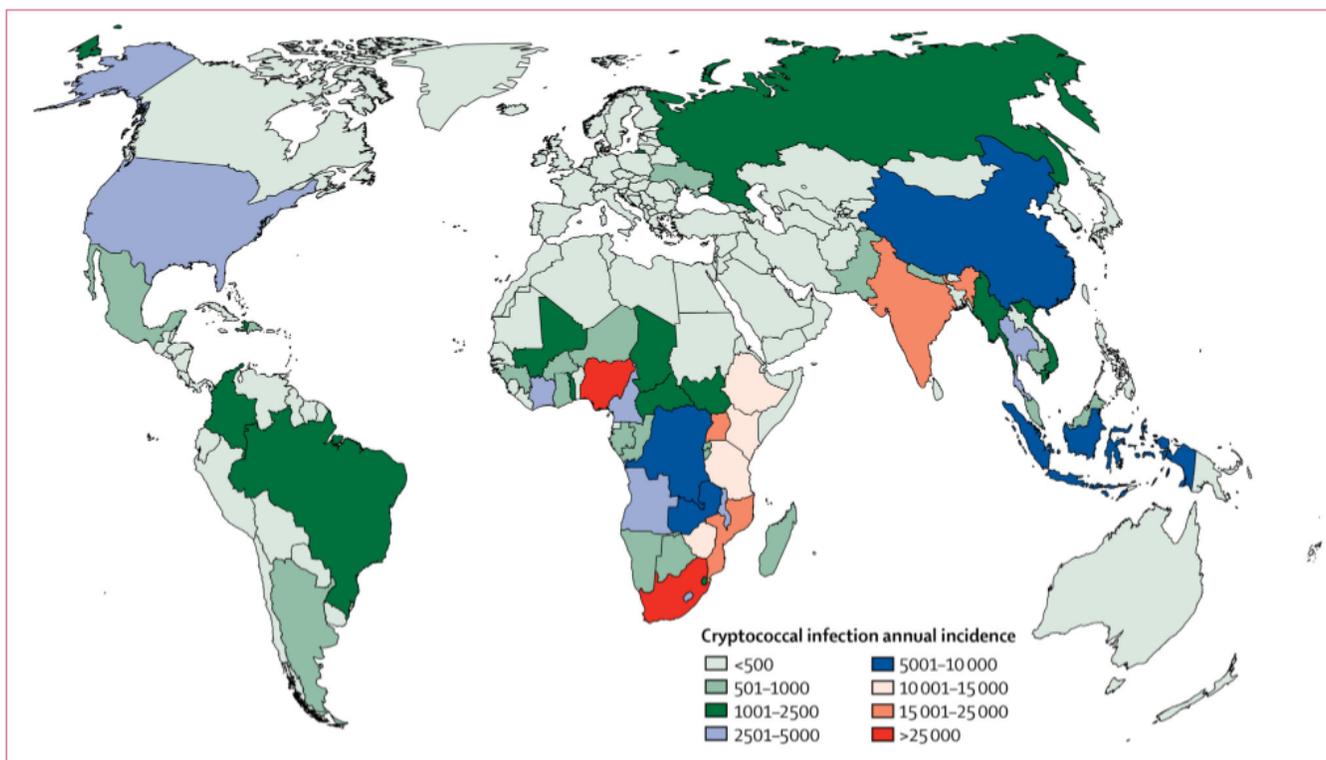
As leveduras do gênero *Candida* fazem parte da microbiota normal que pode colonizar a mucosa oral, gastrointestinal e vaginal, além da pele de indivíduos saudáveis (TALAPKO et al., 2021). Além de casos de infecções de mucosas, quadros mais severos ocasionados por leveduras desse gênero ocorrem por infecções invasivas, como quadros de candidemia, no qual a levedura infecta o sistema circulatório do paciente, e está associado com uma mortalidade de mais de 40% mesmo após a implementação de uso de novos agentes antifúngicos (FIRACATIVE, 2020). Além disso, infecções invasivas por leveduras desse gênero estão especialmente associadas com o desenvolvimento de infecções hospitalares, tornando-as um problema de saúde pública. Enquanto que *Candida albicans* é o

principal agente associado com infecções invasivas, a emergente espécie *Candida auris* vem rapidamente ganhando atenção da comunidade científica, com o número de países com casos de infecção por *C. auris* relatados aumentando de 2 em 2009 para 39 em 2020 (Figura 1), estando entre estes últimos representados países de todos os continentes com exceção da Antártida (DU et al., 2020). Além do rápido crescimento em número de casos reportados, uma das maiores preocupações com relação à *C. auris* é sua ampla resistência aos fármacos antifúngicos disponíveis atualmente, com algumas linhagens apresentando resistência à todas as classes de antifúngicos disponíveis.



**Figura 1: Impacto científico e epidemiológico de *C. auris*.** (A) Número de artigos científicos publicados tendo como tema *C. auris* entre janeiro de 2009 e junho 2020. (B) Países com casos reportados de infecção por *C. auris* entre janeiro de 2009 e junho 2020. Vermelho denota o ano no qual o país teve seu primeiro caso registrado. (ARE) Emirados Árabes Unidos, (AUS) Austrália, (AUT) Áustria, (BEL) Bélgica, (BGD) Bangladesh, (CAN) Canadá, (CHE) Suíça, (CHL) Chile, (CHN) China, (COL) Colômbia, (CRI) Costa Rica, (DEU) Alemanha, (EGY) Egito, (ESP) Espanha, (FRA) França, (GBR) Reino Unido, (GRC) Grécia, (IND) Índia, (IRN) Irã, (ISR) Israel, (ITA) Itália, (JPN) Japão, (KEN) Quênia, (KOR) Coreia do Sul, (KWT) Kuwait, (MYS) Malásia, (NLD) Países Baixos, (NOR) Noruega, (OMN) Omã, (PAK) Paquistão, (PAN) Panamá, (POL) Polônia, (RUS) Rússia, (SAU) Arábia Saudita, (SDN) Sudão, (SGP) Singapura, (THA) Tailândia, (USA) Estados Unidos da América, (VEN) Venezuela, (ZAF) África do Sul. (A e B) Adaptados de Du et al. 2020 (DU et al., 2020) (C) Casos clínicos reportados de *C. auris* nos Estados Unidos da América entre Janeiro e Dezembro de 2021, conforme o CDC. Adaptado de [www.cdc.gov/fungal/candida-auris/tracking-c-auris.html](http://www.cdc.gov/fungal/candida-auris/tracking-c-auris.html)

*Cryptococcus neoformans* e *Cryptococcus gattii* são as principais espécies patogênicas do gênero *Cryptococcus*, cujo quadro de infecção é denominado criptococose, caracterizada por um caráter sistêmico, podendo acometer diversos órgãos (KWON-CHUNG et al., 2014). A criptococose usualmente se restringe à infecção dos pulmões, sendo a via de entrada majoritária das leveduras no organismo dos hospedeiros, podendo se disseminar para o sistema nervoso central. A infecção do sistema nervoso central, denominada meningite criptocócica, é o quadro mais grave da doença, podendo atingir uma mortalidade de mais de 70% em países de baixa renda, e é responsável por 15% das mortes associadas à HIV em um nível global (PASQUIER et al., 2018; RAJASINGHAM et al., 2017; TSAI et al., 2021). Enquanto *C. neoformans* é o principal causador de quadros de criptococose e *C. gattii* não apresenta tropismo pelo sistema nervoso central, de maneira a não produzir usualmente casos de meningite criptocócica, *C. gattii* apresenta linhagens hiper-virulentas responsáveis com surtos da ilha de Vancouver no Canadá e na região noroeste do Pacífico nos Estados Unidos, além de uma maior associação com infecções de indivíduos aparentemente imunocompetentes, ao contrário de *C. neoformans*, que usualmente acomete indivíduos imunocomprometidos (KWON-CHUNG et al., 2014). Assim, dada a alta taxa de mortalidade por criptococose e a existência de linhagens hiper-virulentas capazes de infectar indivíduos imunocompetentes, leveduras patogênicas do gênero *Cryptococcus* apresentam grande relevância para a saúde pública de diversos países (Figura 2).



**Figura 2: Estimativa de infecções anuais por *Cryptococcus* em 2014.** Obtido de Rajasingham *et al.* 2017 (RAJASINGHAM *et al.*, 2017). Estimativa do número de pessoas com antigenemia para *Cryptococcus* em 2014, com 95% de intervalo de confiança.

Mesmo com crescente relevância de fungos patógenos para a saúde humana, as alternativas de tratamento para fungos não seguem a mesma tendência, sendo ainda extremamente escassas e restritas a apenas 5 grupos de moléculas: azóis, equinocandinas, polienos, alilaminas e análogos de pirimidina (HOKKEN *et al.*, 2019). O reduzido número de fármacos antifúngicos disponíveis se deve majoritariamente à conservação de muitos processos biológicos e de proteínas entre organismos fúngicos e seus hospedeiros, de maneira que compostos com atividade antifúngica podem apresentar alta toxicidade a células humanas, tendo como exemplo a anfotericina B, que apesar de ser um fármaco com eficácia no tratamento de fungos é utilizado como segunda linha de tratamento devido à toxicidade renal (FRANÇA *et al.*, 2014). Além do número de fármacos disponíveis, outro problema é a disponibilidade e acessibilidade aos principais antifúngicos do mercado. Enquanto certos países apresentam indisponibilidade de algumas das principais opções de tratamento para micoses fatais

por falta de registro, o elevado custo dos fármacos dificulta ainda mais sua administração correta, com estimativas de até U\$ 450,00 para uma terapia convencional de 2 semanas levando em consideração apenas a fase de indução do tratamento (KNEALE et al., 2016). Por fim, a presença e desenvolvimento de resistência por diversos fungos clinicamente relevantes agrava ainda mais o cenário de baixa variedade de fármacos, sendo reportados fungos resistentes para todas as classes de fármacos atualmente aplicados no tratamento de doenças fúngicas, além do crescente número de linhagens e espécies que apresentam essas resistências (HOKKEN et al., 2019; PERLIN; RAUTEMAA-RICHARDSON; ALASTRUEY-IZQUIERDO, 2017; WIEDERHOLD, 2017), evidenciando a urgente necessidade por novas alternativas de tratamento.

Um ponto crucial para o estabelecimento de infecções e que pode ser visado para o desenvolvimento de novas alternativas de tratamento são os processos de adaptação pelos quais patógenos passam durante a interação patógeno-hospedeiro. A partir do momento no qual o patógeno entra em contato com seu hospedeiro, o patógeno passa por alterações em seu padrão de expressão gênica e metabolismo na tentativa de estabelecer e manter um ambiente de infecção favorável, ao mesmo tempo que evadindo as defesas do organismo hospedeiro (BROCK, 2009; CHAI et al., 2009). Assim como com as moléculas efetoras e as alterações metabólicas e fenotípicas geradas por esse processo, o estudo dos sistemas de regulação da expressão gênica que governam o processo podem levar à identificação de pontos passíveis de intervenção por novas terapias.

## 1.2. Pequenos RNAs Regulatórios

Os RNAs podem ser agrupados em duas principais classes de acordo com sua capacidade de codificar proteínas, sendo esses grupos os RNAs codificantes, que apresentam como finalidade a sua tradução em proteínas, grupo composto exclusivamente por mRNAs, e os RNAs não-codificantes, composto por todos os RNAs sem capacidade de codificar para proteínas (DINGER et al., 2008). Os RNAs

não-codificantes podem apresentar papéis que variam desde componentes estruturais de complexos macromoleculares, como o RNA longo não-codificante NEAT1, envolvido da formação dos corpúsculos subnucleares denominados *paraspeckles* (BOND; FOX, 2009), até a atuação direta por meio da atividade enzimática do próprio RNA, como as ribozimas (DOHERTY; DOUDNA, 2001). Para os RNAs não-codificantes, existe uma subdivisão do grupo de acordo com o tamanho dos RNAs, sendo RNAs com mais de 200 nucleotídeos classificados como RNAs longos não-codificantes, e RNAs com tamanho menor do que esse limite, classificados como pequenos RNAs (sRNAs, do inglês *small RNAs*) (MA; BAJIC; ZHANG, 2013).

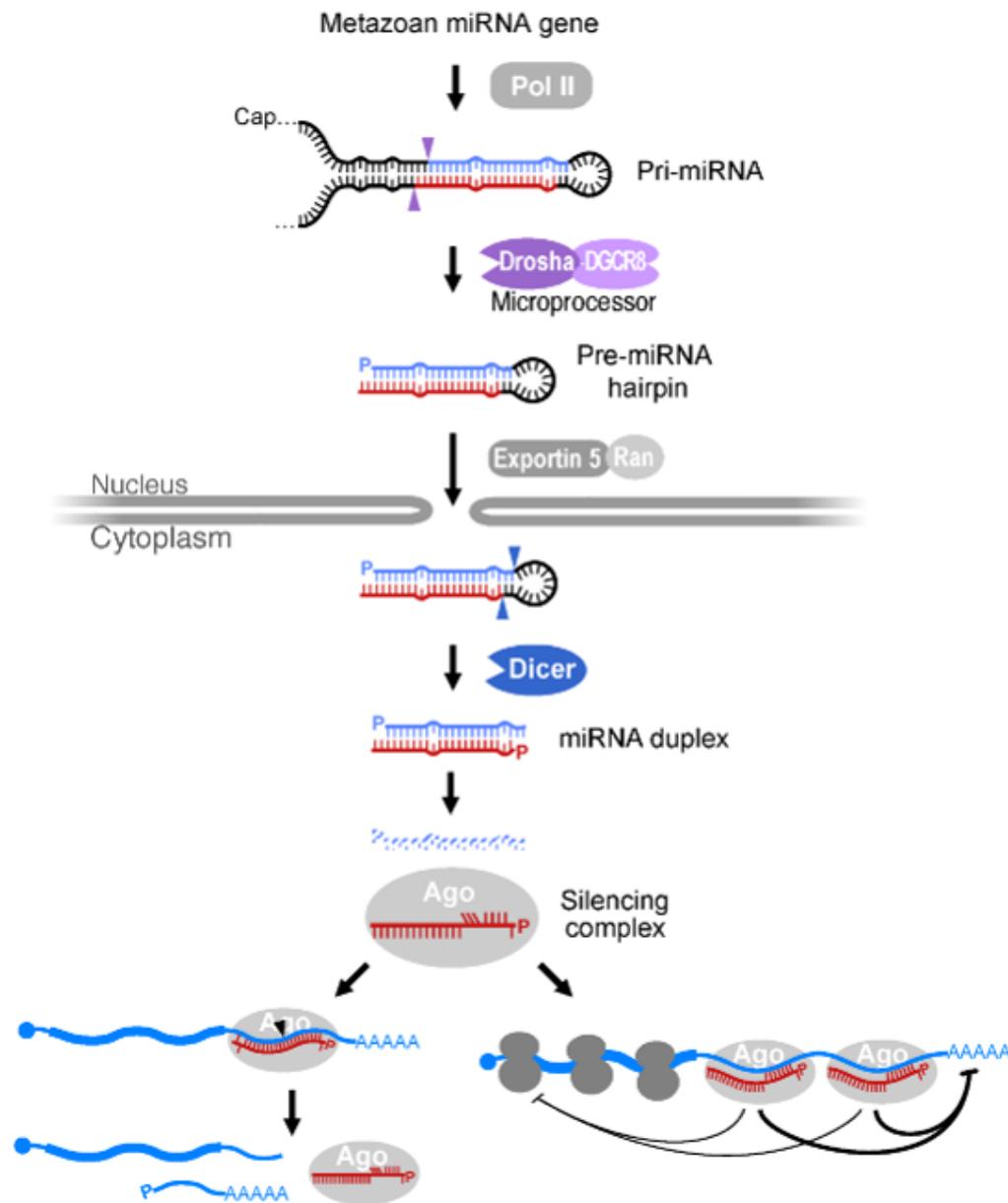
Apesar de que alguns tipos de sRNAs terem sido caracterizados funcionalmente há décadas, como os RNAs de transferência (tRNAs) e alguns RNAs ribossomais (rRNAs), a maioria dos sRNAs não tiveram grande atenção até a descoberta de lin-4, o primeiro microRNA (miRNA) identificado (LEE; FEINBAUM; AMBROS, 1993). Com a descoberta de miRNAs e a constatação de que sRNAs podem atuar na regulação da expressão gênica, uma maior atenção passou a ser dada a estas moléculas, com consequente identificação de novas classes de sRNAs regulatórios em organismos de todos os domínios da vida, assim como dos mecanismos a elas associadas (ATAYDE; TSCHUDI; ULLU, 2011; BABSKI et al., 2014; DUTTA; SRIVASTAVA, 2018; GULERIA et al., 2011; TORRES-MARTÍNEZ; RUIZ-VÁZQUEZ, 2017; ZHANG, 2009). Diversas funções associadas a esses sRNAs regulatórios já foram descritas, podendo eles atuarem na regulação da expressão gênica a nível pós-transcricional, remodelamento de cromatina e controle da transcrição, proteção contra ácidos nucleicos exógenos e defesa do genoma contra elementos de transposição (CHEKULAEVA; FILIPOWICZ, 2009; GIRARD; HANNON, 2008; KARGINOV; HANNON, 2010; MOAZED, 2009), evidenciando a essencialidade dos sRNAs na biologia celular e manutenção da homeostase dos organismos.

sRNAs são moléculas com alta diversidade funcional e estrutural, cujas diferenças se caracterizam pelo tamanho, constituição e função. siRNAs são sRNAs produzidos a partir de RNAs de fita dupla e associados com a defesa a RNAs exógenos (CARTHEW; SONTHEIMER, 2009). piRNAs estão relacionados especialmente ao remodelamento de cromatina e controle de elementos de

transposição em células germinativas (FU; WANG, 2014; WEICK; MISKA, 2014). Já miRNAs são sRNAs regulatórios simples-fita com tamanho usual em torno de 22 nucleotídeos e papel proeminente na regulação pós-transcricional de mRNAs, sendo estimados que 60% do transcrito humano seja regulado pelos mesmos, assim como 17% em *C. elegans* e 27% no gênero *Drosophila* (BARTEL, 2018). Além de sua função na regulação da taxa de meia vida de mRNAs, os miRNAs apresentam uma ampla gama de funções, como a regulação da transcrição e remodelamento de cromatina e a defesa contra elementos repetitivos e RNAs exógenos (HUANG; LI, 2012; JEANG, 2012; O'CONNOR; VANICEK; MURPHY, 2014; PU et al., 2019; TRISTÁN-RAMOS et al., 2020), ressaltando a importância dessa classe de sRNAs.

Dentre as vias possíveis de atuação dos microRNAs está a via de RNA de interferência, via da qual diversas outras classes de sRNAs regulatórios podem participar, mesmo que não da mesma maneira que os microRNAs (BARTEL, 2018; WILSON; DOUDNA, 2013). Essa via é conservada entre os diferentes taxa eucarióticos, apesar de pequenas variações em alguns pontos da via serem observadas entre os mesmos, assim como existem organismos nos quais a via é incompleta, caso comum em fungos (NICOLÁS; GARRE, 2016). Utilizando a via associada aos miRNAs de metazoários como exemplo (Figura 3) (BARTEL, 2018), o passo inicial ocorre logo após a transcrição do gene de miRNA pela RNA polimerase II (Pol II) em um pri-miRNA, molécula de RNA de aproximadamente 70 nucleotídeos que possui uma estrutura de grampo. O pri-RNA é clivado pelo complexo microprocessador, composto pelas proteínas Drosha e DGCR8, clivagem que ocorre na porção dupla-fita do pri-miRNA e libera um grampo de RNA. Esse grampo apresenta tamanho reduzido quando comparado ao pri-miRNA e não possui grandes extremidades simples-fita (com a exceção de 2 nucleotídeos livres na extremidade 3'), sendo denominado pré-miRNA. O pré-miRNA é então exportado para o citoplasma pelas proteínas exportina-5 e Ran, e é novamente processado, sendo clivado pela proteína Dicer na extremidade contendo o *loop*, produzindo um RNA dupla-fita com 2 nucleotídeos livres em ambas as extremidades. Uma das fitas é considerada o miRNA funcional, estando essa ainda pareada ao miRNA “estrela” (ou miRNA\*), produto do processamento do pré-miRNA usualmente não-funcional. Através de extremidades não-pareadas da dupla-fita de

miRNA-miRNA\*, ocorre a interação com proteínas do complexo de silenciamento mediado por RNA (RISC), especificamente com proteínas Argonata (Ago), sendo selecionada preferencialmente a fita que apresente extremidade 5'-A ou 5'-U, assim como a que apresentar a extremidade 5' da região dupla-fita com menor estabilidade de pareamento. Usualmente, por essas características, a fita considerada miRNA é carregada na proteína Argonata, sendo o miRNA\* descartado e degradado. Por fim, o complexo RISC carregado com miRNA interage com mRNAs até a identificação de seu mRNA-alvo por meio do pareamento de bases entre o miRNA carregado e o mRNA alvo. Esse pareamento de bases usualmente ocorre na região 3'-UTR de mRNAs para alvos de miRNA, sendo primariamente regido pelo pareamento do *seed* do miRNA, que corresponde aos nucleotídeos 2 a 7 da extremidade 5' do miRNA, podendo haver a extensão do pareamento entre os RNAs além dessa porção do miRNA. Após a identificação, dependendo da Argonata utilizada pelo complexo RISC e da extensão do pareamento entre miRNA e alvo, o mRNA-alvo pode ser hidrolisado diretamente pela proteína Argonata e conseqüentemente degradado, ou pode ocorrer apenas a inibição da tradução e aceleração da degradação do mRNA pela interação com RISC, por meio do recrutamento por parte do RISC de proteínas acessórias associadas a essas funções.



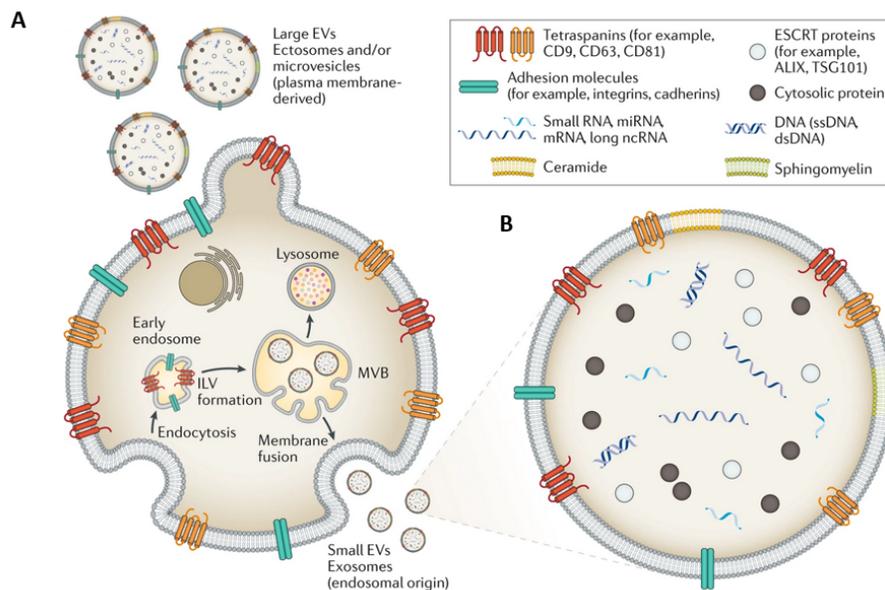
**Figura 3: A produção e funcionamento de microRNAs associada à via de RNA de interferência.** Representação gráfica dos passos principais envolvidos na produção e funcionamento de microRNAs pela via canônica de RNA de interferência de metazoários, conforme previamente descrito. Adaptado de Bartel 2018 (BARTEL, 2018).

### 1.3. Pequenos RNAs e a Interação Patógeno-Hospedeiro

Da mesma forma como sRNAs têm a capacidade de atuar sobre a regulação da expressão gênica relacionada a diversos processos celulares, sua função se estende à interação patógeno-hospedeiro, seja no desenvolvimento da infecção em patógenos ou na defesa contra infecções nos hospedeiros. Apesar do limitado número de estudos que avaliam o potencial de sRNAs como reguladores de determinantes de virulência e do fenótipo de virulência em patógenos (CHABELSKAYA; GAILLOT; FELDEN, 2010; EISENBART et al., 2020; MURPHY; PAYNE, 2007; PATEL; NAIR, 2021; RAMAN et al., 2017), existe uma extensa literatura demonstrando a produção de sRNAs com funções associadas à imunidade por parte do hospedeiro durante as interações com patógenos em diversos tipos de infecções (ACUÑA; FLOETER-WINTER; MUXEL, 2020; CROSTON et al., 2018; GUO; LI; DING, 2019). Dada a presença ubíqua de sRNAs regulatórios, sendo relatados em todos os domínios da vida, e a demonstrada participação dos mesmos no processo de regulação do sistema imunitário por parte de hospedeiros durante infecções, é possível hipotetizar que patógenos produzam sRNAs que mimetizam ou bloqueiem a função de sRNAs regulatórios do sistema imune do hospedeiro.

Uma das formas pelas quais sRNAs podem ser transmitidas entre patógenos e hospedeiros são por meio das vesículas extracelulares. Vesículas extracelulares são um grupo heterogêneo de partículas delimitadas por uma bicamada lipídica, produzidas por células e sem capacidade de replicação. Apesar da variação na literatura quanto à delimitação de classes de vesículas extracelulares, consistentemente duas classes principais de vesículas são relacionadas à comunicação intercelular (Figura 4A): (i) os exossomos, com tamanho entre 30-100 nm (intervalo que varia entre diferentes autores em até 100 nm), produzidas a partir de endossomos por meio de invaginações dos mesmos após sua maturação, formando corpos multivesiculares, sendo posteriormente liberadas a partir da fusão entre os corpos multivesiculares e a membrana plasmática, processo o qual é evolutivamente conservado; e (ii) microvesículas, com tamanho de 50 nm até 1 µm, produzidas a partir de brotamentos na membrana plasmática (LIXWEILER; JUNKER, 2020; MUNHOZ DA ROCHA et al., 2020; ROBBINS;

MORELLI, 2014; VAN DER POL et al., 2012). Vesículas extracelulares não apenas são produzidas por organismos de todos os domínios da vida, como também têm papel relevante durante a interação patógeno-hospedeiro. Vesículas extracelulares de patógenos apresentam relação com o processo de infecção, participando de processos como a formação de biofilmes, aquisição de nutrientes, transporte de fatores de virulência e modulação da resposta imune do hospedeiro (MUNHOZ DA ROCHA et al., 2020). Da mesma forma, vesículas extracelulares dos hospedeiros atuam na auto-regulação do sistema imunitário durante infecções, participando de processos como a apresentação de antígenos e modulação da resposta por meio de moléculas sinalizadoras presentes nessas vesículas (ROBBINS; MORELLI, 2014). De maneira interessante, tanto vesículas produzidas por patógenos quanto por hospedeiros apresentam carga de RNA, incluindo pequenos RNAs (Figura 4B), com demonstração da transferência de RNAs presentes em vesículas para células receptoras das mesmas (MUNHOZ DA ROCHA et al., 2020).



**Figura 4: Geração de vesículas extracelulares e representação de sua estrutura e conteúdo.** (A) Processo geral de formação dos dois principais tipos de vesículas extracelulares, sendo eles microvesículas e exossomos. (B) Representação das características usuais de exossomos, como tipos de moléculas presentes em sua superfície, além de possíveis cargas presentes nas mesmas, como pequenos RNAs (*small RNA* e *miRNA*). (A e B) Adaptados de Linxweiler e Junker 2020 (LINXWEILER; JUNKER, 2020).

Apesar de alguns exemplos da presença de sRNAs patógenos sendo liberados para o hospedeiro durante infecções existirem há vários anos (BUCK et al., 2014; CHENG et al., 2013), o assunto ainda parece não possuir a atenção devida. Mesmo com indicativos de que essa comunicação ocorre de maneira ubíqua, sendo uma estratégia utilizada durante infecções tanto em plantas quanto em metazoários, assim como em infecções fúngicas e bacterianas, pouca informação e desenvolvimento de trabalhos nessa área existem (BUCK et al., 2014; CAI; HE; JIN, 2019; CHENG et al., 2013; CUI et al., 2019; KOEPPEN et al., 2016; WALZER; CHI, 2017; WANG et al., 2021). Em especial, apenas uma porção extremamente restrita desses trabalhos envolvem patógenos humanos e a utilização dos sRNAs na infecção. Vista a relevância do intercâmbio de biomoléculas por intermédio de vesículas extracelulares para o desfecho de interações patógeno-hospedeiro, assim como a documentada presença de sRNAs nas mesmas e o impacto dessa classe de moléculas na regulação da expressão gênica de programas de expressão associados ao combate a infecções em hospedeiros, a avaliação do conteúdo de sRNAs de vesículas produzidas por patógenos e seu possível papel na infecção é de grande valia para o melhor entendimento do processo.

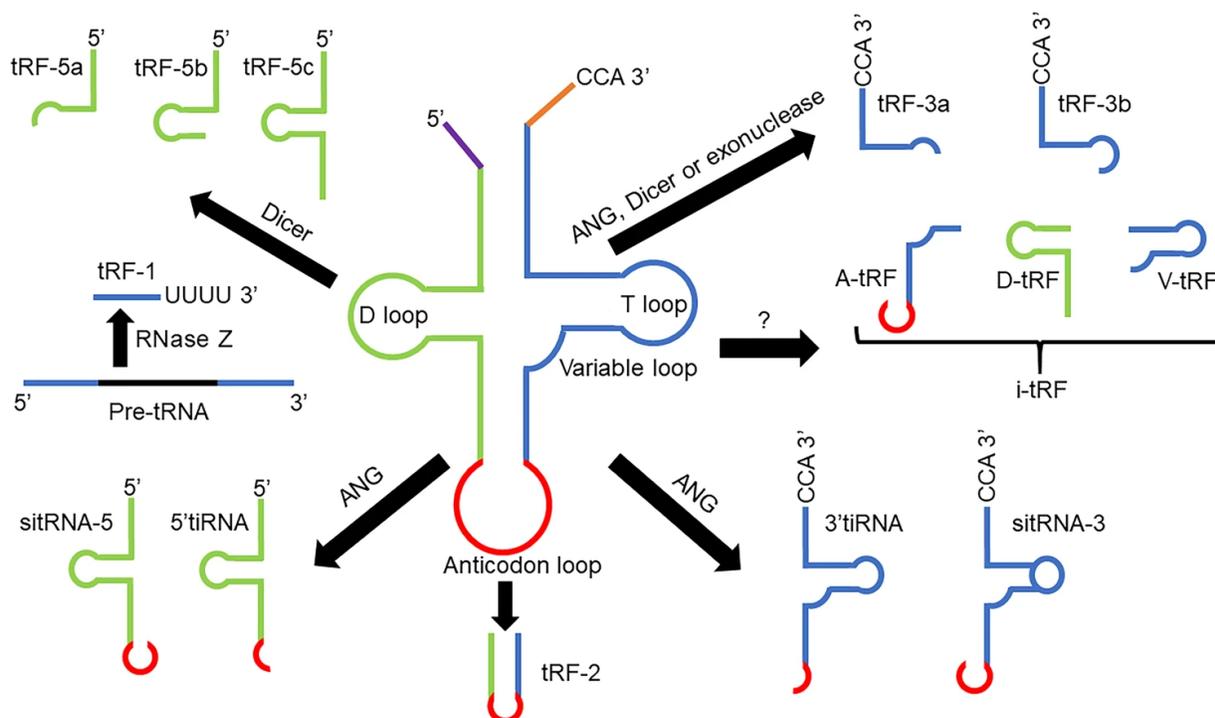
#### 1.4. Fragmentos de RNA derivados de tRNA

Além dos tipos clássicos de sRNAs regulatórios citados previamente, como os miRNAs e piRNAs, novos tipos de sRNAs vêm sendo descobertos, especialmente os originados de outras moléculas de RNA, como rRNAs, snRNAs (do inglês *small nuclear RNAs*, pequenos RNAs nucleares), snoRNAs (do inglês *small nucleolar RNAs*, pequenos RNAs nucleolares) e tRNAs (LI et al., 2012). Isso se deve tanto à crescente facilidade na geração de dados de sequenciamento de pequenos RNAs (sRNA-seq, do inglês *small RNA sequencing*) e conseqüente acúmulo de dados desse tipo disponíveis para análise, quanto por mudanças na abordagem de análise desses dados, visto que muitas dessas novas classes de sRNAs regulatórios tem por origem outros RNAs e eram anteriormente consideradas como produtos de degradação (COLE et al., 2009). Dentre as principais classes de novos sRNAs provenientes de outros tipos de RNA se

encontram os pequenos RNAs derivados de tRNA (tsRNAs, dos inglês *tRNA-derived small RNAs*), pequenos RNAs gerados a partir de uma clivagem altamente específica de tRNAs. Apesar da primeira demonstração da existência de tsRNAs datar da década de 1970 (BOREK et al., 1977; SPEER et al., 1979), até 2009 (LEE et al., 2009) esses fragmentos, como na maioria das novas classes de pequenos RNAs, eram tratados como produtos da degradação de tRNAs. Porém, após a demonstração da participação do tRF-1001 no funcionamento do ciclo celular de linhagens celulares de câncer de próstata (LEE et al., 2009), um número crescente de estudos em diversos organismos demonstrou a presença e função, pelos mais diversos mecanismos, de tsRNAs (GEBETSBERGER et al., 2012; KEAM et al., 2017; KIM et al., 2017; KUMAR et al., 2014; KUSCU et al., 2018; LUO et al., 2018; MLECZKO; CELICHOWSKI; BAŁKOWSKA-ŻYWICKA, 2018; WANG et al., 2019).

As nomenclaturas e definições sobre características dos tsRNAs variam dentro da literatura (KUMAR; KUSCU; DUTTA, 2016; LI; XU; SHENG, 2018; OBERBAUER; SCHAEFER, 2018; XU et al., 2017), de maneira a dificultar uma definição clara, baseada em trabalhos prévios, sobre essa classe. Isso pode ser explicado tanto pelo alto grau de variação nessa classe de sRNAs, como também por ser um campo de estudos novo e ainda não completamente consolidado. Porém, de maneira razoável, podemos delimitar dois grandes grupos de tsRNAs, sendo eles as metades de tRNAs (ou tiRNAs, do inglês *tRNA-derived stress-induced RNAs*) e os fragmentos de RNA derivados de tRNA (tRFs). Enquanto que as metades de tRNAs são fragmentos maiores, tendo em média de 30 a 40 nucleotídeos, produzidos a partir da clivagem dos tRNAs em seu anticódon, e estando majoritariamente associados com a clivagem por enzimas como a angiogenina em situações de estresse, os tRFs são moléculas menores, de 14 a 30 nucleotídeos, não tendo sido ainda identificada uma via clara para a produção dos mesmos, tendo porém demonstrada sua atuação na regulação celular e produção mediante a condições específicas (BAŁKOWSKA-ŻYWICKA et al., 2016; KUSCU et al., 2018; MEGEL et al., 2019; XIE et al., 2020). Os tRFs podem ainda ser subdivididos em tipos de acordo com a região das quais eles foram produzidos: (i) tRFs-5, produzidos a partir da extremidade 5' do tRNA maduro; (ii) tRFs-3, produzidos a partir da extremidade 3' do tRNA maduro e contendo a sequência CCA em sua

extremidade; (iii) tRFs-1, produzidos a partir da região *trailer* do transcrito de tRNA primário; e (iv) i-tRFs, produzidos a partir de regiões internas do tRNA maduro, sem conter extremidades do tRNA (XIE et al., 2020). Apesar das delimitações de origens mais comuns propostas, e a classificação atual acomodar tRFs oriundos de distintas regiões dos tRNAs, é comumente aceito propor que tRFs possam ser originados de qualquer fração do tRNA maduro ou primário, vista a grande variedade origens reportadas para tRFs na literatura (Figura 5).



**Figura 5: Classes e subclasses de tRFs e tsRNAs já reportadas e suas origens no tRNA parental.** Demonstração da origem dentro do tRNA parental das classes principais de tRFs já descritas na literatura, como i-tRFs, tRFs-5, tRFs-3, tRFs-2 e tRFs-1, assim como de subclasses (como para tRFs-5, tRFs-5 de “a” a “c”). Também são representados tsRNAs de outras categorias, como metades de tRNA (tiRNAs) e tsRNAs induzidos por estresses (sitRNAs). Nas setas, são apresentadas proteínas associadas com o processamento das classes e subclasses às quais as setas apontam. (ANG) Angiogenina. Obtido de Xie *et al* 2020 (XIE et al., 2020).

Assim como sua nomenclatura, a biogênese dos tRFs ainda não é plenamente estabelecida, com exceção da biogênese de tRFs-1 a partir da região *trailer* do transcrito de tRNA primário, associados com a ação da RNase Z (HARTMANN et al., 2009). Enquanto existem evidências de que a produção de alguns tRFs dependem de DICER (COLE et al., 2009; DI FAZIO et al., 2022) para o seu processamento a partir dos tRNAs progenitores, outros estudos demonstram que a produção de alguns tRFs é

independente de DICER (KUSCU et al., 2018; MEGEL et al., 2019). Assim, é provável que não exista apenas uma via responsável pela produção de tRFs, mas sim que diferentes tRFs sejam produzidos a partir de diferentes mecanismos, incluindo mecanismos ainda não identificados. Da mesma forma, os mecanismos de ação dos tRFs, apesar da existência de exemplos bem elucidados, parecem variar entre diferentes tRFs. Enquanto já foi demonstrada a existência de mecanismos de regulação por meio da associação direta com outros RNAs impedindo a formação de estruturas secundárias (KIM et al., 2017), e a interação com proteínas e ribossomos impedindo a interação de macromoléculas relacionadas com a tradução (GEBETSBERGER et al., 2012; GOODARZI et al., 2015; KEAM et al., 2017), o mecanismo associado com tRFs que recebe maior atenção da comunidade científica é a atuação na repressão por meio da via de RNA de interferência de uma maneira similar a miRNAs (KUSCU et al., 2018). Além da demonstração de que tRFs possam atuar na repressão de alvos de mRNA por meio da interação com Argonata da mesma forma que miRNAs, tRFs apresentam tamanho similar ao de miRNAs e ampla associação com proteínas Argonata (KUMAR et al., 2014). Assim, dada a relevância dos miRNAs na regulação da expressão gênica e sua popularidade, existe um certo apelo à atribuição de funcionalidade de tRFs à via de RNA de interferência. Porém, apesar da possibilidade de tRFs atuarem pela via de RNAi como miRNAs, as demonstrações diretas dessa funcionalidade são pontuais (HAUSSECKER et al., 2010; KUSCU et al., 2018; MAUTE et al., 2013), já que evidências indiretas como a associação com proteínas Argonatas não representa funcionalidade por meio das mesmas. Em vista da grande variabilidade presente na classe dos tRFs, em termos de características e biogênese, e a demonstração de diversos mecanismos alternativos à atuação pela via de RNAi, assim como a presença de tRFs em organismos que não apresentam uma via de RNAi canônica funcional (BAKOWSKA-ŻYWICKA et al., 2016; WANG et al., 2019), torna-se necessária a avaliação extensiva da associação entre tRFs e a via de RNAi, a fim de definir se essa associação é restrita a tRFs específicos ou uma regra e evitar a atribuições de função errôneas baseadas em evidências indiretas.

Um ponto que dificulta o avanço do conhecimento sobre tRFs é a indisponibilidade de ferramentas para a identificação dos mesmos. Uma das estratégias utilizadas para a identificação de tRFs são análises *in silico* envolvendo dados de sRNA-seq, sendo posteriormente confirmados os tRFs identificados por outras metodologias *in vivo* ou *in vitro*. Apesar da grande disponibilidade de dados e da disponibilidade de ferramentas de análise de tRFs, existe uma limitação em termos de programas disponíveis para fazer sua identificação, com poucos programas disponíveis que, em sua maioria, apresentam dificuldades para a execução ou algoritmos de baixa confiabilidade (LOHER; TELONIS; RIGOUTSOS, 2017; RYVKIN et al., 2014; SELITSKY; SETHUPATHY, 2015; SHI et al., 2018; WANG et al., 2022; XU et al., 2017; ZHENG et al., 2016). Dentre as dificuldades para a execução encontra-se a necessidade de preparação de arquivos complexos e não-usuais para análises de bioinformática, como no caso do programa MINTmap quando executado para organismos que não os quais já possuem esses arquivos prontos fornecidos pelo autor (LOHER; TELONIS; RIGOUTSOS, 2017), levando à necessidade de desenvolvimento de novos códigos para processamento de dados ou trabalho manual no lugar da utilização de arquivos em formato padrão, como fasta e gff/gtf ou genbank, o que é especialmente problemático para a utilização por parte de usuários não experientes. Ainda no tema de execução em organismos que não os quais utilizados pelo autor para o desenvolvimento do programa, um caso ainda mais extremo ocorre com a versão *online* do tsRfun (WANG et al., 2022), no qual apenas dados de *Homo sapiens* podem ser analisados. Esse é um problema extremamente relevante, visto que ele afeta diretamente a expansão do conhecimento sobre tRFs em organismos que não de uma seleta minoria de organismos-modelo. Além dos problemas de execução, boa parte dos programas baseiam a predição de tRFs em algoritmos ingênuos e de baixa confiabilidade, como nos casos da versão local do tsRfun e o tDRmapper (SELITSKY; SETHUPATHY, 2015), que não levam em consideração a possibilidade dos *reads* alinharem em outros locais do genoma que não os tRNAs, gerando falsos positivos. Porém, a maior limitação é a escassez de programas que façam a identificação de tRFs baseado em testes estatísticos que suportem a presença dos tRFs. Enquanto a grande maioria dos preditores de tRFs, como o tDRmapper, SPORTs (SHI et al., 2018)

e MINTmap, indicam tRFs baseados na abundância relativa de *reads* em uma região do tRNA com relação às demais ou sequer fazem testes, relatando todos os *reads* alinhados contra tRNAs, apenas o tsRfun se baseia em testes estatísticos, descritos em sua versão prévia denominada tRF2Cancer (ZHENG et al., 2016), para diferenciar tRFs de regiões contendo *reads* provenientes da degradação dos tRNAs. Assim, um passo de grande importância para o desenvolvimento da área é a geração de programas ou *pipelines* mais acessíveis e robustos, que facilitem a identificação de tRFs e expansão do conhecimento sobre os mesmos.

## 2. OBJETIVOS

Avaliar a presença e identidade de pequenos RNAs regulatórios em fungos patógenos humanos, além de desenvolver novas metodologias *in silico* para a identificação de pequenos RNAs regulatórios.

### **Específicos:**

- Identificar tRFs em *C. gattii* R265 e *C. neoformans* H99, assim como seus respectivos mRNAs alvos próprios;
- Identificar tRFs e miRNAs em vesículas extracelulares de *C. auris*, assim como seus mRNAs alvos em *M. musculus* e *H. sapiens*;
- Desenvolver novas estratégias de processamento de dados de sequenciamento e ferramentas *in silico* para a análise e predição de tRFs de maneira organismo-independente.

### 3. RESULTADOS

Os resultados obtidos na presente tese estão organizados em 3 capítulos, cada qual abordando um problema específico relacionado aos objetivos propostos anteriormente. Na exceção de capítulos apresentados na forma de artigo científico e discussão geral do artigo, a organização interna de cada capítulo seguirá a estrutura: (i) metodologias aplicadas para o desenvolvimento dos resultados obtidos para o capítulo em questão, (ii) resultados obtidos relativos ao ponto abordado pelo capítulo, (iii) discussão dos resultados observados frente ao conhecimento atual sobre o tema.

#### **Temas dos capítulos:**

##### 1. Avaliação de tRFs em *Cryptococcus spp*

Análise *in silico* de tRFs baseada em dados de sRNA-seq em *Cryptococcus gattii* R265 e *Cryptococcus neoformans* H99 a fim de avaliar o impacto da perda da via de RNAi em *C. gattii* R265 sobre a produção de tRFs. Desenvolvimento de novas estratégias de processamento de dados a serem utilizados na identificação de tRFs com a finalidade de reduzir a taxa de falsos-positivos por programas disponíveis na literatura. Esse capítulo é apresentado na forma de artigo, já publicado (STREIT et al., 2021).

##### 2. Avaliação de sRNAs regulatórios em vesículas extracelulares de isolados *Candida auris*

Utilização da mesma abordagem desenvolvida no Capítulo 1 para a identificação de tRFs, mas tendo como objeto de estudos o conjunto de sRNAs isolados de vesículas extracelulares de isolados de *C. auris*, também identificando por meio de metodologias pertinentes a presença de miRNAs nas mesmas. Busca por alvos para os sRNAs regulatórios entre os mRNAs de hospedeiros, *Homo sapiens* e *Mus musculus*.

##### 3. Desenvolvimento de *pipeline* para identificação de tRFs a partir de dados de sRNA-seq

Desenvolvimento de *pipeline* para identificação de tRFs baseado nos testes estatísticos do *software* tsRfun, nas limitações apresentadas por diversos programas descritas previamente e nas práticas de processamento de dados aplicadas nos capítulos 2 e 3. Comparação com tsRfun e aplicação em dados de *C. auris*.

### 3.1. Capítulo 1: Identificação de tRFs em *Cryptococcus spp*

Esse capítulo é apresentado na forma de artigo, sendo o mesmo já publicado no periódico *Fungal Biology*, sob o DOI (10.1016/j.funbio.2020.12.003). Nele é avaliada a presença de tRFs em duas leveduras patogênicas evolutivamente próximas, *C. gattii* linhagem R265 e *C. neoformans* linhagem H99. A via de RNAi é não-funcional na primeira, sendo possível avaliar o impacto da perda da mesma sobre os tRFs. Inicialmente é proposta uma abordagem de processamento de dados que reduza a taxa de falsos-positivos na predição de tRFs pelo *software* tDRmapper, sendo conseguinte a mesma aplicada para a identificação de tRFs *in silico* por meio de dados de RNA-seq. Em seguida, são analisadas as características dos tRFs identificados em ambos os organismos, sendo também feita a identificação *in silico* de possíveis alvos de mRNA para os tRFs e a avaliação da relação do nível de expressão entre tRFs e seus alvos. São também realizados experimentos *in vitro* para verificar a presença dos tRFs preditos, a fim de confirmar as predições de tRFs, assim como a avaliação da clivagem dos alvos de tRFs. Por fim, os resultados obtidos são discutidos frente ao cenário da perda da via de RNAi e as características esperadas de pequenos RNAs que por nela atuam.



# Analysis of tRNA-derived RNA fragments (tRFs) in *Cryptococcus* spp.: RNAi-independent generation and possible compensatory effects in a RNAi-deficient genotype



Rodrigo Silva Araujo Streit<sup>a</sup>, Patrícia Aline Gröhns Ferrareze<sup>a</sup>,  
Marilene Henning Vainstein<sup>a,b</sup>, Charley Christian Staats<sup>a,b,\*</sup>

<sup>a</sup> Programa de Pós-Graduação em Biologia Celular e Molecular, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

<sup>b</sup> Departamento de Biologia Molecular e Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

## ARTICLE INFO

### Article history:

Received 2 April 2020

Received in revised form

30 November 2020

Accepted 18 December 2020

Available online 4 January 2021

### Keywords:

Gene expression regulation

sRNAs

RACE

Transcriptomics

Transposons

Fungi

## ABSTRACT

Small RNAs (sRNAs) are key factors in the regulation of gene expression. Recently, a new class of regulatory sRNAs derived from tRNAs was described, the tRNA-derived RNA fragments (tRFs). Such RNAs range in length from 14 to 30 nucleotides and are produced from both mature and primary tRNA transcripts, with very specific cleavage sites along the tRNA sequence. Although several mechanisms have been proposed for how tRFs mediate regulation of gene expression, the exact mechanism of tRF biogenesis and its dependency upon the RNAi pathway remain unclear. *Cryptococcus gattii* and *Cryptococcus neoformans* are basidiomycetous yeasts and important human pathogens. While *C. neoformans* is RNAi proficient, *C. gattii* VGII has lost essential RNAi genes. Here, we sought to identify the tRF production profile in *C. gattii* VGII and *C. neoformans* in order to assess the RNAi-dependency of tRF production in these fungal species. We developed a RNA-sequencing-based tRF prediction workflow designed to improve the currently available prediction tools. Using this methodology, we were able to identify tRFs in both organisms. Despite the loss of the RNAi pathway, *C. gattii* VGII displayed a number of identified tRFs that did not differ significantly from those observed in *C. neoformans*. The analysis of predicted tRF targets revealed that a higher number of targets was found for *C. gattii* VGII tRFs compared to *C. neoformans* tRFs. These results support the idea that tRFs are at least partially independent of the canonical RNAi machinery, raising questions about possible compensatory roles of alternative regulatory RNAs in the absence of a functional RNAi pathway.

© 2020 British Mycological Society. Published by Elsevier Ltd. All rights reserved.

## 1. Introduction

Small RNAs (sRNAs) are a long-known class of non-coding RNAs smaller than 200 nucleotides in length (Torres-Martínez and Ruiz-Vázquez, 2017), with the most classical representatives being rRNAs and tRNAs. After the first reports of sRNAs acting on gene regulation in *Caenorhabditis elegans* (Feinbaum et al., 1993), sRNAs went from the position of an inert structural RNA molecule to one of the main players in gene expression regulation (Friedman et al., 2009). Although extensive research over the last decades has

elucidated much about the biogenesis and function of the most well-known regulatory sRNAs such as microRNAs, siRNAs, and piRNAs (Bartel, 2018; Carthew and Sontheimer, 2009; Weick and Miska, 2014), the advances in small RNA sequencing techniques and computational analysis, as well as the increase in the amount of data available, has led to the identification of novel regulatory sRNAs derived from classical RNAs such as tRNAs, rRNAs, snRNAs and snoRNAs (Li et al., 2012).

tRNA-derived small RNAs (tsRNAs) are a class of small RNAs with potential regulatory functions that have been identified in several organisms from all domains of life (Kumar et al., 2014). Although initial reports on tsRNAs date back to the 1970s (Phillip and Troll, 1977; Speer et al., 1979), they were believed to be only random degradation products until 2009 (Lee et al., 2009) and were often disregarded in sRNA sequencing studies (Cole et al., 2009). There is some evidence favoring the hypothesis that these fragments are

\* Corresponding author. Universidade Federal do Rio Grande do Sul, Centro de Biotecnologia, Avenida Bento Gonçalves, 9500 - Bloco IV, Prédio 43421, Laboratório 219, Caixa Postal 15005, Porto Alegre, RS, 91501-970, Brazil.

E-mail address: [staats@ufrgs.br](mailto:staats@ufrgs.br) (C.C. Staats).

indeed a type of regulatory sRNA: (i) tsRNAs undergo precise processing in specific regions within the tRNA rather than at random cleavage sites (Cole et al., 2009; Lee et al., 2009; Zheng et al., 2016), (ii) they are differentially expressed in different tissues and conditions (Li et al., 2018; Luo et al., 2018; Nunes et al., 2011), (iii) they can bind to proteins that are key factors in gene expression regulation (Goodarzi et al., 2015), and (iv) their presence or absence was shown to impact processes ranging from genomic surveillance against transposable elements to post-transcriptional regulation (Kuscu et al., 2018). There is still not a clear consensus on the nomenclature of these sRNAs, but tsRNAs may be classified into two different subsets based on their size. tRNA halves (thRNAs) are fragments produced by the cleavage of mature tRNAs on the anticodon mainly during stress conditions by angiogenin in humans and Rny1p in yeast (they are also known as tRNA-derived stress-induced RNAs, or tiRNAs), and they have lengths between 28 and 36 nucleotides. tRNA-derived RNA fragments (tRFs) are smaller fragments, with lengths ranging from 14 to 30 nucleotides (Li et al., 2018).

tRFs are produced from both mature and primary tRNA transcripts. There have been reports of many possible regions of origin for tRFs within the tRNA sequence (La Ferlita et al., 2019; Sun et al., 2018), which could arguably pose the hypothesis that they can be produced from virtually any region within the tRNA sequence. However, it is widely accepted that tRFs are produced mainly from the 5' region of the tRNA sequence (5' tRFs), the 3' region of the tRNA sequence containing a CCA extremity (3' tRFs), and the 3' trailer of the primary tRNA transcript (1' tRNA) (Zheng et al., 2016). Their biogenesis has not yet been entirely elucidated, and there is conflicting evidence on the participation of RNA interference (RNAi) components such as Dicer/Dcl (Kuscu et al., 2018; Megel et al., 2019). Due to their size and capability to bind Argonaute proteins (Ago) (Kumar et al., 2014), much of the work conducted on tRFs focuses on their action through the RNAi pathway. There is clear evidence of tRFs binding to different Ago proteins and post-transcriptionally regulating mRNAs in a microRNA-like fashion (Kuscu et al., 2018); however, tRFs are present in organisms that lack a functional RNAi pathway (Wang et al., 2019), and there are reports of alternative mechanisms of action for tRFs (Gebetsberger et al., 2012; Goodarzi et al., 2015; Keam et al., 2017; Kim et al., 2017; Mleczko et al., 2018).

The basidiomycetous yeasts *Cryptococcus gattii* and *Cryptococcus neoformans* are the main pathogenic agents of cryptococcosis, a systemic mycosis that is estimated to be the cause of 15% of AIDS-related deaths (Kwon-Chung et al., 2014; Rajasingham et al., 2017). The R265 strain of *C. gattii*, which is representative of the VGII molecular type of the *C. gattii* species complex (Cuomo et al., 2018), is characterized by the absence of the majority of genes essential to a functional RNAi pathway. Such genes include the RNA-dependent RNA polymerase (RdRp) and Ago, rendering it RNAi-deficient, while the closely related species *C. neoformans* and other *C. gattii* molecular types (VGI, VGIII, and VGIV) are RNAi proficient (Feretzaki et al., 2016). As there are evidences that tRFs are conserved between related species (Luo et al., 2018; Wang et al., 2016), this RNAi-loss scenario makes these yeasts invaluable models to assess the RNAi-dependency of tRFs. Here, we propose and apply a stringent *in silico* strategy employing small RNA-seq libraries to identify tRFs in *C. gattii* VGII and *C. neoformans*, aiming to evaluate the role of RNAi-dependence of tRFs in fungi by examining possible differences in the tRF production profiles of these organisms.

## 2. Methods

### 2.1. RNA-seq libraries

The Illumina *C. gattii* VGII R265 small RNA library was obtained from previous studies of our group (Ferrareze et al., 2017), which

was conducted by culturing *C. gattii* in YNB medium for 18 h. The Illumina *C. neoformans* H99 small RNA library was downloaded from the Sequence Read Archive under the accession number SRX215659, which was conducted by culturing *C. neoformans* in YPD. The H99 small RNA library was adapter trimmed using Cutadapt (version 1.18) (Martin, 2011), excluding reads that did not match the size selection applied to the library preparation (Dumesic et al., 2013). All the libraries were then filtered to remove low-quality reads with the fastq\_quality\_filter tool from FASTX-Toolkit (version 0.0.14) (Gordon and Hannon, 2010), filtering out reads with any base that scored below Q30.

The Illumina *C. gattii* R265 RNA libraries were generated by our group (Diehl, unpublished results), consisting of 3 replicates paired-end sequencing of *C. gattii* R265 grown in YNB. The Illumina *C. neoformans* H99 RNA libraries were downloaded from the Sequence Read Archive under the accession numbers SRX6799215, SRX6799216 and SRX6799217, which represents growth in YPD. The H99 RNA libraries were processed using Cutadapt (version 1.18), removing adapter sequences, 3' quality trimming reads with a Q28 cutoff, and removing reads smaller than 40 nucleotides. These libraries were used for mRNA target expression evaluation.

### 2.2. tRF prediction workflow

Initially, we predicted the tRNA genes for *C. gattii* VGII R265 and *C. neoformans* H99 from the genomic sequences, using the assemblies *C. deuteroformans*\_R265\_chr (GenBank accession GCA\_002954075.1) and CNA3 (GenBank accession GCA\_000149245.3), respectively. The prediction was performed using tRNAscan-SE (version 2.0) (Chan et al., 2019) with default parameters. All the tRNA predictions were accepted, including predicted pseudogenes, as they could still be transcriptionally active. We then generated the mature and primary tRNA sequences for each organism using Bedtools (version 2.25.0) (Quinlan and Hall, 2010). For mature tRNA sequences, we used the getfasta function of Bedtools to extract the FASTA sequence from the genome (using the -s option) using the tRNA coordinates file from the tRNAscan-SE output. In order to remove the introns from the sequence, the -split option was used. After the tRNA sequence extraction, the 3' CCA was added when not coded by the gene, and for histidine tRNAs, a 5' G was added, both manually (Cooley et al., 1982). For primary tRNA sequences, a new coordinate file was created manually in which the coordinates from the beginning and ending of the tRNA were altered by -40 and 40, respectively, in an attempt to include the 5' leader as well as the 3' trailer sequences, and intron coordinates were not included. We then used Bedtools to extract the FASTA sequence from the primary tRNAs using these new files, using the -s option but not the -split option. Finally, redundancy was removed by selecting only one of the redundant tRNAs as a representative, generating a unique tRNA set. For the tRF prediction with tDRmapper, the anticodon coordinate was placed in the header of the FASTA sequence of mature tRNAs, and the mature and primary tRNAs were placed in the same file. A local bidirectional best-hit BLASTn (version 2.6.0) (Altschul et al., 1990) analysis was done with mature and primary tRNA sequences separately to identify correspondent sequences in the *C. neoformans* and *C. gattii* VGII genomes.

In order to filter our small RNA-seq libraries from reads with dubious origins, we generated a tRNA-free genome by removing the mature or primary tRNA sequences from the genomic sequence with Bedtools, using the coordinates from the tRNA prediction. We then aligned the small RNA libraries against the tRNA-free genome with Bowtie2 (version 2.3.4.1) (Langmead and Salzberg, 2012) using the -end-to-end and -very-sensitive parameters, and we generated new libraries with Samtools (version 1.9) (Li et al., 2009)

and Seqkit (version 0.10.0) (Shen et al., 2016) composed only of reads that failed to align or aligned with some mismatch or gap in the tRNA-free genome sequences. The new libraries were aligned against the mature or primary tRNA sequences using Bowtie2, and new libraries were generated by Samtools and Seqkit containing only reads that aligned perfectly to tRNAs. The libraries from the mature and primary sequence alignments were then merged, and the new libraries composed from reads that perfectly aligned to tRNAs but not to non-tRNA locations were used for tRF prediction.

The tRF prediction was performed employing tDRmapper (Selitsky and Sethupathy, 2015), using the tRNA-only aligned libraries and the tDRmapper tRNA sequence file. tRFs were defined as the location of the majority of the coverage within the tRF prediction site (locMajorityCov) from the software output and classified based on their origin within the tRNA. The following classification used was: (i) 5' leader for tRFs produced from any region within the primary tRNA transcript that included the 5' leader, (ii) 5' terminus for tRFs that contained the 5' extremity of the mature tRNA, (iii) internal for tRFs produced from any region of the mature tRNA that did not contain the 5' extremity or 3' CCA extremity, (iv) intronic for tRFs produced from any region within the primary tRNA transcript that included the intron sequence, (v) 3' terminus for tRFs that contained the 3' CCA extremity of the mature tRNA, and (vi) 3' trailer for tRFs produced from any region within the primary tRNA transcript that included the 3' trailer. Redundancy was removed by combining equal tRFs predicted from regions conserved among tRNA isoforms. tRFs with a length smaller than the minimal read length in the library, tRFs that presented gaps in comparison with the parental tRNA sequence, and tRFs with ambiguous origin (e.g., contained two different regions from two different tRNAs) were excluded. The transcripts per million (TPM) value for each tRF was calculated as follows:

$$\text{TPM} = (\text{tRF read count/clean library size}) \times 10^6$$

where the tRF read count was the number of reads for a given tRF with complete correspondence to the tRF in the quality filtered library (in both length and sequence) and the clean library size was the total number of reads within the quality filtered library. A 10 TPM cutoff was applied to the predicted tRFs, and tRFs with lower TPMs were removed from the prediction.

### 2.3. RT-PCR

In order to confirm the data generated in our workflow, some tRFs from both cryptococcal strains were evaluated using a stem-loop RT-PCR analysis originally described to detect microRNAs (Varkonyi-Gasic et al., 2007). *C. neoformans* H99 and *C. gattii* R265 strains were grown in YPD for 18 h. An aliquot of 1 mL of the culture was washed with ultrapure water and lyophilized. RNA was isolated using Trizol (Thermo Fisher), after cell lysis by bead beating (6 cycles of 30 s at full vortex speed and 1 min in ice). The first strand synthesis was conducted using each tRF specific stem-loop primer using 250 ng of RNA and 1  $\mu$ L of ImProm-II Reverse Transcriptase (Promega), using the parameters already described (Varkonyi-Gasic et al., 2007). For the qPCR analysis, an aliquot of 2  $\mu$ L of each reverse transcription reaction was mixed with 5 pmol of Universal reverse primer and 5 pmol of each tRF specific forward primer in a total volume of 5  $\mu$ L, followed by the addition of an equal volume of 2 X Power SYBR Green Master Mix (ThermoFisher). The cycling was carried out according the original protocol (Varkonyi-Gasic et al., 2007) in a 7500 Fast Real-Time PCR System (ThermoFisher), followed by a melt curve analysis to determine the  $T_m$  of each amplicon. The primers sequences (Supplementary Table 1) were

designed using the miRNA Primer Design Tool (Czimmerer et al., 2013).

### 2.4. Target prediction

To predict mRNA targets for tRFs, we employed two software: psRNAtarget (Dai et al., 2018) and TAPIR (Bonnet et al., 2010). psRNAtarget was executed from its online platform. Each tRF was analyzed individually using either the R265 transcription annotation of our group (Ferrareze et al., 2020) or the H99 annotation from FungiDB Release 43 (Basenko et al., 2018), with the following parameter alterations: (i) number of top targets set to 1000, (ii) HSP size as the length of the analyzed tRF (or the closest the software accepted), and (iii) enabling the calculation of target accessibility. TAPIR was executed locally with default parameters for all tRFs. Only target sites predicted by both methods were accepted; these were determined by an automatized comparison of the first position in the mRNA target region between the two programs, with an allowed three-nucleotide difference in the predicted beginning of the target site due to minor differences in the handling of the alignment and report between the two programs. The mRNA targets of each tRF were evaluated for Gene Ontology (GO) enrichment in biological processes in the FungiDB website, using a cutoff of 0.05 in a Bonferroni test and a minimal of two mRNA targets.

We evaluated target expression for each tRF target with RNA-seq libraries from *C. gattii* R265 and *C. neoformans* H99. For this step, we used the same R265 genome assembly and annotation as before, and the H99 FungiDB's genome assembly and annotation (release 43), with both annotations in GFF format. First, we aligned our libraries to their respective genomes using STAR (version 2.7.3a). For STAR genome indexing, we provided the annotation GFF file, and set genomeSAindexNbases to 11 and sjdbOverhang to 75 for *C. gattii* and 100 for *C. neoformans*. For STAR alignment, we set outFilterType to BySJout, alignIntronMin to 10, alignIntronMax 3000 and outFilterIntronMotifs to RemoveNoncanonical. We then converted the sam alignment file to bam and sorted it using Samtools (version 1.9), and estimated the transcriptional expression using Stringtie (version 2.1.2) with -e option, setting the -rf flag for R265. The target TPM was defined as the mean TPM value calculated by Stringtie for transcript isoforms of a given gene in all replicates, although multiple transcript isoforms are present only in H99 annotation. For tRF-target expression comparison, the tRF TPM was the TPM from the tRF prediction step.

Transposition element targets for tRFs were predicted with psRNAtarget only. The analysis was executed as for mRNAs, but an expectation value of three instead of five was used in order to be more stringent, as the prediction was based entirely on this tool. The transposition elements for both *C. gattii* VGII and *C. neoformans* were predicted using TransposonPSI, and their sequences were extracted from the best hit per loci gff3 file (Supplementary files 1 and 2) using Bedtools.

### 2.5. 5'-RACE

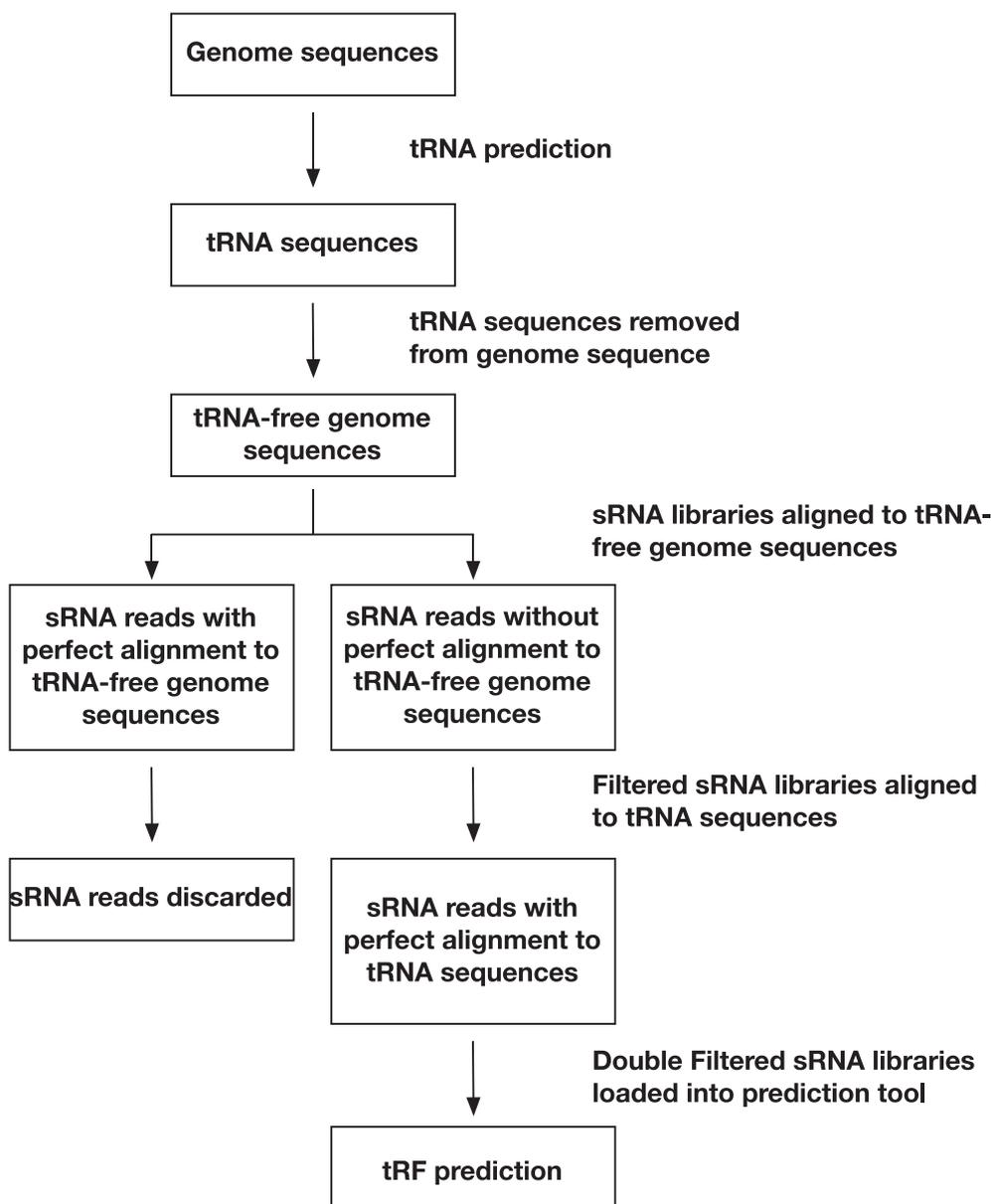
The 5' Rapid amplification of cDNA ends (RACE) was performed to detect potential cleavage activity driven by tRFs. A total of 500 ng of DNase-treated total RNA samples was used for cDNA synthesis using the SMART™ RACE cDNA amplification kit (Clontech). Confirmation of mRNA cleavage was evaluated using PCR amplification of cDNA with a gene specific primer and a primer that hybridize in the template shift primer (Supplementary Table 1). Amplification reactions were analyzed by electrophoresis in 1.5% TAE electrophoretic gels.

### 3. Results

#### 3.1. tRF prediction

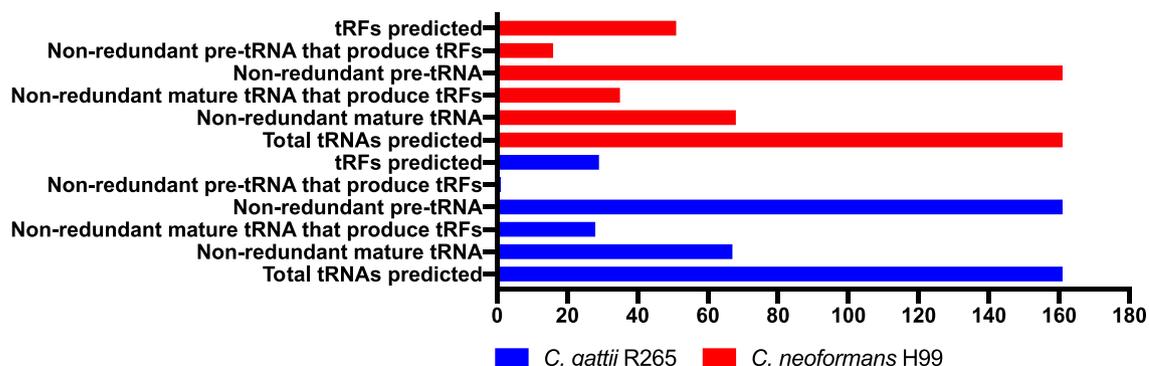
Using our proposed tRF prediction workflow (Fig. 1), we were able to predict tRFs in both *C. gattii* R265 and *C. neoformans* H99 (Supplementary Table 2), identifying 29 and 51 predicted tRFs, respectively (Fig. 2A). The majority of these tRFs were derived from mature tRNA (Fig. 2A). The identification of a comparable number of tRFs in *C. gattii* VGII and *C. neoformans* suggests that tRFs do not entirely depend on RNAi machinery to be produced or stabilized. In line with this, when comparing the parental tRNA from R265 and H99 tRFs, we observed that only 15 corresponding tRNA pairs produced tRFs in both organisms (Supplementary Table 3), and even in these cases, only 4 of the tRNA pairs produced tRFs from the

same region. These differences were not restricted to the tRNA isoacceptors that underwent processing, but they were observed in the processing itself. While the majority of the tRFs from *C. neoformans* originated from the 5' extremity of tRNAs, most of tRFs predicted in *C. gattii* VGII were from internal regions of the tRNA. Additionally, only one tRF arose from *C. gattii* VGII primary tRNA transcripts, despite the considerable abundance of this class in *C. neoformans* (Fig. 2B). Also, tRFs from R265 were preferentially 15, 17, and 18 nts in length; smaller than tRFs from *C. neoformans* that were mainly processed in 19, 22, and 24 nt fragments (Fig. 2C). Although the difference in size preference between these organisms could be due to the difference in size range of the libraries, as the smallest sRNAs in *C. neoformans* libraries are 18 nts, it is possible to assume that the tRF production profile of these organisms could be different.

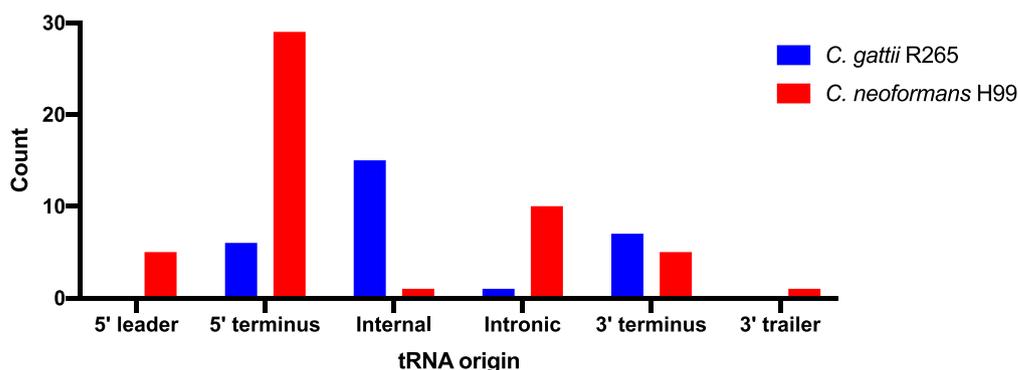


**Fig. 1. tRF prediction workflow.** tRNA genes were predicted with the genomic sequence of the organisms. Using the coordinates from the predicted tRNA genes, the tRNA sequences were removed from the genome. The sRNA read library was aligned against the tRNA-free genome, and reads that aligned perfectly to this sequence were removed from the library, generating a non-tRNA-free library. The new library was aligned against the predicted tRNA sequences, and reads that did not perfectly align against the predicted tRNA sequences were removed from the library, generating a tRNA-exclusive library that was loaded into a tRF prediction tool.

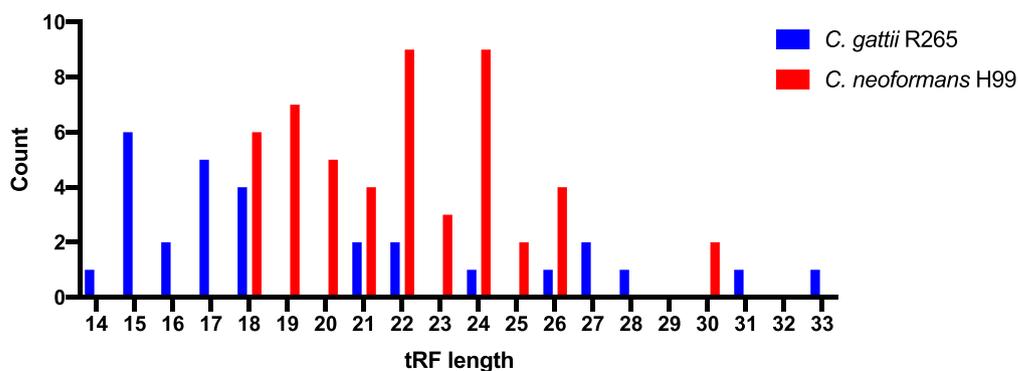
A



B



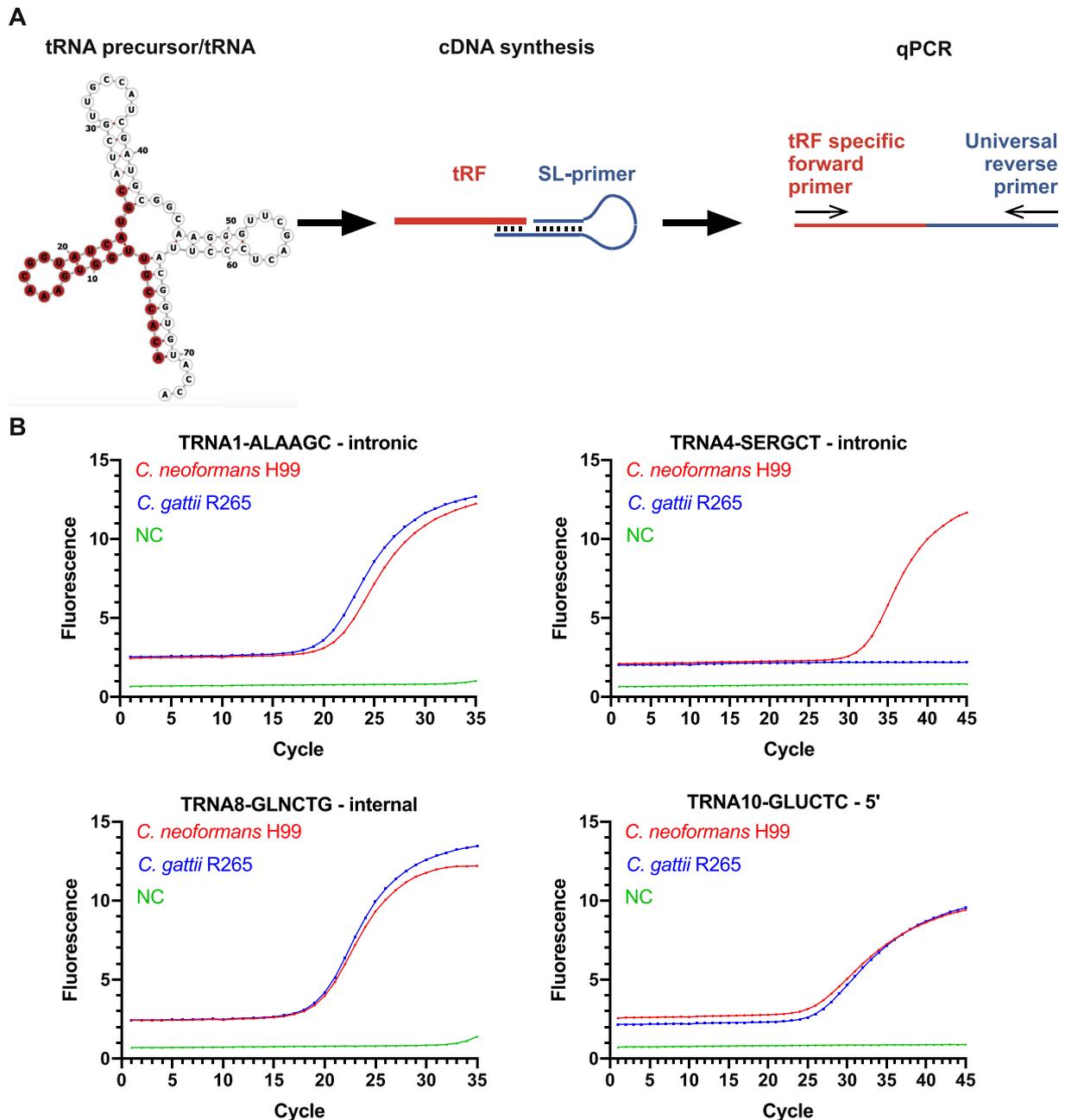
C



**Fig. 2. tRF production profile in *Cryptococcus* spp.** (A) Comparison of the counts of predicted tRNA, tRNA that produces tRFs and tRFs in *C. gattii* R265 and *C. neoformans* H99. (B) Location of predicted tRFs within tRNA sequences for both *C. gattii* R265 and *C. neoformans* H99. (C) Length (in nt) of predicted tRFs in *C. gattii* R265 and *C. neoformans* H99.

In order to evaluate the validity of our workflow, we employed a quantitative RT-PCR to detect small RNAs, based on the principle of stem-loop reverse transcription coupled to PCR (Varkonyi-Gasic et al., 2007). A potential tRF is hybridized, by its last nucleotides, to a tRF specific stem-loop primer, which is associated with the first strand synthesis. Then, a universal reverse primer and a tRF specific forward primer are used to amplify the chimeric molecule for further detection (Fig. 3A). Using this approach, we select four distinct tRFs originating from distinct sections of the tRNA or pre-tRNA molecule (intron, internal or 5') and detected in the *C. neoformans* (pre-CP003827.1.TRNA4-SERGCT,

CP003820.1.TRNA10-GLUGTC) or *C. gattii* (pre-CP025759.1.TRNA1-ALAAGC, CP025765.1.TRNA8-GLNCTG). We evaluated the expression of such tRFs in either *C. neoformans* and *C. gattii* grown in YPD and found that the internal and 5' tRFs (originating from CP025765.1.TRNA8-GLNCTG and CP003820.1.TRNA10-GLUGTC, respectively), displayed almost the same transcriptional in both organisms (Fig. 3B). However, for the intronic tRF derived from pre-CP003827.1.TRNA4-SERGCT, found in the *C. neoformans* libraries, we could only observe amplification from *C. neoformans* RNA (Fig. 3B). In the opposite direction, the intronic tRF derived from pre-CP025759.1.TRNA1-ALAAGC, found in *C. gattii*, displayed increased



**Fig. 3.** Detection of the predicted tRFs by RT-qPCR. (A) Illustration of the stem-loop method to detect small RNAs. A tRF (red) derived from a hypothetical tRNA is hybridized to a tRF-specific stem-loop primer, followed by reverse transcription and consequent amplification by a universal reverse primer and a tRF specific forward primer. (B) Amplification profile of four distinct tRFs from distinct tRNAs with distinct origins. (For interpretation of the references to color/colour in this figure legend, the reader is referred to the Web version of this article.)

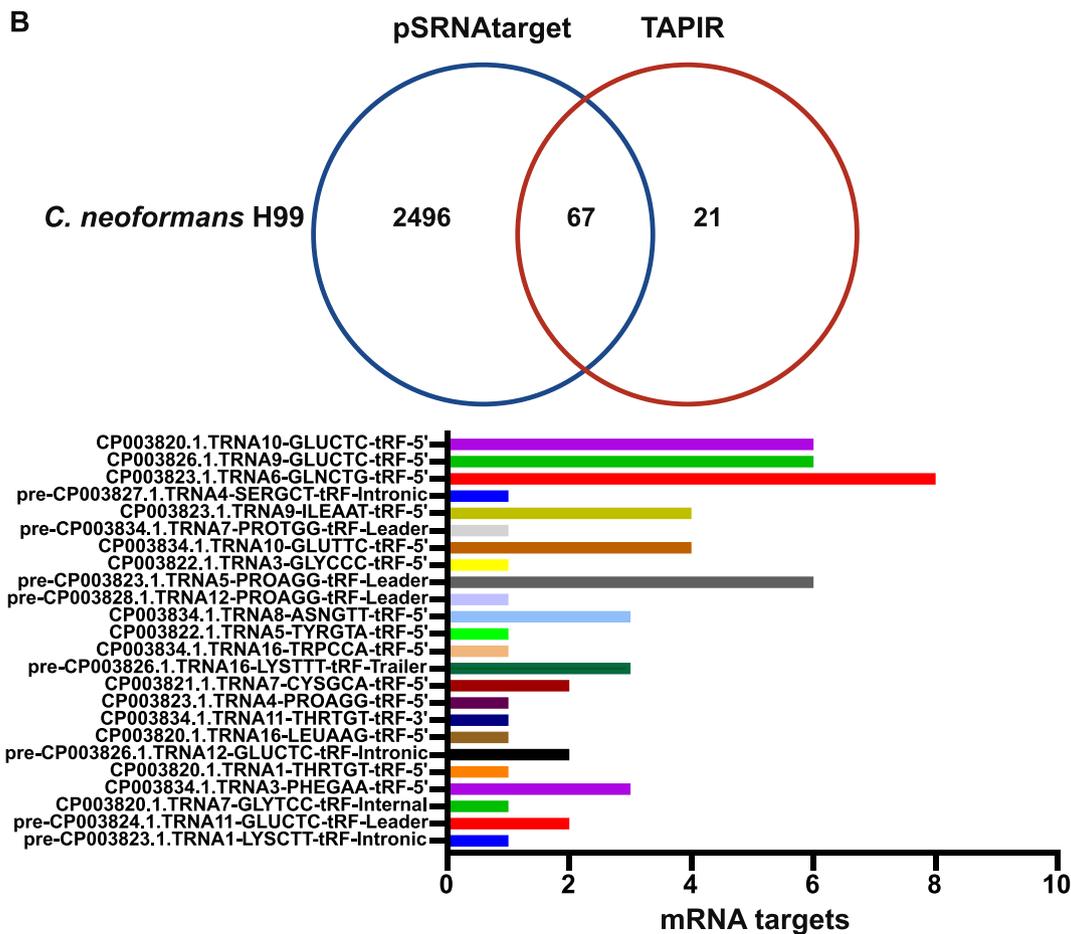
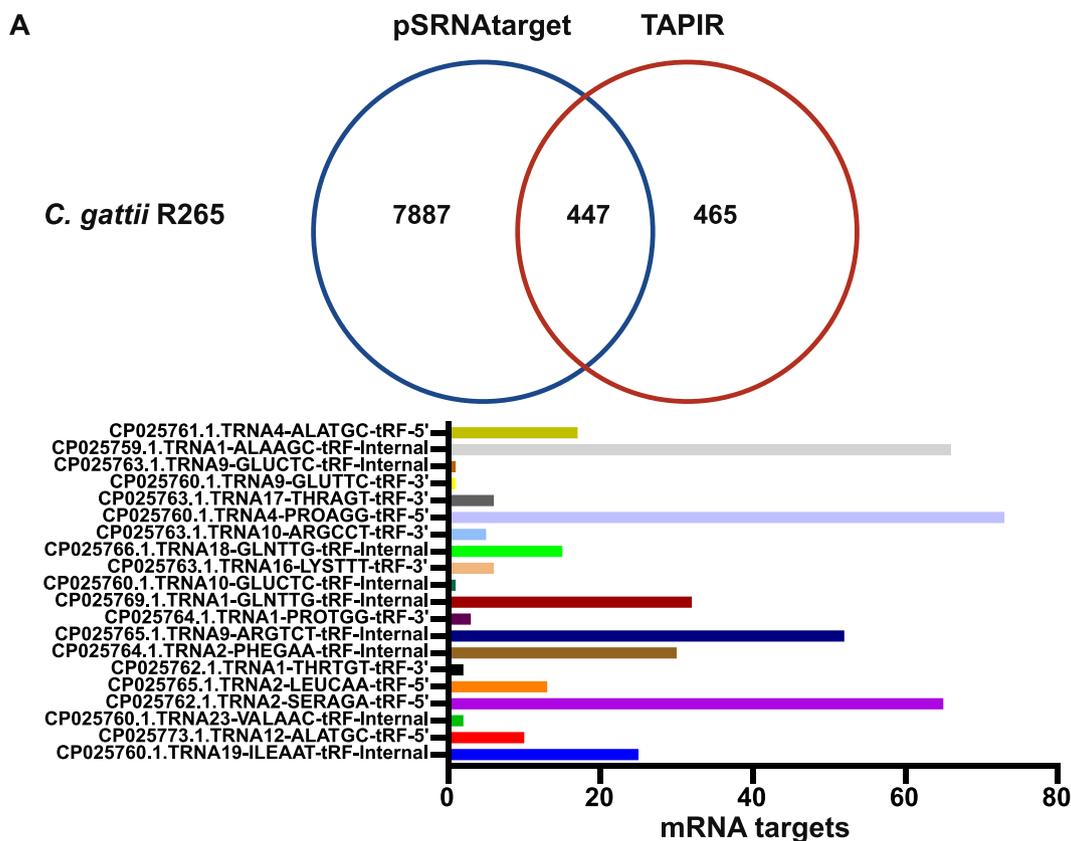
transcript levels compared to those in *C. neoformans*, as detected by a near 1 cycle in the quantification fluorescence threshold (Fig. 3B).

### 3.2. tRF target prediction

In order to assess possible functions of the predicted tRFs, we evaluated their capability to target both endogenous mRNAs and transposable elements. Using two independent methods to evaluate potential targets (pSRNATarget and TAPIR), we found contrasting values for the prediction of each tRF-mRNA pair. Hence, we use the intersect of both methods to infer the total number of mRNA targets of the predicted tRFs (Fig. 4 and Supplementary

Table 4). Both organisms had relatively close numbers of tRFs targeting mRNAs, with 24 for *C. neoformans* H99 and 20 for *C. gattii* R265. Moreover, we found that tRFs in *C. gattii* can display a high range of targets (up to 73), while this number is drastic lower in *C. neoformans* (up to 8 targets – Fig. 4). We were unable to find any enriched biological process associated with the mRNAs potentially targets by such tRFs in both organisms.

In order to evaluate whether this increased number of RNA targets in *C. gattii* R265 could reflect a potential mechanism of gene regulation independent of the activity of the RNAi pathway, we determined the mRNA levels of the putative tRF targets based on public available RNA-Seq datasets. We found that the expression



levels of tRFs from the two species has a range of near 2 orders of magnitude in *C. gattii* and 3 orders of magnitude, in *C. neoformans* (Fig. 5A and B). The expression of putative targets in both species also displayed a high degree of variance (Supplementary Table 4). We could not detect a correlation between expression of tRFs and the predicted targets ( $R^2 = 0.0046$  and  $R^2 = 0.0044$  for *C. gattii* and *C. neoformans* data, respectively). We then analyzed each tRF–target pair by evaluating the ratio between the mRNA and tRF transcript levels (Fig. 5A and B) and stratified such data into three major arbitrary categories according to the ratio between the TPM of such pair: high ratio (HR – ratio  $\geq 1$ ), intermediate ratio (IR –  $10 > \text{ratio} \geq 0.001$ , and low ratio (LR – ratio  $< 0.001$ ). Chi-squared analysis revealed that there is a significant ( $\chi^2 = 24.663$ ,  $df = 2$ ,  $p\text{-value} < 0.001$ ) positive association between low ratio between predict targets and tRFs expression ratio in *C. neoformans* (Fig. 5C), but not in *C. gattii* nor in others ranges of ratios analyzed.

The analysis of tRFs potential to target retrotransposons revealed that TAPIR was too stringent, as no targets could be found. Therefore, we employed only psRNAtarget, but with a higher score cutoff to partially compensate the lack of comparison with other programs. We found that while one tRF was predicted to target 7 transposable elements in *C. gattii* VGII, no tRF was observed to target transposable elements in *C. neoformans* (Supplementary Table 5). This observation further highlights the differences in targeting potential observed between the tRFs from these organisms.

To evaluate if tRFs may be involved in regulation of gene expression dependent upon the activity of RNAi pathway in cryptococcal species, we performed 5' RACE reactions using RNA isolated from both the RNAi proficient *C. neoformans* H99 and the RNAi deficient *C. gattii* R265. We selected a target of tRF TRNA10-GLUCTC, which was the most abundant tRF identified in the RNA-seq libraries of *C. neoformans* H99 (Supplementary Table 4), and could also be detected in *C. gattii* (Fig. 3B). Based on the predicted hybridization site of this tRFs in the CNAG\_05202 mRNA sequence (*C. neoformans*), we designed primers that would amplify a near 700 bp cDNA fragment derived from the gene specific primer (reverse primer) and the primer that hybridize in the SMART template shift sequence (forward primer – Fig. 6A). The same set of primers were also used in the analysis of the ortholog of CNAG\_05202 in *C. gattii* (CNBG\_9479). We could only detect amplification of potential cleavage mRNA-derived cDNAs in samples collected from *C. neoformans*, but not in *C. gattii* (Fig. 6B). This suggest that cryptococcal cells can exploit tRFs to regulate gene expression in a RNAi pathway-dependent fashion.

#### 4. Discussion

Research on tRFs is hampered by the lack of *in silico* strategies to identify them. Although there are several studies on tRFs, many rely solely on a nearly manual analysis of alignments to identify possible tRFs (Hirose et al., 2015; Kumar et al., 2014; Wang et al., 2013, 2019), a practice that could lead to inconsistencies due to its subjectivity. Also, with the increasing ease of generating and obtaining high-throughput data from RNA sequencing, automated solutions for massive tRF prediction could circumvent the difficulties of manual prediction of tRFs in multiple samples. Although the tDRmapper software can handle data from different organisms with reasonable ease, one of its main drawbacks is that it handles the alignments naively, aligning the reads directly to the tRNA without taking into account the possibility of those reads being originated from other regions of the genome (Xu et al., 2017).

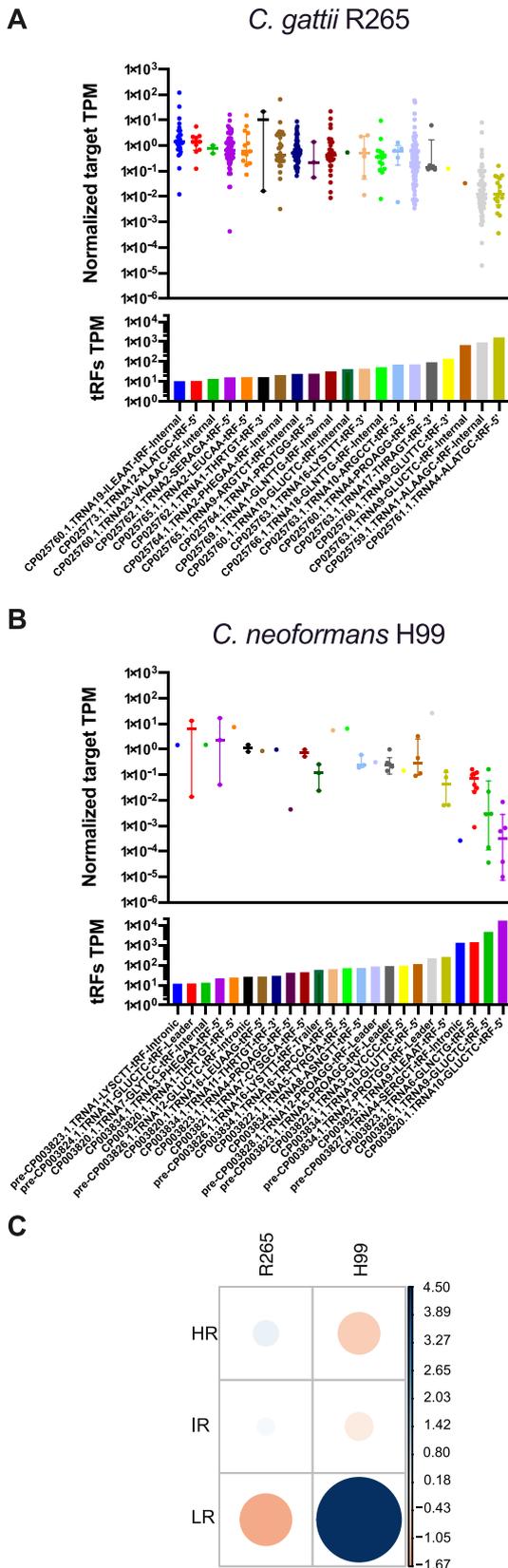
In an effort to increase the accuracy of tDRmapper, we designed a workflow (Fig. 1) for filtering the input reads of tDRmapper, which would reduce the false positive prediction rate. Although our stringent workflow may not fully represent the entire tRF composition, as reads that would align equally to tRNA and non-tRNA genomic regions could origin from within tRNAs, and modified bases could be misinterpreted as mismatches (Ryvkin et al., 2013), this stringency is indispensable as this is a poorly characterized and highly variable sRNA class. We also chose to use all of the predicted tRNAs for our analysis, including those marked as possible pseudogenes, because tRNA-derived sequences could be a source of sRNAs (Surbanovski et al., 2016).

The relatively recent discovery of tRFs is not only reflected in the scarcity of available tools to study them but also in the conflicting data about their biogenesis and function. One of these inconsistent points is the relationship between tRFs and the RNAi pathway. Despite their microRNA-like characteristics and the evidence of some tRFs depending on RNAi machinery for their biogenesis or function (Cole et al., 2009; Kuscus et al., 2018), other evidence for RNAi independency and alternative mechanisms of action and reports of tRFs in organisms that lack the canonical RNAi pathway (Mleczo et al., 2018; Wang et al., 2019) questions if the observed RNAi dependency of some tRFs is an exception rather than the rule.

To address this question, we applied our tRF prediction strategy on *C. gattii* VGII and *C. neoformans* small RNA-seq data and compared their tRF production profiles. As about 83% of tRFs are shared between the phylogenetic close organisms *Drosophila virilis* and *D. melanogaster* (Luo et al., 2018), it is expected that few changes in the tRF profile would have occurred between *C. gattii* VGII and *C. neoformans*, unless the change is related to the absence or presence of a functional RNAi pathway even when the data is derived from a complex culture medium (YPD) for *C. neoformans* and from a defined culture medium (YNB) for *C. gattii*. We observed that despite the close relationship, *C. gattii* VGII displayed a preference towards shorter tRFs and production of tRFs from internal regions of the tRNA, while *C. neoformans* tRFs were usually longer and biased towards the 5' terminus of the tRNA. There was also a partial shift in the set of tRNAs processed into tRFs. Furthermore, the differences extended to the targeting potential of the tRFs, with the *C. gattii* VGII tRFs being theoretically able to target a considerably increased number of self-mRNAs when compared to *C. neoformans* tRFs and also showed potential to target transposable elements, which was not observed for any *C. neoformans* tRF. These differences could be related not only to differences in the tRF sequences but also to the shorter tRFs in *C. gattii* VGII. We should consider that despite our *in silico* prediction pointed towards a very distinct tRF set production between *C. gattii* and *C. neoformans*, we observed in our confirmatory RT-qPCR experiments that most of our evaluated tRFs were actually present in both organisms. Although the extension of this phenomena among the predicted tRFs is unknown, this inconsistency may be explained at least partially by the differences in the growth media that was employed in each RNA-seq experiments, as it was shown that tRF production is sensible to growth condition (Bąkowska-Żywicka et al., 2016). Despite the presence of some unpredicted tRFs, the identification of all tested tRFs in the corresponding organism confirms our prediction.

Our results allowed us to hypothesize at least two non-exclusive scenarios: first, the tRF profile shift in *C. gattii* VGII is due to the loss of RNAi machinery that acted upon tRFs during their biogenesis, stabilization, or function, which abolished the former tRF set and

**Fig. 4. tRF target prediction.** tRF target prediction against self-transcripts for (A) *C. gattii* R265 and (B) *C. neoformans* H99 showing the numbers of all mRNAs targeted by tRFs in both predictors (top) and the number of targets by each tRF pairs (bottom).



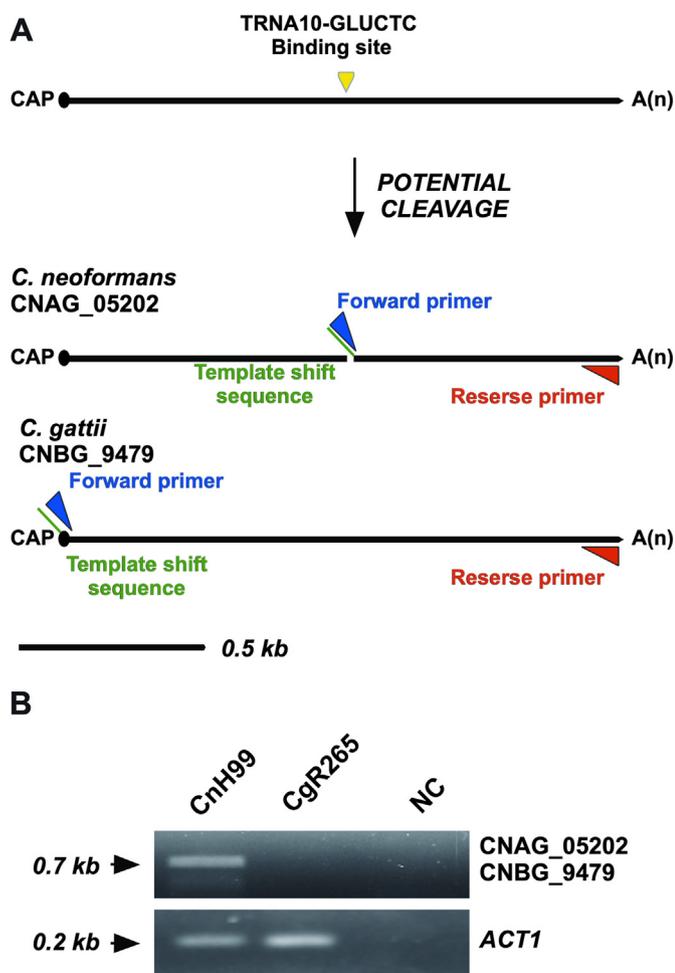
**Fig. 5. Profile of tRF and predicted targets mRNA levels.** The ratio between each predicted target mRNA to the corresponding tRF is displayed as along with the median and interquartile range (upper panel), as well as the tRF RNA levels (bottom panel) for *C. gattii* R265 (A) and *C. neoformans* H99 (B). All the individual ratios of *C. gattii* and *C. neoformans* were divided into three distinct categories: high ratio (HR, where ratio  $\geq 1$ ), intermediate ratio (IR, where  $1 > \text{ratio} \geq 0.001$ ), and low ratio (LR, where

led to its replacement by an RNAi-independent one; second, the loss of a functional RNAi pathway in *C. gattii* VGII led to the adoption of a novel tRF set that not only is RNAi-independent but also may compensate for the lack of RNAi regulation and thus render the organism viable in its absence. Consistent with the first scenario, the tRF with the highest expression in *C. neoformans* (corresponding to almost 2% of the entire sRNAseq library, TPM = 16,861) (Supplementary Table 1), had a size of 18 nt and a U at its 5' terminus, making it very similar to microRNAs and therefore likely to act through the RNAi pathway (Mi et al., 2008). Moreover, this tRF potentially act in the cleavage of its target RNAs in *C. neoformans* H99, but not in *C. gattii* R265. Although these characteristics were not shared among all *C. neoformans* tRFs, the 5' terminal preference in tRF biogenesis, which is compatible with Dicer-1-like processing (Wilson and Doudna, 2013), and the fact that many of the tRFs had a size compatible with RNAi machinery reinforces the hypothesis that H99 tRFs are likely associated with the RNAi pathway. We also observed that the absence of functional RNAi proteins that may produce and bind tRFs in RNAi-proficient organisms did not abolish the production of tRFs in *C. gattii* VGII. Rather, an almost entirely different set of tRFs was found: compared to *C. neoformans* tRFs, tRFs from *C. gattii* VGII had a different size and originated from different parts of the tRNA sequences, and they also arose from different tRNAs. These differences could indicate the RNAi independency of tRFs in *C. gattii* VGII.

In accordance with our second hypothesized scenario, we observed that even though *C. neoformans* presented a wider variety of tRFs as well as more tRFs predicted to target mRNAs, the average number of mRNA targets per tRF was considerably higher for the tRFs from *C. gattii* VGII. This possible gain in regulatory capabilities of *C. gattii* VGII tRFs may be due to the adoption of a new tRF set that could at least partially compensate for the loss of RNAi regulation possibly by inhibition of protein synthesis instead of mRNA degradation. In line with this assumption, we found that the application of a simple mathematical transformation, which is the ratio between the mRNA levels of the putative targets and the tRFs levels, followed by stratification in distinct ranges of ratios, allow us to find a statistically significant fraction of the genes that displayed a positive association in *C. neoformans* only in the low ratio range (ratio  $< 0.001$ ), as revealed by a chi-squared test. This suggests that tRFs could also be used for RNAi pathway in *C. neoformans* H99 by active degradation of their mRNA targets, as already demonstrated for intron-derived siRNAs in the SCANR model (Dumesic et al., 2013). In addition, tRFs in the budding yeast *Saccharomyces cerevisiae*, which also lacks a RNAi pathway (Drinnenberg et al., 2009), are known to be involved in inhibition of protein synthesis (Bakowska-Zywicka et al., 2016; Mleczko et al., 2018).

In conclusion, we were able to demonstrate the presence and possible functionality of tRFs in the absence of a functional RNAi pathway. Further studies on tRF functions may reveal novel mechanisms behind this RNAi-independent action and could show ancestrally conserved mechanisms, as tRFs are present in organisms evolutionarily earlier than the RNAi pathway. Much is yet to be elucidated about tRF biogenesis and function, and although the high variation in several aspects of tRFs hinders their study, they may eventually be understood as a thrilling new world of sRNA regulation.

ratio  $< 0.001$ ). A Pearson's Chi-squared test were performed and residuals plotted. Positive significant association between a cryptococcal strain and the corresponding ratio ranges is depicted as blue circles with Chi-square residuals higher than 2 (C). (For interpretation of the references to color/colour in this figure legend, the reader is referred to the Web version of this article.)



**Fig. 6.** Analysis of potential tRF functionality by 5' RACE. (A) Schematic representation of TRNA10-GLUCTC mRNA targets. The binding sites of tRF, as well as the reverse (red) and forward (blue) primers are shown. (B) Agarose gel image of 5' RACE reaction to detect cleavage of *C. neoformans* CNAG\_05202 and *C. gattii* CNBG\_9479 transcripts. Amplicon sizes were inferred by comparison with Quick-load Purple 1 kb Plus DNA Ladder (NEB) molecular size marker. Similar RNA loads were confirmed by amplification of a segment of *ACT1* transcript. (For interpretation of the references to color/colour in this figure legend, the reader is referred to the Web version of this article.)

### Declaration of competing interest

The authors declare no conflicts of interest. The authors are solely responsible for the content and the writing of the manuscript.

### Acknowledgments

This project was funded by the Advanced Network of Computational Biology (RABICÓ - Biocomputacional Grant 23038.010041/2013–13) from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). CCS was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) grants (404141/2016–1 and 309897/2017–3).

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.funbio.2020.12.003>.

### References

- 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. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Bakowska-Zywicka, K., Kasprzyk, M., Twardowski, T., 2016. TRNA-derived short RNAs bind to *Saccharomyces cerevisiae* ribosomes in a stress-dependent manner and inhibit protein synthesis in vitro. *FEMS Yeast Res.* <https://doi.org/10.1093/femsyr/fow077>.
- Bakowska-Zywicka, K., Mleczko, A.M., Kasprzyk, M., Machtel, P., Zywicki, M., Twardowski, T., 2016. The widespread occurrence of tRNA-derived fragments in *Saccharomyces cerevisiae*. *FEBS Open Bio* 6, 1186–1200. <https://doi.org/10.1002/2211-5463.12127>.
- Bartel, D.P., 2018. Metazoan MicroRNAs. *Cell* 173, 20–51. <https://doi.org/10.1016/j.cell.2018.03.006>.
- Bonnet, E., He, Y., Billiau, K., van de Peer, Y., 2010. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics* 26, 1566–1568. <https://doi.org/10.1093/bioinformatics/btq233>.
- Carthew, R.W., Sontheimer, E.J., 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell* 136, 642–655. <https://doi.org/10.1016/j.cell.2009.01.035>.
- Chan, P.P., Lin, B.Y., Mak, A.J., Lowe, T.M., 2019. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *bioRxiv* 614032. <https://doi.org/10.1101/614032>.
- Cole, C., Sobala, A., Lu, C., Thatcher, S.R., Bowman, A., Brown, J.W.S., Green, P.J., Barton, G.J., Hutvagner, G., 2009. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* 15, 2147–2160. <https://doi.org/10.1261/rna.1738409>.
- Cooley, L., Appel, B., Soll, D., 1982. Post-transcriptional nucleotide addition is responsible for the formation of the 5' terminus of histidine tRNA. *Proc. Natl. Acad. Sci. U.S.A.* 79, 6475–6479. <https://doi.org/10.1073/pnas.79.21.6475>.
- Cuomo, C.A., Rhodes, J., Desjardins, C.A., 2018. Advances in *Cryptococcus* genomics: insights into the evolution of pathogenesis. *Mem. Inst. Oswaldo Cruz* 113. <https://doi.org/10.1590/0074-02760170473>.
- Czimmerer, Z., Hulvely, J., Simandi, Z., Varallyay, E., Havelda, Z., Szabo, E., Varga, A., Dezso, B., Balogh, M., Horvath, A., Domokos, B., Torok, Z., Nagy, L., Balint, B.L., 2013. A versatile method to Design stem-loop primer-based quantitative PCR assays for detecting small regulatory RNA molecules. *PLoS One*. <https://doi.org/10.1371/journal.pone.0055168>.
- Dai, X., Zhuang, Z., Zhao, P.X., 2018. PsRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* 46, W49–W54. <https://doi.org/10.1093/nar/gky316>.
- Drinnenberg, I.A., Weinberg, D.E., Xie, K.T., Mower, J.P., Wolfe, K.H., Fink, G.R., Bartel, D.P., 2009. RNAi in budding yeast. *Science* (80-. ) 326, 544–550. <https://doi.org/10.1126/science.1176945>.
- Dumesic, P.A., Natarajan, P., Chen, C., Drinnenberg, I.A., Schiller, B.J., Thompson, J., Moresco, J.J., Yates, J.R., Bartel, D.P., Madhani, H.D., 2013. Stalled spliceosomes are a signal for RNAi-mediated genome defense. *Cell* 152, 957–968. <https://doi.org/10.1016/j.cell.2013.01.046>.
- Feinbaum, R.L., Ambros, V., Lee, R.C., Feinbaum, R.L., Ambros, V., 1993. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843–854. [https://doi.org/10.1016/0092-8674\(93\)90529-Y](https://doi.org/10.1016/0092-8674(93)90529-Y).
- Feretzkaki, M., Billmyre, R.B., Clancey, S.A., Wang, X., Heitman, J., 2016. Gene Network polymorphism illuminates loss and retention of novel RNAi silencing components in the *Cryptococcus* pathogenic species complex. *PLoS Genet.* 12 <https://doi.org/10.1371/journal.pgen.1005868>.
- Ferrareze, P.A.G., Maufrais, C., Streit, R.S.A., Priest, S.J., Cuomo, C., Heitman, J., Staats, C.C., Janbon, G., 2020. Application of an optimized annotation pipeline to the *Cryptococcus deuterogattii* genome reveals dynamic primary metabolic gene clusters and genomic impact of RNAi loss. *bioRxiv* 2020.09.01.278374. <https://doi.org/10.1101/2020.09.01.278374>.
- Ferrareze, P.A.G., Streit, R.S.A., Santos, P.R., dos Santos, F.M., dos Almeida, R.M.C. de, Schrank, A., Kmetzsch, L., Vainstein, M.H., Staats, C.C., 2017. Transcriptional analysis allows genome reannotation and reveals that *Cryptococcus gattii* VGII undergoes nutrient restriction during infection. *Microorganisms* 5, 49. <https://doi.org/10.3390/microorganisms5030049>.
- Friedman, R.C., Farh, K.K.H., Burge, C.B., Bartel, D.P., 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19, 92–105. <https://doi.org/10.1101/gr.082701.108>.
- Gebetsberger, J., Zywicki, M., Künzi, A., Polacek, N., 2012. tRNA-derived fragments target the ribosome and function as regulatory non-coding RNA in *haloferax volcanii*. *Archaea* 1–11. <https://doi.org/10.1155/2012/260909>.
- Goodarzi, H., Liu, X., Nguyen, H.C.B., Zhang, S., Fish, L., Tavazoie, S.F., 2015. Endogenous tRNA-derived fragments suppress breast cancer progression via YBX1 displacement. *Cell* 161, 790–802. <https://doi.org/10.1016/j.cell.2015.02.053>.
- Gordon, A., Hannon, G.J., 2010. Fastx-toolkit. FASTQ/A short-reads pre-processing tools. (Unpublished). [http://hannonlab.cshl.edu/fastx\\_toolkit](http://hannonlab.cshl.edu/fastx_toolkit) 5.
- Hirose, Y., Ikeda, K.T., Noro, E., Hiraoka, K., Tomita, M., Kanai, A., 2015. Precise mapping and dynamics of tRNA-derived fragments (tRFs) in the development of *Triops cancriformis* (tadpole shrimp). *BMC Genet.* 16, 83. <https://doi.org/10.1186/s12863-015-0245-5>.
- Keam, S.P., Sobala, A., Ten Have, S., Hutvagner, G., 2017. TRNA-derived RNA fragments associate with human multisynthetase complex (MSC) and modulate ribosomal protein translation. *J. Proteome Res.* 16, 413–420. <https://doi.org/10.1021/acs.jproteome.6b00267>.

- Kim, H.K., Fuchs, G., Wang, S., Wei, W., Zhang, Y., Park, H., Roy-Chaudhuri, B., Li, P., Xu, J., Chu, K., Zhang, F., Chua, M.S., So, S., Zhang, Q.C., Sarnow, P., Kay, M.A., 2017. A transfer-RNA-derived small RNA regulates ribosome biogenesis. *Nature* 552, 57–62. <https://doi.org/10.1038/nature25005>.
- Kumar, P., Anaya, J., Mudunuri, S.B., Dutta, A., 2014. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Med.* 12, 78. <https://doi.org/10.1186/s12915-014-0078-0>.
- Kuscu, C., Kumar, P., Kiran, M., Su, Z., Malik, A., Dutta, A., 2018. tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA* 24, 1093–1105. <https://doi.org/10.1261/rna.066126.118>.
- Kwon-Chung, K.J., Fraser, J.A., Doering, T.L., Wang, Z.A., Janbon, G., Idnurm, A., Bahn, Y.-S., 2014. *Cryptococcus neoformans* and *Cryptococcus gattii*, the etiologic agents of cryptococcosis. *Cold Spring Harb. Perspect. Med.* 4, a019760. <https://doi.org/10.1101/cshperspect.a019760>. –a019760.
- La Ferlita, A., Alaimo, S., Veneziano, D., Nigita, G., Balatti, V., Croce, C.M., Ferro, A., Pulvirenti, A., 2019. Identification of tRNA-derived ncRNAs in TCGA and NCI-60 panel cell lines and development of the public database tRFExplorer. *Database*. <https://doi.org/10.1093/database/baz115>.
- Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. <https://doi.org/10.1038/nmeth.1923>.
- Lee, Y.S., Shibata, Y., Malhotra, A., Dutta, A., 2009. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* 23, 2639–2649. <https://doi.org/10.1101/gad.1837609>.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAM-tools. *Bioinformatics* 25, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
- Li, S., Xu, Z., Sheng, J., 2018. tRNA-derived small RNA: a novel regulatory small non-coding RNA. *Genes* 9, 246. <https://doi.org/10.3390/genes9050246>.
- Li, Z., Ender, C., Meister, G., Moore, P.S., Chang, Y., John, B., 2012. Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. *Nucleic Acids Res.* 40, 6787–6799. <https://doi.org/10.1093/nar/gks307>.
- Luo, S., He, F., Luo, J., Dou, S., Wang, Y., Guo, A., Lu, J., 2018. Drosophila tsRNAs preferentially suppress general translation machinery via antisense pairing and participate in cellular starvation response. *Nucleic Acids Res.* 46, 5250–5268. <https://doi.org/10.1093/nar/gky189>.
- Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* 17, 10. <https://doi.org/10.14806/ej.17.1.200>.
- Megel, C., Hummel, G., Lalande, S., Ubrig, E., Cognat, V., Morelle, G., Salinas-Giegé, T., Duchêne, A.M., Maréchal-Drouard, L., 2019. Plant RNases T2, but not Dicer-like proteins, are major players of tRNA-derived fragments biogenesis. *Nucleic Acids Res.* 47, 941–952. <https://doi.org/10.1093/nar/gky1156>.
- Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C., Chen, S., Hannon, G.J., Qi, Y., 2008. Sorting of small RNAs into Arabidopsis Argonaute complexes is directed by the 5' terminal nucleotide. *Cell* 133, 116–127. <https://doi.org/10.1016/j.cell.2008.02.034>.
- Mleczko, A.M., Celichowski, P., Bąkowska-Żywicka, K., 2018. Transfer RNA-derived fragments target and regulate ribosome-associated aminoacyl-transfer RNA synthetases. *Biochim. Biophys. Acta - Gene Regul. Mech.* 1861, 647–656. <https://doi.org/10.1016/j.bbagr.2018.06.001>.
- Nunes, C.C., Gowda, M., Sailsbery, J., Xue, M., Chen, F., Brown, D.E., Oh, Y.Y., Mitchell, T.K., Dean, R.A., 2011. Diverse and tissue-enriched small RNAs in the plant pathogenic fungus *Magnaporthe oryzae*. *BMC Genom.* 12, 288. <https://doi.org/10.1186/1471-2164-12-288>.
- Phillip, W., Troll, W., 1977. High turnover rate of transfer RNA in tumor tissue. *Canc. Res.* 37, 3362–3366.
- Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. <https://doi.org/10.1093/bioinformatics/btq033>.
- Rajasingham, R., Smith, R.M., Park, B.J., Jarvis, J.N., Govender, N.P., Chiller, T.M., Denning, D.W., Loyse, A., Boulware, D.R., 2017. Global burden of disease of HIV-associated cryptococcal meningitis: an updated analysis. *Lancet Infect. Dis.* 17, 873–881. [https://doi.org/10.1016/S1473-3099\(17\)30243-8](https://doi.org/10.1016/S1473-3099(17)30243-8).
- Ryvkin, P., Leung, Y.Y., Silverman, I.M., Childress, M., Valladares, O., Dragomir, I., Gregory, B.D., Wang, L.S., 2013. HAMR: high-throughput annotation of modified ribonucleotides. *RNA* 19, 1684–1692. <https://doi.org/10.1261/rna.036806.112>.
- Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* 11, e0163962. <https://doi.org/10.1371/journal.pone.0163962>.
- Speer, J., Gehrke, C.W., Kuo, K.C., Waalkes, T.P., Borek, E., 1979. tRNA breakdown products as markers for cancer. *Cancer* 44, 2120–2123. [https://doi.org/10.1002/1097-0142\(197912\)44:6<2120::AID-CNCR2820440623>3.0.CO](https://doi.org/10.1002/1097-0142(197912)44:6<2120::AID-CNCR2820440623>3.0.CO).
- Sun, C., Fu, Z., Wang, S., Li, J., Li, Y., Zhang, Y., Yang, F., Chu, J., Wu, H., Huang, X., Li, W., Yin, Y., 2018. Roles of tRNA-derived fragments in human cancers. *Canc. Lett.* 414, 16–25. <https://doi.org/10.1016/j.canlet.2017.10.031>.
- Šurbanovski, N., Brilli, M., Moser, M., Si-Ammour, A., 2016. A highly specific microRNA-mediated mechanism silences LTR retrotransposons of strawberry. *Plant J.* 85, 70–82. <https://doi.org/10.1111/tpj.13090>.
- Torres-Martínez, S., Ruiz-Vázquez, R.M., 2017. The RNAi universe in fungi: a varied landscape of small RNAs and biological functions. *Annu. Rev. Microbiol.* 71, 371–391. <https://doi.org/10.1146/annurev-micro-090816-093352>.
- Varkonyi-Gasic, E., Wu, R., Wood, M., Walton, E.F., Hellens, R.P., 2007. Protocol: a highly sensitive RT-PCR method for detection and quantification of microRNAs. *Plant Methods* 3, 12. <https://doi.org/10.1186/1746-4811-3-12>.
- Wang, Q., Lee, I., Ren, J., Ajay, S.S., Lee, Y.S., Bao, X., 2013. Identification and functional characterization of tRNA-derived RNA fragments (tRFs) in respiratory syncytial virus infection. *Mol. Ther.* 21, 368–379. <https://doi.org/10.1038/mt.2012.237>.
- Wang, Q., Li, T., Xu, K., Zhang, W., Wang, X., Quan, J., Jin, W., Zhang, M., Fan, G., Wang, M.B., Shan, W., 2016. The tRNA-derived small RNAs regulate gene expression through triggering sequence-specific degradation of target transcripts in the oomycete pathogen *Phytophthora sojae*. *Front. Plant Sci.* 7 <https://doi.org/10.3389/fpls.2016.01938>.
- Wang, Zhensheng, Wei, C., Hao, X., Deng, W., Zhang, L., Wang, Zenglei, Wang, H., 2019. Genome-wide identification and characterization of transfer RNA-derived small RNAs in *Plasmodium falciparum*. *Parasites Vectors* 12, 36. <https://doi.org/10.1186/s13071-019-3301-6>.
- Weick, E.-M., Miska, E.A., 2014. piRNAs: from biogenesis to function. *Development* 141, 3458–3471. <https://doi.org/10.1242/dev.094037>.
- Wilson, R.C., Doudna, J.A., 2013. Molecular mechanisms of RNA interference. *Annu. Rev. Biophys.* 42, 217–239. <https://doi.org/10.1146/annurev-biophys-083012-130404>.
- Xu, W.L., Yang, Y., Wang, Y.D., Qu, L.H., Zheng, L.L., 2017. Computational approaches to tRNA-derived small RNAs. *Non-coding RNA* 3, 2. <https://doi.org/10.3390/ncrna3010002>.
- Zheng, L.L., Xu, W.L., Liu, S., Sun, W.J., Li, J.H., Wu, J., Yang, J.H., Qu, L.H., 2016. tRF2Cancer: a web server to detect tRNA-derived small RNA fragments (tRFs) and their expression in multiple cancers. *Nucleic Acids Res.* 44, W185–W193. <https://doi.org/10.1093/nar/gkw414>.

### 3.1.1. Discussão geral do artigo

O primeiro passo na geração de conhecimento sobre produção e funcionamento de tRFs em qualquer organismo é a identificação dos mesmos, seja por metodologias *in silico* (XU et al., 2017) ou *in vitro* (BAKOWSKA-ŻYWICKA et al., 2016), de maneira que a precisão do conhecimento gerado é diretamente relacionada à precisão da metodologia escolhida para a identificação dos mesmos. Infelizmente, as metodologias *in silico* atualmente disponíveis para a identificação de tRFs a partir de dados de sRNA-seq apresentam limitações como a difícil ou impossibilitada execução em organismos que não os para os quais a ferramenta foi desenvolvida (WANG et al., 2022), assim como a utilização de algoritmos ingênuos (SELITSKY; SETHUPATHY, 2015) ou falta de testes para a identificação de tRFs (SHI et al., 2018). A fim de contornar grande parte desses problemas, desenvolvemos um *workflow* de processamento de dados de sRNAs a ser aplicado antes da predição de tRFs pela ferramenta tDRmapper (SELITSKY; SETHUPATHY, 2015), selecionada pela facilidade de execução em qualquer organismo, de maneira a corrigir a abordagem simplista aplicada pela ferramenta. Apesar de estridente, nossa abordagem não apenas foi capaz de gerar predições de tRFs para ambos os organismos, como foram demonstrados por meio de metodologias *in vitro* a presença de tRFs preditos por essa abordagem.

O campo de pesquisa sobre tRFs ainda apresenta diversas lacunas no conhecimento, assim como inconsistências. Dentre elas, se encontra a associação dos tRFs com a via canônica de RNAi, visto que da mesma forma que temos evidência da associação de tRFs com argonautas e seu funcionamento pela via de RNAi (COLE et al., 2009; DI FAZIO et al., 2022), existem também evidências da independência de algumas proteínas da via de RNAi para sua produção (KUSCU et al., 2018; MEGEL et al., 2019), assim como seu funcionamento por mecanismos completamente alheios à via de RNAi (GEBETSBERGER et al., 2012; KIM et al., 2017; MLECZKO; CELICHOWSKI; BAKOWSKA-ŻYWICKA, 2018). A avaliação da dependência da via de RNAi é ainda mais relevante para fungos, visto que apesar de que a mesma ser usualmente associada com a defesa e estabilização do genoma, existem diversos

eventos de perda da via de RNAi canônica por parte de fungos (NICOLÁS; GARRE, 2016).

Dada a aparente conservação de tRFs entre organismos do mesmo gênero (LUO et al., 2018; WANG et al., 2016), na ausência de dependência da via de RNAi para a produção e funcionamento de tRFs seria esperado que o conjunto de tRFs entre *C. gattii* R265 e *C. neoformans* H99 não apresentasse grande variação. Porém, foi observado que não apenas o conjunto de tRNAs que produz tRFs varia entre esses organismos, mas também a preferência por região da qual se originam os tRFs dentro do tRNA e o tamanho dos tRFs. De maneira interessante, *C. neoformans*, que possui uma via de RNAi funcional, apresentou preferência de produção de tRFs a partir das extremidades dos tRNAs e de tamanhos entre 18 e 24 nucleotídeos, padrão condizente com o processamento por Dicer (LANGENBERGER et al., 2013; MENG et al., 2017). Nessa hipótese, em *C. neoformans*, Dicer reconheceria as extremidades do tRNA, associadas em uma dupla-fita assim como pré-miRNAs, e faria a clivagem das mesmas, produzindo uma dupla-fita composta por um tRF e um “tRF\*”, de maneira similar aos miRNAs (BARTEL, 2018), sendo o tRF posteriormente carregado em Argonata.

Essa associação entre os tRFs de *C. neoformans* e a via de RNAi se estendem também à análise de possíveis alvos para os tRFs. Enquanto que os tRFs de *C. gattii* não apresentaram nenhum tipo de correlação com os níveis de mRNA de seus alvos putativos, tRFs de *C. neoformans* apresentaram uma correlação negativa com os níveis de mRNA de seus possíveis alvos, indicando um potencial na atuação sobre a regulação dos níveis de mRNA dos mesmos. Essa hipótese é reforçada pela demonstração de clivagem em um alvo de tRF no sítio de interação predito para o mesmo com o mRNA em *C. neoformans*, indicando que nesse organismo os tRFs possivelmente apresentam capacidade de direcionamento de mecanismos de clivagem para seus alvos por meio de pareamento de base da mesma forma que miRNAs, podendo essa maquinaria ser o RISC, componente da via de RNAi.

Juntos, esses resultados apontam para uma associação, parcial ou completa, entre tRFs e a via de RNAi em *C. neoformans*, indicando que na presença da mesma tRFs podem utilizá-la para atuar na regulação da expressão gênica como predito para

outros organismos, podendo o mesmo ser aplicável para outros fungos. Apesar disso, tornam-se necessárias ainda avaliações mais diretas da associação de tRFs com a via de RNAi nesse organismo, como a avaliação da presença de tRFs associados a Argonautas e do funcionamento desses tRFs por meio de Argonauta sobre alvos de mRNA.

Enquanto que *C. neoformans* apresenta uma associação com a via de RNAi, *C. gattii* manteve a produção de tRFs na ausência da mesma por mecanismos alternativos, podendo estes estarem presentes também em *C. neoformans*, visto que não existe evidência da extensão da associação de tRFs nesse organismo com a via de RNAi. Apesar da possibilidade da presença também em *C. neoformans*, visto que os conjuntos de tRF entre os dois organismos são extremamente diferentes, a produção e funcionamento de tRFs por mecanismos que não a via de RNAi ocorrem de maneira diferente entre os mesmos, seja pela utilização de mecanismos alternativos à via de RNAi diferentes ou por variações significativas em um mecanismo utilizado por ambos, visto que seria esperado a identificação de um conjunto razoável de tRFs comuns a ambas as espécies se os mecanismos de produção e funcionamento independentes de RNAi fossem compartilhados. Assim, é possível que os tRFs em *C. gattii* apresentem um mecanismo de produção não presente em *C. neoformans*, ou não utilizado pelo mesmo na produção de tRFs.

Além da variação do conjunto de tRFs devido a mecanismos alternativos que não dependam da via de RNAi, a variação do conjunto de tRFs em *C. gattii* também pode estar associada a adoção de sRNAs passíveis de produção e funcionamento na ausência da via de RNAi como mecanismos compensatórios para a perda da regulação pela via de RNAi. Apesar da usual associação da via de RNAi com mecanismos de defesa do genoma e não de regulação gênica em fungos, a ampla utilização dessa via em diversos organismos na regulação da expressão gênica demonstra o potencial da mesma para essa finalidade (BARTEL, 2018), podendo ela ser utilizada, mesmo que em uma menor extensão, na regulação gênica de fungos (NICOLÁS; GARRE, 2016). Dessa maneira, a perda da via de RNAi poderia acarretar a perda de redes de regulação da expressão gênica, que por sua vez poderiam ter um impacto negativo no

*fitness* do organismo em seu nicho, o que não é observado em *C. gattii* R265 mesmo na ausência da via de RNAi.

Em linha com esse racional, na análise de alvos de mRNA para tRFs em *C. gattii* foi visto que tRFs nesse organismo apresentam uma quantidade maior de alvos que tRFs de *C. neoformans*. Além disso, apenas em *C. gattii* foi identificado um possível alvo para tRFs em sequências de elementos de transposição. Assim, é possível que após a perda da via de RNAi, *C. gattii* teve selecionado um novo conjunto de tRFs, com sequências distintas dos tRFs de *C. neoformans*, assim como um tamanho reduzido que facilitaria o pareamento de bases ao longo de sua extensão e permitiria uma maior quantidade de alvos, de maneira que esse novo conjunto apresentasse um papel compensatório pela perda de via de RNAi. Apesar de evidências contra a regulação à nível de mRNA, visto que tRFs em *C. gattii* aparentam não ter capacidade de clivar seus alvos, outras formas de regulação podem ser empregadas, as quais não são possíveis de serem avaliadas pelos níveis de mRNA, como na tradução.

Enquanto que a extensão da associação de tRFs com a via de RNAi em *C. neoformans* não pode ser analisada e o conjunto de resultados apenas nos permite hipotetizar a associação de tRFs com essa via nesse organismo, a clara presença de tRFs em *C. gattii* deixa evidente a independência da via de RNAi de ao menos parte dos tRFs. Essa demonstração é extremamente valiosa para o estudo de fungos, visto que diversos fungos apresentam a perda da via de RNAi, de maneira que a presença de sRNAs que independem dessa via para seu funcionamento em fungos abre novas possibilidades de mecanismos a serem explorados e utilizados para desenvolvimento de tecnologias nesses organismos. Apesar dessa demonstração, os dados obtidos nesse estudo não puderam elucidar os mecanismos alternativos envolvidos com tRFs, com exceção da evidência de que, ao menos em *C. gattii*, tRFs não apresentam capacidade de reduzir os níveis de mRNAs de seus alvos, qualquer seja o mecanismo pelo qual atuem. Assim, são necessários novos estudos a fim de elucidar os mecanismos envolvidos na regulação exercida pelos tRFs, que podem levar à descoberta ou confirmação de mecanismos ancestrais, visto evidências da presença de tRFs em todos os reinos da vida (BAKOWSKA-ŻYWICKA et al., 2016; GEBETSBERGER et al., 2012; KUMAR et al., 2014; LUO et al., 2018; WANG et al.,

2019), assim como de novos mecanismos de regulação por sRNAs gênero ou espécie-específicos que utilizem tRFs, visto que vias de RNAi não-canônicas já foram identificadas em fungos (CÁNOVAS-MÁRQUEZ et al., 2021; PÉREZ-ARQUES et al., 2020).

### 3.2. Capítulo 2: Avaliação de sRNAs regulatórios em vesículas extracelulares de isolados *C. auris*

Nesse capítulo, é avaliado o conteúdo de sRNAs de vesículas extracelulares de diferentes isolados de *C. auris* na presença ou ausência do antifúngico caspofungina por meio de análises *in silico* de dados de sRNA-seq. Utilizando uma adaptação da abordagem aplicada no Capítulo 1, são identificados os tRFs presentes nas vesículas, sendo também feita posteriormente a identificação de miRNAs e análise das características dos sRNAs identificados. Em seguida, são preditos alvos para os sRNAs identificados entre os mRNAs de hospedeiros mamíferos, *H. sapiens* e *M. musculus*, e é executada a análise de enriquecimento funcional do conjunto de alvos identificados, a fim de avaliar possíveis vias-alvo para sRNAs fúngicos nesses hospedeiros. Ao final, são discutidos a relevância dos achados para o entendimento da interação patógeno-hospedeiro e do papel dos sRNAs nesse contexto, assim como especificamente o impacto que sRNAs de *C. auris* podem exercer durante a infecção.

### 3.2.1. Materiais e métodos

#### 3.2.1.1. Dados utilizados

A sequência genômica de *C. auris* utilizada para a predição de tRFs e miRNAs foi recuperada do NCBI no formato fasta, sob o acesso GCF\_002775015.1, enquanto as sequências de mRNA de *Homo sapiens* e *Mus musculus* utilizadas para a predição de alvos dos sRNAs vesiculares identificados, também obtidas no NCBI em formato fasta, encontram-se respectivamente sob os acessos GRCh38.p13 e GRCm39. Os dados de sRNA-seq utilizados para a predição de sRNAs foram disponibilizados pela Prof. Lysangela Alves do Instituto Carlos Chagas da Fiocruz (Curitiba, PR) (MUNHOZ DA ROCHA et al., 2021), que correspondem a bibliotecas de sequenciamento preparadas em triplicatas técnicas provenientes de vesículas extracelulares dos isolados CDC385, CDC387, MMC1 e MMC2, na presença ou ausência de caspofungina.

#### 3.2.1.2. Predição de tRFs em vesículas extracelulares

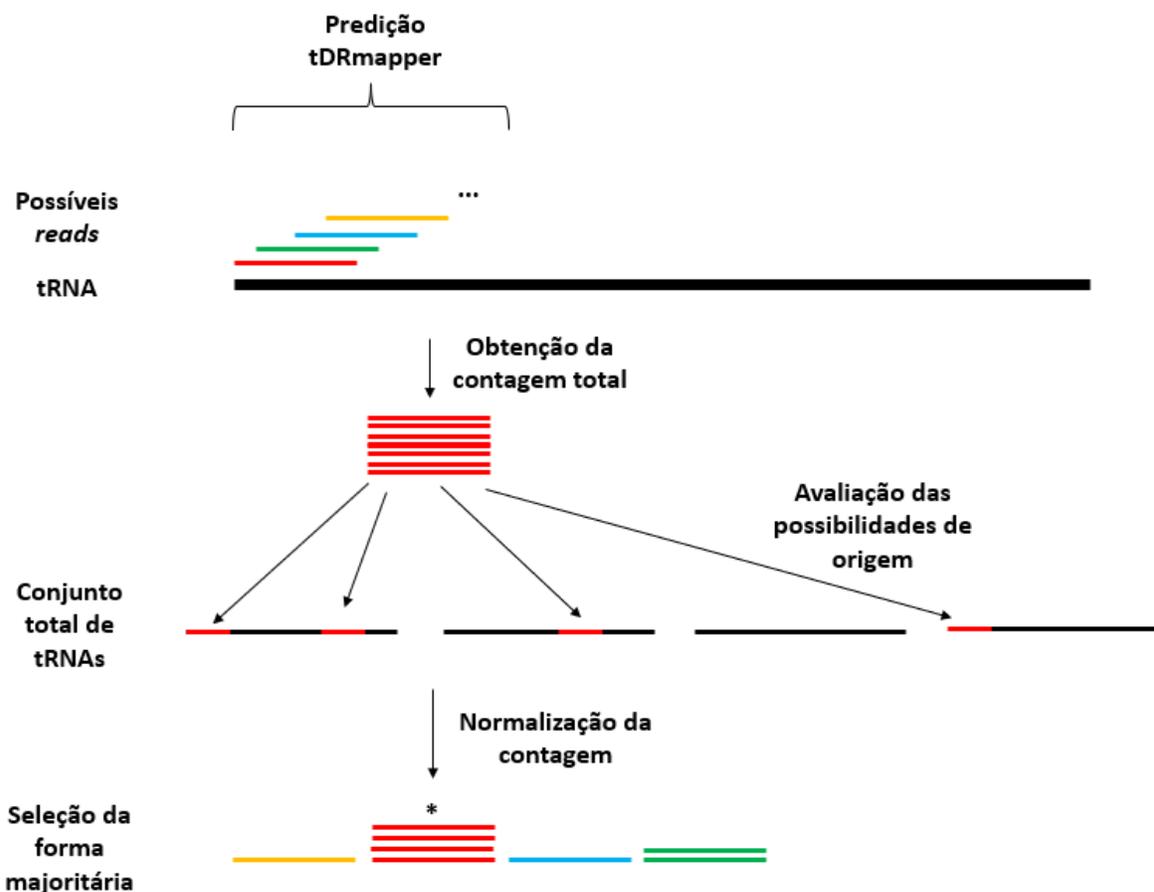
A predição de tRFs foi realizada para cada biblioteca de *reads* de sRNAs separadamente, sem a junção das triplicatas, aplicando a mesma metodologia descrita no artigo correspondente ao Capítulo 1 (STREIT et al., 2021), ocorrendo variação apenas nos passos finais de delimitação de tRF e filtragem de tRFs baseado em sua expressão. Visto que a região predita como tRF pelo tDRmapper (SELITSKY; SETHUPATHY, 2015) pode corresponder a um agrupamento de diversos *reads* diferentes, não necessariamente a cobertura da região correspondendo a *reads* com a mesma sequência, foi avaliada a contagem de *reads* para cada *read* possível de ser encontrado inteiramente dentro da região predita como tRF (Figura 6). A contagem foi feita obtendo o número de cópias que cada *read* da região possuía dentro da biblioteca analisada, normalizando a contagem pelo número de vezes que aquele *read* poderia alinhar perfeitamente com diferentes regiões de diferentes tRNAs, uma vez que diversos tRNAs partilham sequências em comum e não é possível atribuir de maneira precisa a origem de cada *read* dessas regiões. Após a obtenção das contagens normalizadas, dado o princípio de que tRFs são produzidos em alta quantidade e de maneira precisa (COLE et al., 2009; ZHENG et al., 2016), o *read* majoritário da região

predita como tRF foi delimitado como candidato à tRF. Todos os candidatos a tRF de todas as bibliotecas foram então avaliados com relação a porcentagem que sua contagem normalizada representava com relação à contagem de *reads* da região predita como tRF pelo tDRmapper, obtida pela soma das contagens normalizadas de cada *read* avaliado no passo anterior, assim como com relação à contagem de reads do tRNA como um todo, obtida a partir do alinhamento dos *reads* da biblioteca contra as sequências de tRNA utilizando bowtie2 (LANGMEAD; SALZBERG, 2012) e consequente obtenção das contagens para cada tRNA utilizando samtools idxstats (DANECEK et al., 2021). Utilizando a mediana dos valores de porcentagem da região e do tRNA obtidos, sendo os valores respectivamente 45.95% e 17.05%, todos os candidatos a tRF de todas as bibliotecas foram conjuntamente divididos em quadrantes, e os candidatos a tRFs localizados no quadrante inferior esquerdo foram excluídos dado o fato de que não se apresentam como *reads* majoritários do tRNA, o que é esperado de *bona fide* tRFs, possuindo baixa confiabilidade. Os demais candidatos foram aceitos como tRFs e analisados quanto às suas características. Para a avaliação da composição de nucleotídeos nas extremidades 5' e 3' dos tRFs identificados, foram extraídos os nucleotídeos correspondentes aos 7 primeiros e 7 últimos nucleotídeos, e avaliados utilizando o servidor WebLogo 3 (CROOKS et al., 2004).

#### 3.2.1.3. Predição de miRNAs em vesículas extracelulares

Para a predição de miRNAs, as bibliotecas tiveram os *reads* atribuídos a tRNAs pelo passo anterior removidos, gerando bibliotecas livres de *reads* de tRNA que pudessem levar a identificação de tRFs como miRNAs. Após isso, as bibliotecas foram utilizadas pelo programa mirnova (VITSIOS et al., 2017) para a identificação de miRNAs, utilizando como modelo de treino o conjunto de dados de *H. sapiens* fornecido pelos desenvolvedores, e sem a utilização de um genoma de referência. Devido à dificuldade em obtenção de programas específicos para fungos passíveis de execução e ao fato de que se objetiva identificar miRNAs em vesículas a serem endereçados a hospedeiros mamíferos, que podem apresentar características semelhantes a miRNAs humanos para sua utilização pelo hospedeiro, a aplicação de um modelo de miRNAs

humanos é razoável. Os miRNAs reportados pelo programa foram filtrados quanto à sua expressão, a fim de remover miRNAs com pouco ou nenhum suporte de *reads*, sendo removidos miRNAs que não apresentassem pelo menos 100 *reads* de suporte em ao menos uma das bibliotecas. Os miRNAs filtrados foram então analisados quanto às suas características. A análise da composição de nucleotídeos das extremidades dos miRNAs foi executada da mesma forma que para os tRFs.



**Figura 6: Método de filtragem dos resultados da predição do tDRmapper.** Passos da filtragem proposta dos resultados da predição de tRFs pelo tDRmapper conforme descrito na metodologia.

#### 3.2.1.4. Predição de alvos para sRNAs em hospedeiros mamíferos

Os alvos para os tRFs e miRNAs identificados foram preditos para os transcritos de *M. musculus* e *H. sapiens* utilizando o *software* miranda (JOHN et al., 2004). Os

alvos identificados foram separados para cada miRNAs e tRFs com tamanhos entre 18 e 24 nucleotídeos, sendo em seguida alimentados na plataforma *online* Panther e executada análise de super representação de termos de ontologias gênicas relacionadas a processos biológicos (MI et al., 2021; THOMAS et al., 2006). Na plataforma, os conjuntos carregados foram atribuídos ao organismo de origem, *Homo sapiens* ou *Mus musculus*, e selecionada a análise “Statistical overrepresentation test” com o conjunto de anotação “GO biological process complete”, utilizando como lista de referência a lista de referência que inclui todos os genes da anotação do organismo em questão, selecionando por fim o teste “Fisher’s Exact” com a correção “Calculate False Discovery Rate”. Para os gráficos no REVIGO (SUPEK et al., 2011), foram fornecidos os termos identificados na forma de seu código (GO:XXXX) para o teste prévio, unindo os resultados de todos os tRFs ou miRNAs, sendo os termos separados por super e sub-representados, associados ao número de vezes que o mesmo termo foi identificado nas análises. Foram selecionados os parâmetros “Small (0.5)” para a simplificação da lista final e “Higher value is better” para os valores associados aos termos nos casos de termos super-representados. Para termos sub-representados, em menor quantidade, o parâmetro de simplificação foi alterado para “Large (0.9)”. Os resultados foram recuperados na forma de *script* para a geração de um “*Tree map*” em R, sendo executado no Rstudio (R versão 4.2.0) e salvo o gráfico gerado.

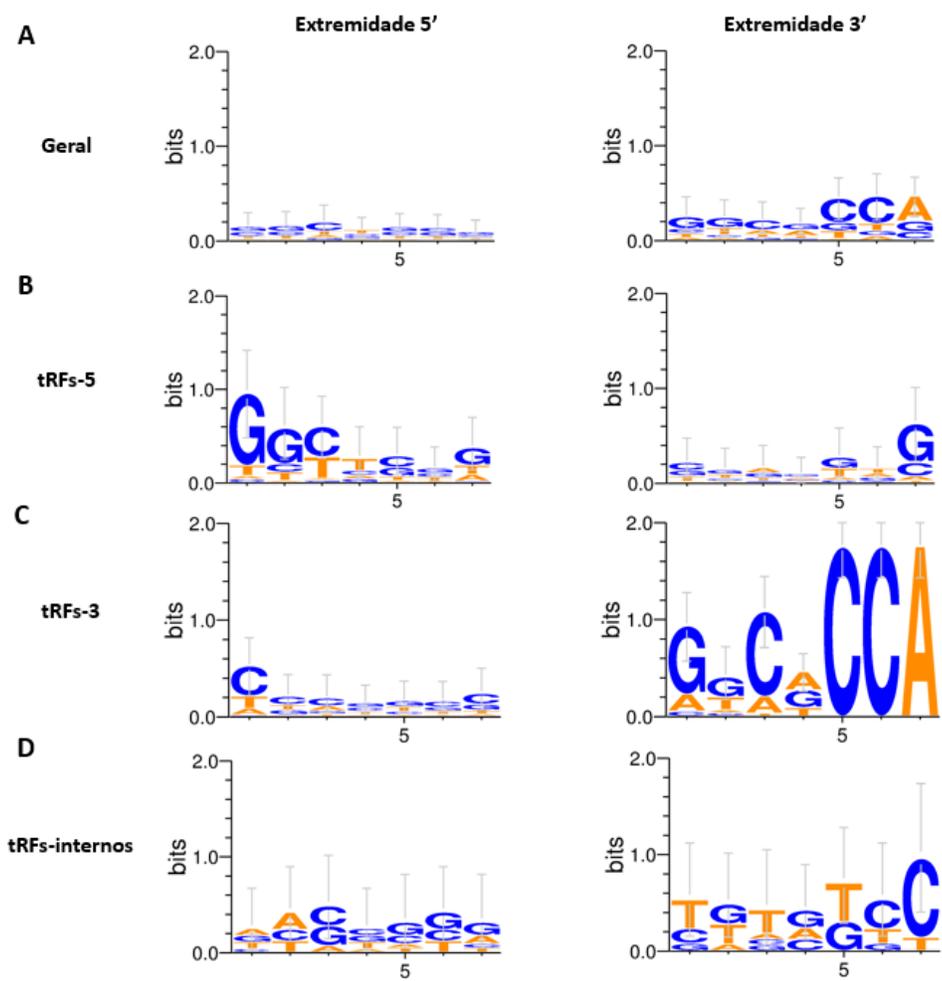
### 3.2.2. Resultados

#### 3.2.2.1. Identificação de tRFs em vesículas de *C. auris*

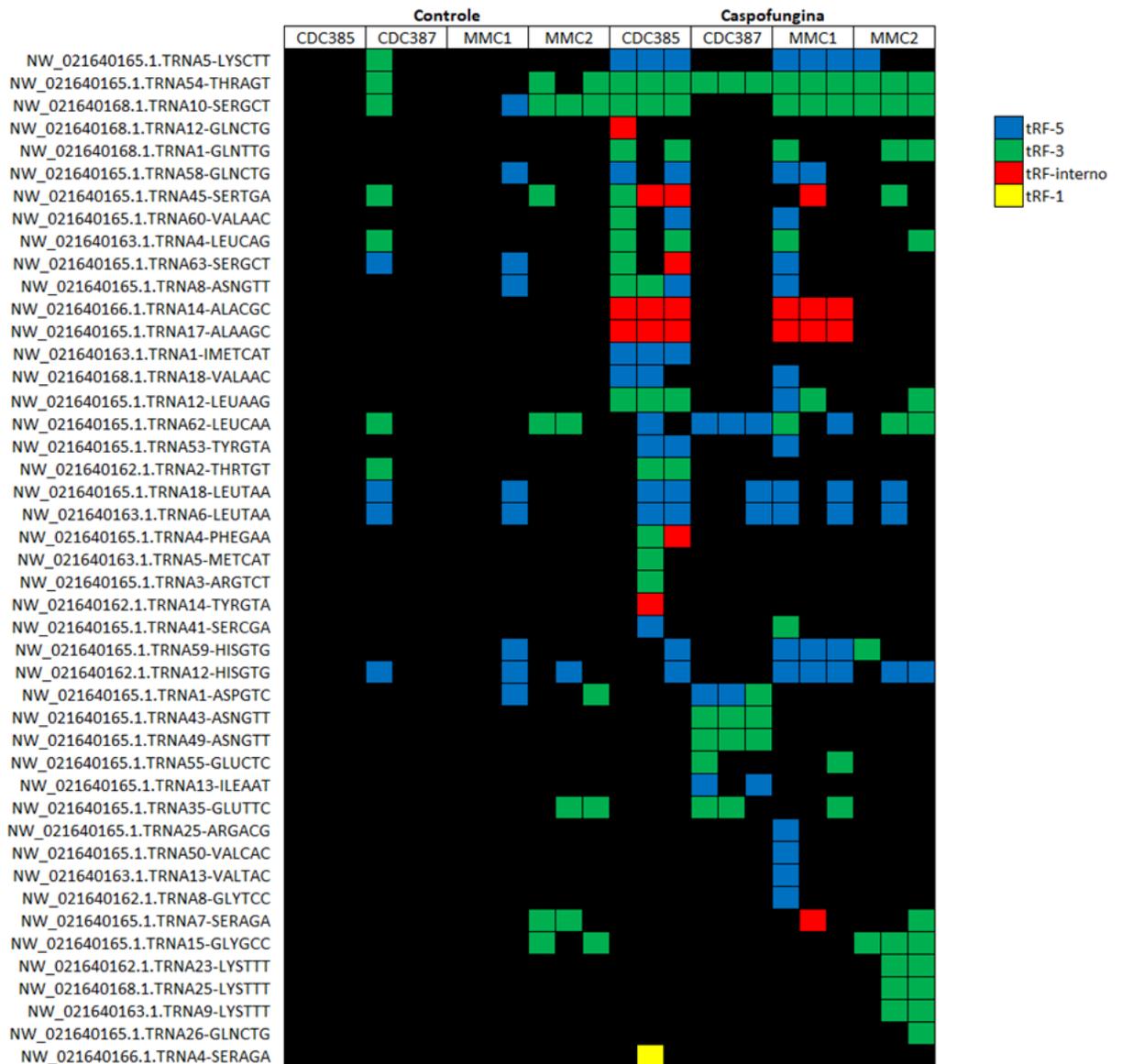
Utilizando nossa abordagem de predição de tRFs, foram identificados tRFs em 17 das 24 bibliotecas de sRNAs analisadas, totalizando 71 sequências únicas de tRF entre todas as condições e linhagens, produzidas a partir de 45 sequências únicas de tRNA. Das 71 sequências únicas identificadas, 70 apresentaram origens em sequências de tRNA maduro, enquanto uma apresentou origem em transcritos de tRNA primário, compreendendo a junção entre a sequência do tRNA maduro e a região 3’ *trailer*, apesar de que esta última tem suporte de apenas uma das três triplicatas referentes à linhagem e condição analisada. Em sua maioria, tRFs foram identificados em bibliotecas correspondentes ao tratamento por caspofungina (Figura 7), as 12

bibliotecas correspondendo esse tratamento apresentando tRFs preditos, sendo, com exceção das bibliotecas da linhagem MMC2 na condição controle, as únicas bibliotecas nas quais foram identificados tRFs nas três triplicatas. Enquanto em 15 das bibliotecas foram identificados tRFs nas três triplicatas referentes à mesma linhagem e condição, apenas 9 dos 71 tRFs identificados apresentaram suporte entre pelo menos três triplicatas da mesma linhagem e condição. Além disso, 8 tRFs foram identificados como sendo produzidos a partir de mais de um tRNA, sendo essa identificação provavelmente relacionada à uma alta similaridade de sequência entre os tRNAs, e 20 tRNAs apresentaram capacidade de produção de mais de um tRF.

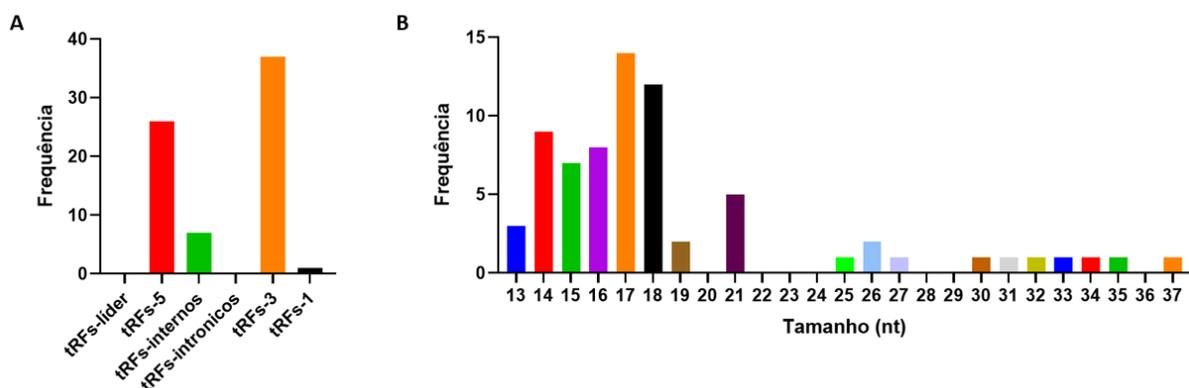
Os tRFs identificados apresentaram uma preferência de produção a partir das extremidades 5' e 3' dos tRNAs, com tRFs-3 representando 52% e tRFs-5 representando 36,6% do total de tRFs (Figura 8A). Enquanto o tamanho de tRFs identificados variou entre 13 e 37 nucleotídeos, foi observada uma preferência por tRFs de tamanho entre 14 e 18 (Figura 8B). Quando analisados quanto à composição dos 7 nucleotídeos de cada extremidade dos tRFs identificados, foi verificado que, quando avaliados em conjunto, não foi identificada preferência por nucleotídeos em nenhuma posição nas extremidades dos tRFs, com exceção da sequência CCA na extremidade 3', provavelmente oriunda dos tRFs 3' (Figura 9A). Quando analisados separadamente os tRFs-5, 3 e internos, foram identificados diferentes padrões para cada classe. Para tRFs-5, foi identificada uma preferência por nucleotídeo G na primeira posição do tRF, seguido, com menor preferência, por G e em seguida C/T, enquanto que na extremidade 3' existe uma preferência também moderada por nucleotídeos G na última posição (Figura 9B). Para tRFs 3', existe uma preferência moderada por C/T no primeiro nucleotídeo, enquanto, além da presença unânime da sequência CCA na extremidade 3', o que define essa classe, foi identificada uma preferência por nucleotídeos C e G nas posições -5 e -7 a partir do fim do tRF, respectivamente (Figura 9C). Por fim, para os tRFs internos, a única preferência clara é a de nucleotídeos C na extremidade 3' dos tRFs (Figura 9D).



**Figura 9: Preferência por bases nas extremidades das diferentes classes de tRFs identificados.** Logos da frequência observada de bases nos 7 nucleotídeos das extremidades 5' e 3' dos tRFs para os conjuntos (A) de tRFs totais, (B) tRFs-5, (C) tRFs-3 e (D) tRFs-internos. Os logos foram gerados no servidor Weblogo3.



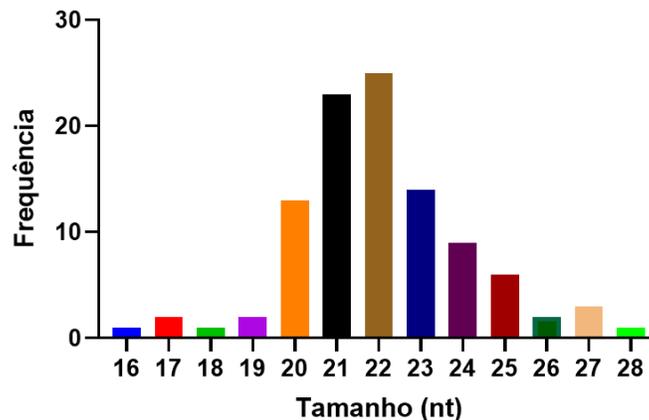
**Figura 7: tRNAs produtores de tRFs de acordo com a predição para cada biblioteca de sRNA-seq.** Representação gráfica da predição da produção de tRFs para cada sequência única de tRNA em *C. auris*. São apresentados apenas os tRNAs de sequência única com tRFs preditos a partir dos mesmos em pelo menos uma das bibliotecas. As condições e linhagens analisadas estão apresentadas sobre o cabeçalho do gráfico, sendo o trio de colunas abaixo de cada divisória de linhagem a representação das triplicatas de sRNA-seq. tRFs-5 são representados em azul, tRFs-3 são representados em verde, tRFs-internos são representados em vermelho e tRFs-1 são representados em amarelo. Espaços pretos denotam a ausência da predição de tRFs para aquele tRNA na linhagem e condição correspondente.



**Figura 8: Características dos tRFs identificados.** Frequência de (A) origem de classes e (B) tamanhos dos tRFs identificados em vesículas extracelulares de *C. auris*, avaliando o conjunto total de tRFs identificados entre todas as bibliotecas. (nt) representa o tamanho em nucleotídeos.

#### 3.2.2.2. Identificação de miRNAs em vesículas de *C. auris*

Foram identificados 102 miRNAs preditos em vesículas, sendo identificados *reads* para pelo menos 1 miRNA em cada biblioteca, independente da quantidade de *reads* presentes. Apesar da presença de *reads* dos miRNAs identificados em todas as bibliotecas, apenas 1 miRNA teve *reads* identificados nas 3 bibliotecas correspondentes à mesma condição, independente de quantidade, enquanto 67 dos 102 miRNAs apresentaram *reads* apenas em 1 biblioteca de *reads*. Os miRNAs identificados apresentaram tamanho variando de 16 a 28 nucleotídeos, com uma preferência de tamanho de 21 e 22 nucleotídeos (Figura 10). Quando avaliados quanto a composição de suas extremidades, não foram identificadas preferências tanto analisando os miRNAs conjuntamente quanto analisando separadamente os miRNAs de tamanho de 21 e 22 nucleotídeos.



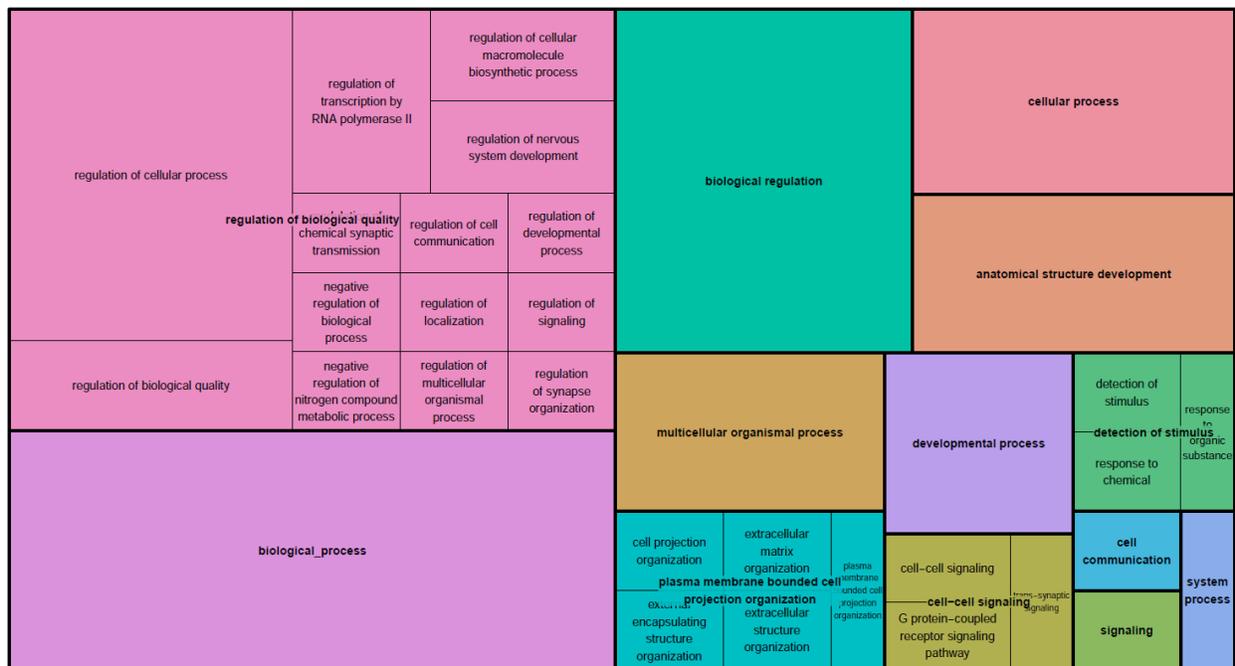
**Figura 10: Características dos miRNAs identificados.** Frequência de tamanhos dos miRNAs identificados em vesículas extracelulares de *C. auris*, avaliando o conjunto total de miRNAs identificados em todas as condições. (nt) representa o tamanho em nucleotídeos.

### 3.2.2.3. Identificação de alvos de sRNAs em hospedeiros mamíferos

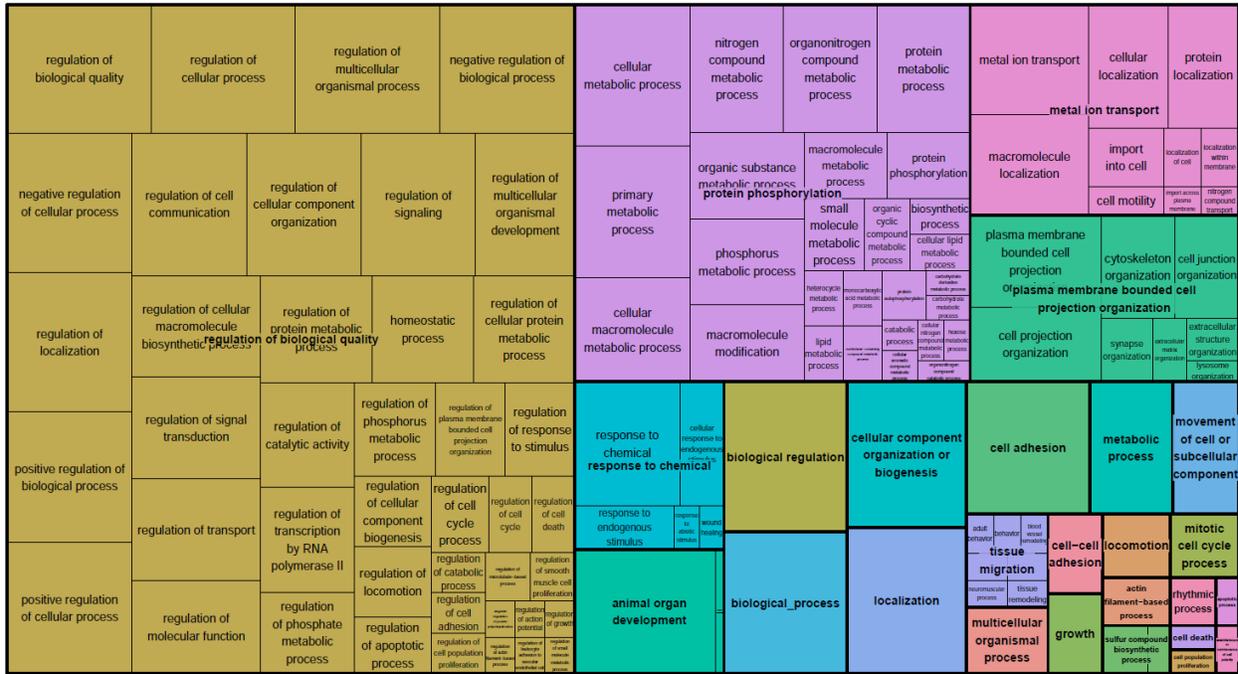
Foram identificados 19 tRFs de tamanho entre 18 e 24 nucleotídeos com alvos em *H. sapiens* e *M. musculus*, apresentando os 19 tRFs alvos em ambos os organismos, enquanto que 87 miRNAs apresentaram alvos nesses organismos, sendo também preditos alvos para ambos os organismos em todos os miRNAs. Para tRFs, a média de alvos de mRNA predita por tRF foi de aproximadamente 40742 mRNAs em *H. sapiens* e 34114 mRNAs em *Mus musculus*, quantidade que representa aproximadamente 24,94% e 26,17%, respectivamente, dos RNAs presentes nas anotações dos transcriptomas utilizados para a predição de alvos. Para miRNAs, o número de alvos foi próximo, com uma média de aproximadamente 43721 alvos por miRNA preditos em *H. sapiens* e 36751 em *M. musculus*.

Quanto ao enriquecimento funcional destes potenciais alvos, enquanto todos os 19 tRFs apresentaram algum enriquecimento de termos de ontologias gênicas informativas (excluído o termo “Unclassified”) em *M. musculus*, apenas 17 dos tRFs apresentaram termos informativos. Enquanto que tanto em *H. sapiens* quanto em *M. musculus*, os principais termos super representados comuns a múltiplos tRFs apresentaram relação com o desenvolvimento (Figuras 11 e 12), termos relacionados à

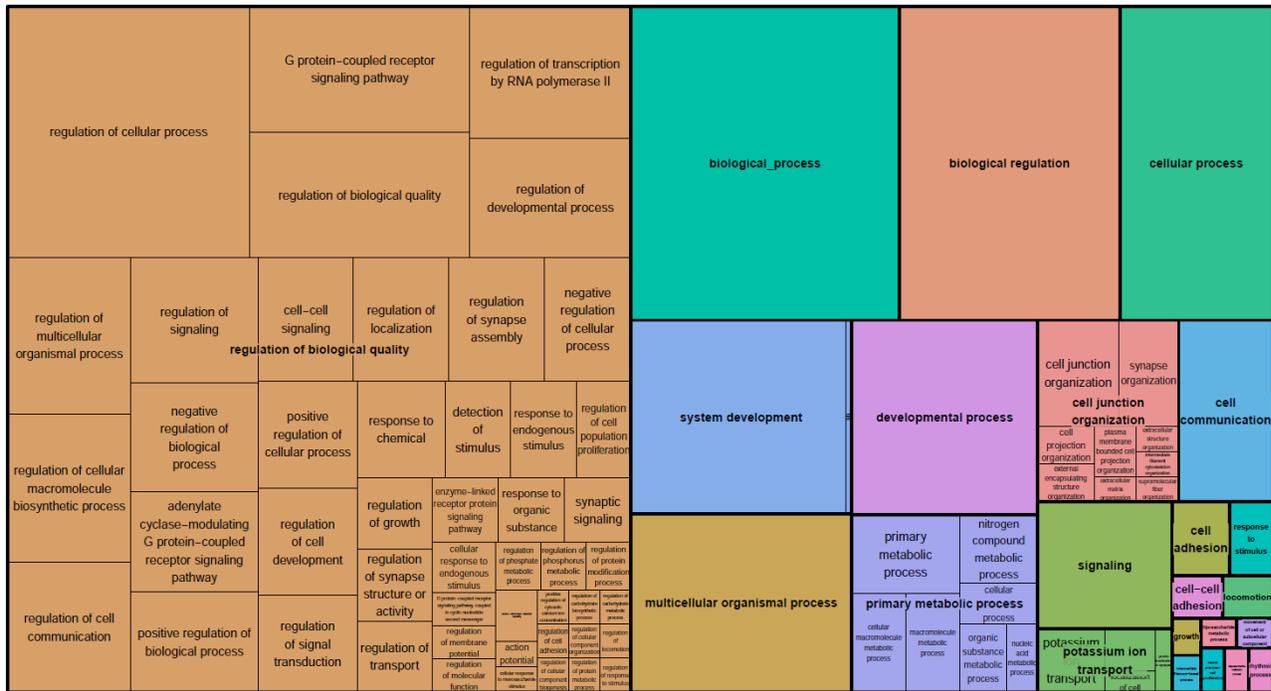
transcrição, metabolismo celular e sinalização intracelular identificados. Para miRNAs, todos apresentaram enriquecimento funcional, seguindo padrões semelhantes de termos super representados aos tRFs para ambos os organismos (Figuras 13 e 14). Quando levados em conta termos sub-representados, diversos termos relacionados com o sistema imunitário são identificados consistentemente entre os diferentes sRNAs em ambos os organismos (Figuras 15 à 18), de maneira que apesar de por meio de um número reduzido de alvos, os sRNAs de vesículas podem apresentar capacidade de regular genes com papéis no funcionamento do sistema imunitário e, portanto, serem de grande relevância para o desenvolvimento de infecções.



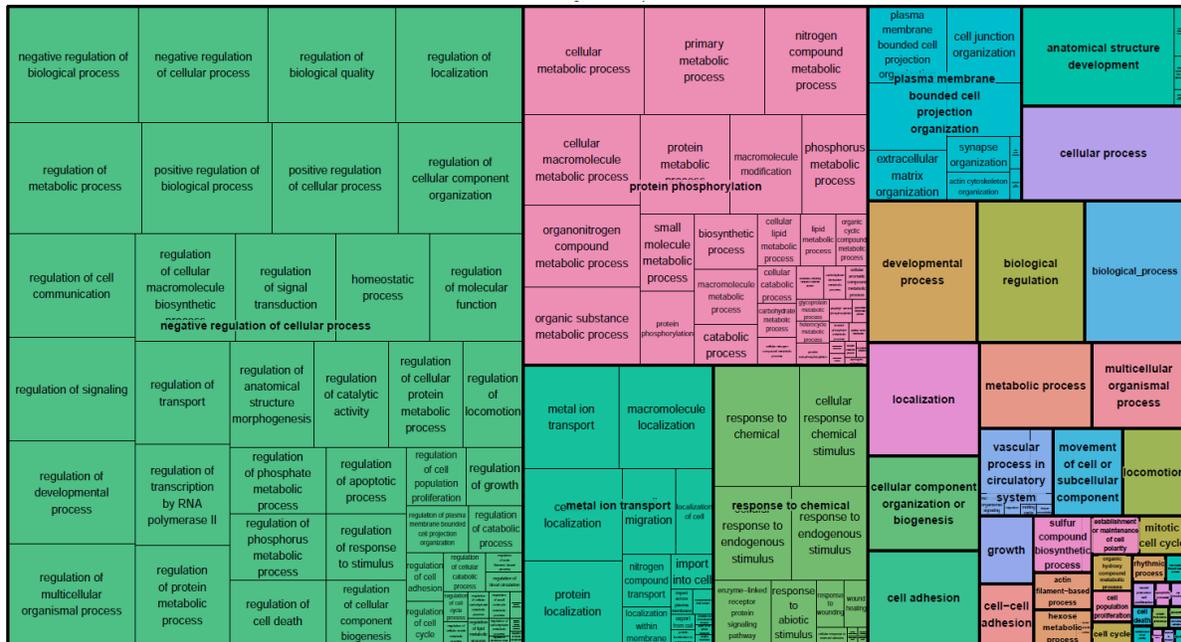
**Figura 11: Termos super-representados para alvos de tRFs em *H. sapiens*.** Representação em *Treemap* dos termos super-representados no conjunto de alvos de tRFs em *H. sapiens* após processo de simplificação pelo REVIGO.



**Figura 12: Termos super-representados para alvos de tRFs em *M. musculus*.** Representação em *Treemap* dos termos super-representados no conjunto de alvos de tRFs em *M. musculus* após processo de simplificação pelo REVIGO.



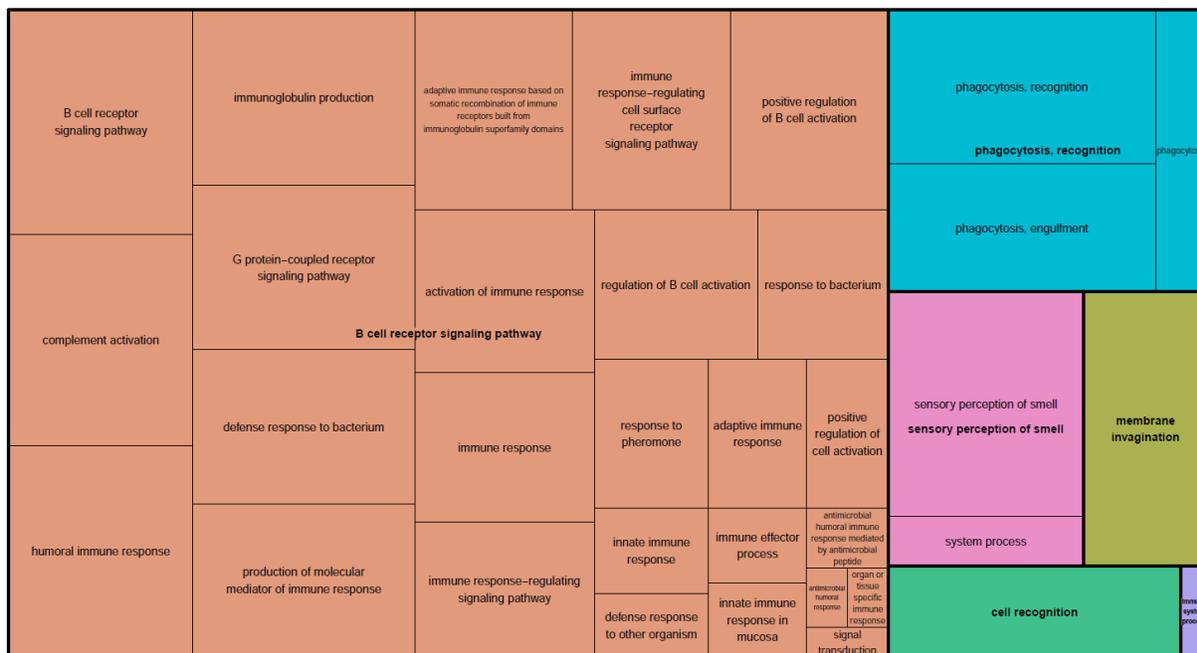
**Figura 13: Termos super-representados para alvos de miRNAs em *H. sapiens*.** Representação em *Treemap* dos termos super-representados no conjunto de alvos de miRNAs em *H. sapiens* após processo de simplificação pelo REVIGO.



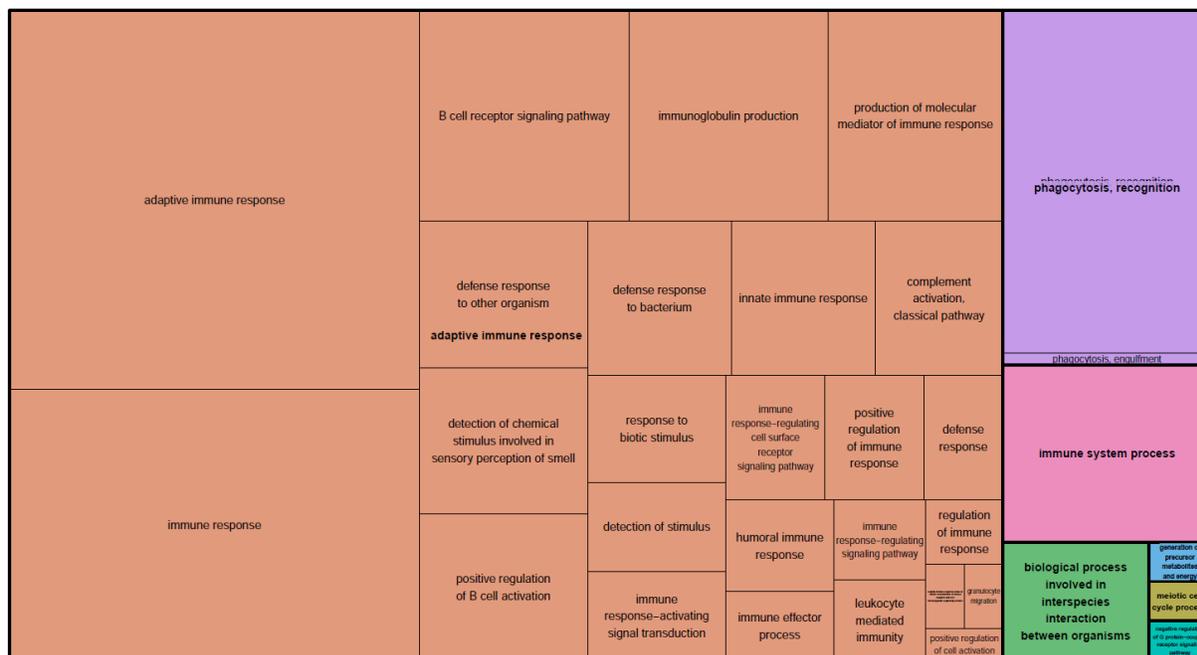
**Figura 14: Termos super-representados para alvos de miRNAs em *M. musculus*.** Representação em *Treemap* dos termos super-representados no conjunto de alvos de miRNAs em *M. musculus* após processo de simplificação pelo REVIGO.



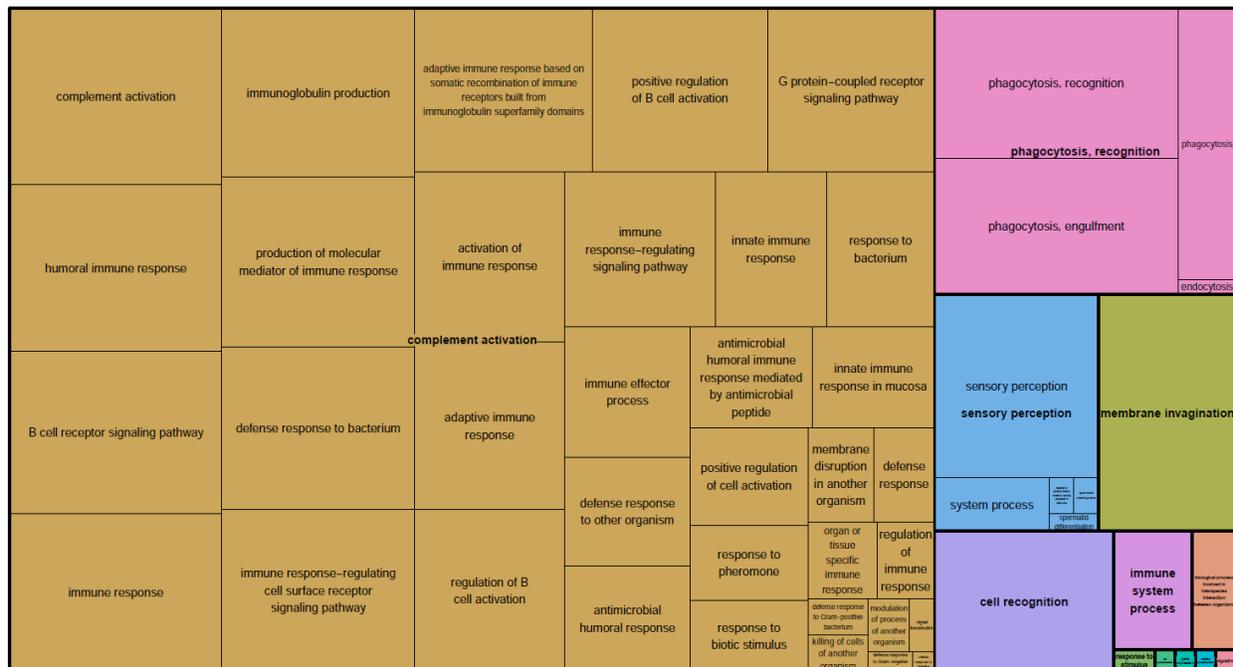
**Figura 15: Termos sub-representados para alvos de tRFs em *H. sapiens*.** Representação em *Treemap* dos termos sub-representados no conjunto de alvos de tRFs em *H. sapiens* após processo de simplificação pelo REVIGO.



**Figura 16: Termos sub-representados para alvos de tRFs em *M. musculus*.** Representação em *Treemap* dos termos sub-representados no conjunto de alvos de tRFs em *M. musculus* após processo de simplificação pelo REVIGO.



**Figura 17: Termos sub-representados para alvos de miRNAs em *H. sapiens*.** Representação em *Treemap* dos termos sub-representados no conjunto de alvos de miRNAs em *H. sapiens* após processo de simplificação pelo REVIGO.



**Figura 18: Termos sub-representados para alvos de miRNAs em *M. musculus*.** Representação em *Treemap* dos termos sub-representados no conjunto de alvos de miRNAs em *M. musculus* após processo de simplificação pelo REVIGO.

### 3.2.3. Discussão

Apesar da relevância demonstrada de sRNAs na regulação da expressão gênica durante infecções (ACUÑA; FLOETER-WINTER; MUXEL, 2020; CHABELSKAYA; GAILLOT; FELDEN, 2010; CROSTON et al., 2018; EISENBART et al., 2020; RAMAN et al., 2017), poucos estudos avaliaram diretamente a utilização de sRNAs de qualquer classe na interação patógeno-hospedeiro, deixando uma lacuna sobre a extensão da relevância de sRNAs. Além disso, entre os poucos estudos disponíveis nesse tema, até o presente momento, não existem estudos demonstrando a interação entre patógenos não-virais e hospedeiros por meio de tRFs. Dessa forma, a avaliação da presença de sRNAs regulatórios em vetores de transporte de macromoléculas com conhecido caráter de atuação inter-específica, como as vesículas extracelulares (MUNHOZ DA ROCHA et al., 2020; ROBBINS; MORELLI, 2014), representa um avanço no conhecimento da área.

Inicialmente, para a predição de tRFs nas vesículas extracelulares, utilizamos novas abordagens junto ao método desenvolvido e aplicado no Capítulo 1 (STREIT et al., 2021). Enquanto os passos iniciais de filtragem e predição de tRFs utilizando o tDRmapper não foram alterados, o passo final de filtragem de possíveis tRFs foi redesenhado, a fim de reduzir a subjetividade da análise e utilizar critérios mais condizentes com essa classe de sRNAs. Na versão anterior do *workflow* de predição de tRFs, os candidatos à tRFs preditos pelo tDRmapper tinham sua expressão avaliada em transcritos por milhão (TPM, do inglês *Transcripts per Million*), sendo em seguida aplicado um corte de 10 TPM para eliminar candidatos com baixo suporte de *reads*, que poderiam ser falsos-positivos. A utilização de um corte fixo, selecionado sem levar em consideração os dados apresenta arbitrariedade e está sujeito à subjetividade do analista dos dados, além de que o TPM, apesar de apresentar relação com a confiabilidade do resultado, não traduz necessariamente os pontos que caracterizam tRFs, como a produção precisa e em quantidades altas quando comparada com outras regiões do mesmo tRNA (ZHENG et al., 2016). Além disso, o aceite direto da região predita como tRF pelo tDRmapper como um tRF, sem levar em consideração a possibilidade de que a região predita seja composta por diversos *reads*, pode levar a falsos negativos, visto que em regiões compostas por múltiplos *reads* podem não apresentar *reads* cuja sequência compreenda a região como um todo em quantidades satisfatórias, sendo os *reads* majoritários da região correspondentes a uma sequência parcial da mesma. Assim, nossa nova abordagem consiste na busca de todas as possibilidades de tRFs dentro de uma região predita como tRF pelo tDRmapper, a fim de identificar o possível tRF com maior número de *reads* dentro da região e evitar falsos negativos, e a filtragem de falsos positivos pela porcentagem a qual representa o possível tRF com relação ao total de *reads* presentes na região predita como tRF pelo tDRmapper, que reflete a precisão do processamento, e a porcentagem com relação ao total de *reads* do tRNA como um todo, que reflete a alta quantidade de *reads* da região produtora de tRF com relação ao restante do tRNA. Por fim, os cortes utilizados são determinados de acordo com a mediana de ambos os valores, de maneira a dividir a população de possíveis tRFs em quadrantes, e eliminar o quadrante correspondente a

candidatos que apresentem os piores valores relacionados às características que definem tRFs dentro da população.

Utilizando essa abordagem, fomos capazes de identificar 71 tRFs em vesículas extracelulares, sendo uma ampla variedade de tRFs quando levada em conta a avaliação de apenas 4 linhagens da mesma espécie de levedura em apenas 2 condições distintas de cultivo. Além da predição de uma grande gama de tRFs, múltiplos tRFs foram preditos em mais de uma das bibliotecas que correspondiam às triplicatas da mesma linhagem e condição, de maneira a elevar a confiabilidade de nossos resultados e abordagem de predição. Assim como na conservação da predição entre triplicatas, os tRFs identificados apresentaram conservação variável entre as diferentes linhagens, com exemplos de tRFs com aparente conservação entre as linhagens, assim como de tRFs possivelmente linhagem-específicos. Essas variações na presença de tRFs entre diferentes linhagens podem apresentar implicações relevantes na interação patógeno hospedeiro, visto que por um lado, tRFs conservados podem apresentar sua conservação de produção entre diferentes linhagens devido a uma função essencial desenvolvida pelos mesmos, enquanto que tRFs linhagem-específicos podem compor os mecanismos responsáveis por diferenças de respostas e de virulência entre diferentes linhagens.

De maneira interessante, foi observada uma grande preferência pela produção de tRFs em condição de exposição à caspofungina para todas as linhagens avaliadas, sendo identificados tRFs em apenas 5 das 12 bibliotecas da condição controle. tRFs apresentam uma produção condição-específica (BAKOWSKA-ŻYWICKA et al., 2016), de maneira que a variação na identificação de tRFs entre condições controle e na interação com caspofungina pode ser explicada por uma adaptação das leveduras ao ambiente por meio da produção de um conjunto distinto de tRFs. Apesar disso, a baixa quantidade de tRFs, sendo até mesmo ausentes em vesículas da linhagem CDC385, é extremamente improvável, vista a presença praticamente ubíqua de tRFs nos mais diversos organismos e condições de cultivo (BAKOWSKA-ŻYWICKA et al., 2016; GEBETSBERGER et al., 2012; KUMAR et al., 2014; LUO et al., 2018; WANG et al., 2019), assim como seu papel na regulação de funções básicas das células (GEBETSBERGER et al., 2012; KEAM et al., 2017; KIM et al., 2017; LEE et al., 2009).

Dessa maneira, é possível hipotetizar que não apenas ocorra a produção diferencial de tRFs na presença de caspofungina, mas também alterações na seleção de macromoléculas a serem carregadas em vesículas extracelulares de maneira a permitir a entrada de tRFs nas mesmas. Porém, na ausência de análise do conteúdo intracelular de tRFs, é impossível afirmar se essa variação observada tem relação exclusivamente com a variação no conteúdo intracelular de tRFs, se ocorre variação na seleção de tRFs a serem carregados em vesículas, e se apenas alguns tRFs são preferencialmente carregados em vesículas ou é um carregamento indiscriminado.

Na predição de miRNAs vesiculares, foram identificados também diversos miRNAs, totalizando 102 miRNAs. Diferente de tRFs, os quais não foram observados em múltiplas bibliotecas de condição controle, os miRNAs foram identificados em todas as bibliotecas. Apesar disso, os miRNAs identificados em sua grande maioria não apresentaram suporte entre as triplicatas, reduzindo a confiabilidade dessa análise. Isso pode ocorrer devido ao fato de que a predição foi executada de maneira independente de um genoma e com uma ferramenta desenhada para miRNAs humanos. Enquanto diversas ferramentas de predição de miRNAs estão disponíveis para *H. sapiens*, ainda existe uma escassez de ferramentas computacionais para a predição de miRNAs em outros organismos, especialmente para miRNAs fúngicos, de maneira a dificultar a identificação dos últimos. Assim, a reduzida disponibilidade de abordagens de predição, somado ao fato de que o objetivo do presente estudo é a identificação de sRNAs com potencial para a atuação em células de hospedeiro mamífero e estes potencialmente apresentem características semelhantes aos sRNAs do hospedeiro para a utilização da maquinaria da célula hospedeira para gerar seus efeitos, torna a utilização de ferramentas para miRNAs humanos a escolha mais viável de predição apesar de suas limitações.

Além da identificação dos sRNAs em vesículas, diversos sRNAs com tamanhos condizentes com a atuação pela via de RNAi de metazoários foram identificados. Apesar de que a composição de nucleotídeos das extremidades dos sRNAs identificados não aparenta uma clara preferência posicional quando tratados em grandes grupos, à exceção da extremidade 3' CCA de tRFs-3, e que alguns dos grupos não o padrão de extremidade 5' como o de miRNAs de metazoários (BARTEL, 2018),

alguns padrões de interesse foram identificados. tRFs-3 apresentaram preferência moderada por T no primeiro nucleotídeo, semelhante a miRNAs e associado com carregamento em Argonautas (BARTEL, 2018), enquanto que tRFs-5 demonstraram uma preferência por G no primeiro nucleotídeo, semelhante a tRFs associados com Argonauta em *Arabidopsis thaliana* (LOSS-MORAIS; WATERHOUSE; MARGIS, 2013). Dessa forma, é possível que apesar de que tRFs em *C. auris* apresentem uma grande variação de características, subgrupos dentro de cada tipo de tRF podem apresentar uma composição de nucleotídeos que permitam a sua utilização de maneira mais otimizada pela via de RNAi canônica de hospedeiros metazoários.

A fim de melhor explorar o potencial de regulação de sRNAs de *C. auris* sobre hospedeiros mamíferos, foram identificados alvos para sRNAs de tamanho condizente com a utilização pela via de RNAi em *H. sapiens* e *M. musculus*. Apesar de que sRNAs em um intervalo de tamanho de 13 a 37 nucleotídeos foram identificados e os mesmos podem também atuar por mecanismos alternativos na regulação de células do hospedeiro (GEBETSBERGER et al., 2012; GOODARZI et al., 2015; KIM et al., 2017; MLECZKO; CELICHOWSKI; BAŁKOWSKA-ŻYWICKA, 2018), a predição da efetividade dos sRNAs por esses mecanismos seria de difícil execução, de maneira que objetivamos a identificação de atuação exclusivamente pela via de RNAi, ainda que dessa forma estejamos possivelmente subestimando o potencial dos sRNAs vesiculares. Com essa análise, identificamos diversos alvos para todos os sRNAs avaliados, e mesmo que seja provável que uma grande quantidade de alvos identificados sejam falsos-positivos dada a quantidade de alvos preditos por sRNA, é esperado que sRNAs com tamanho potencial de interação com transcritos do hospedeiro sejam capazes de gerar ao menos parte das interações propostas.

Além da provável capacidade de interação com transcritos hospedeiros, quase todos os sRNAs avaliados quanto ao enriquecimento funcional de seu conjunto de alvos apresentaram super representação de termos de ontologias gênicas. Apesar de que em sua maioria os termos estavam relacionados a processos de desenvolvimento do organismo e aparentemente não apresentam relevância para o processo de infecção, o que pode ser explicado dado o possível grande número de interações espúreas preditas para cada sRNA, alguns termos de relevância para o processo de

infecção, como termos relacionados à transcrição gênica, regulação do metabolismo celular e sinalização celular, também foram super representados em múltiplos sRNAs. A regulação da transcrição é a base para a adaptação celular de qualquer organismo, de maneira que células do sistema imunitário passam por alterações dinâmicas e específicas ao longo do processo de infecção de maneira a produzir respostas protetoras contra patógenos (BENOIT; DESNUES; MEGE, 2008; MUÑOZ et al., 2019; SHOHAM; LEVITZ, 2005). Da mesma forma que a transcrição, o metabolismo celular também é alterado durante infecções, de maneira a compensar a demanda energética requerida para controlar a infecção e a produzir intermediários do metabolismo com potencial na regulação imune (PELLON et al., 2022). Por fim, o processo de ativação de células do sistema imunitário e a comunicação desse sistema dependem de comunicações intercelulares que têm como base mecanismos de transdução de sinais por cascatas de sinalização (KRISHNAN et al., 2007). Dessa forma, os sRNAs identificados apresentam um potencial disruptivo para a homeostase celular que pode interferir na resposta imunitária. Além disso, mesmo que sub representados nas análises de enriquecimento funcional, foram identificados diversos termos relacionados diretamente com o sistema imunitário em múltiplos sRNAs, de maneira que além dos mecanismos citados acima, sRNAs vesiculares podem apresentar ainda interação com transcritos de importância direta para o funcionamento do sistema imunitário.

Apesar de que os resultados e hipóteses levantadas quanto à atuação dos sRNAs sobre transcritos de células hospedeiras sejam teóricas, requerendo ainda a confirmação experimental, eles reiteram a grande importância do estudo dos sRNAs na interface patógeno-hospedeiro. Não apenas sRNAs podem estar envolvidos em diversos mecanismos de evasão ao sistema imunitário, como um ponto crucial para o presente organismo é a associação dessa possibilidade com a demonstração de que alguns tRFs apresentam uma liberação em vesículas extracelulares mais elevadas durante o tratamento com antifúngico. Enquanto que é extremamente improvável que a produção de sRNAs com capacidade de alterar a homeostase de um hospedeiro mamífero seja uma resposta selecionada ao longo da evolução contra a exposição à caspofungina, dada a origem fúngica dessa classe de fármacos (EMRI et al., 2013), o estímulo à produção desses sRNAs por parte do fármaco pode, até certa extensão,

interferir com a eliminação da infecção por parte do fármaco. Além disso, enquanto que a caspofungina aparenta induzir a produção de sRNAs vesiculares, não apenas não sabemos a extensão de antifúngicos que possam produzir efeitos semelhantes, como também esse fenômeno pode ocorrer em situações estressoras presentes durante a interação patógeno-hospedeiro mesmo na ausência do fármaco, visto que não foi avaliado qual estímulo produzido pela caspofungina leva ao aumento na liberação de sRNAs em vesículas extracelulares. Assim, além de demonstrar a presença de sRNAs em vesículas extracelulares e seu potencial na regulação de transcritos de hospedeiros mamíferos, esse estudo ressalta a importância de um maior investimento na elucidação do papel de sRNAs na interação patógeno-hospedeiro durante infecções.

### 3.3. Capítulo 3: Desenvolvimento de *pipeline* para identificação de tRFs a partir de dados de sRNA-seq

Neste capítulo, relatamos o desenvolvimento e teste de um *pipeline* para identificação de tRFs baseada em análise de dados de RNA-seq e com amparo estatístico. Inicialmente, uma breve descrição do fluxo de trabalho deste *pipeline* é descrita. Após isso, variações na metodologia estatística proposta para a análise de tRFs existentes na literatura são propostas juntamente do racional para a implementação das mesmas em termos de acurácia e velocidade de execução. Por fim, as diferentes implementações de testes estatísticos associadas com a *pipeline* são testadas e comparadas com a ferramenta tRF2Cancer com dados simulados e biológicos.

### 3.3.1. Materiais e métodos

#### 3.3.1.1. Dados genômicos e de sequenciamento de sRNAs

Para os testes de avaliação de acurácia e tempo de execução tanto utilizando dados simulados quanto biológicos, foi utilizada a sequência de genoma de *Homo sapiens* GRCh38, sob o acesso do RefSeq GCF\_000001405.26, no formato fasta, a fim de utilizar a mesma versão o genoma humano utilizado pela ferramenta tsRFinder, pertencente ao pacote tsRFun. Para os testes com dados de sequenciamento de origem biológica, foi utilizada a biblioteca de *reads* de sRNAs humanos sob o acesso SRR3235777 do SRA, sendo removidas sequências de adaptadores previamente às análises.

Para a geração de bibliotecas simuladas de *reads* de sRNA com distribuição aleatória, um código em Python foi desenvolvido. Brevemente, foi recuperado o genoma de *H. sapiens* GRCh38 e fornecido ao código de geração de *reads* aleatórios, assim como o intervalo de tamanho de *reads* a ser utilizado de 14 a 30. Para cada sequência de tRNA, o código seleciona um número de *reads* aleatório entre 0 e 20.000 para ser gerada para aquele tRNA. Durante a geração de cada *read*, um tamanho aleatório é selecionado entre 14 e 30, sendo que após a seleção do tamanho são geradas todas as possibilidades de *reads* com a sequência do tRNA e o tamanho selecionado, e uma das possibilidades é selecionada aleatoriamente. Por fim, após a seleção de uma sequência aleatória, é gerada uma linha de qualidade com pontuação em Phred +33 do mesmo tamanho da sequência, contendo valores aleatórios entre 30 e 40 para cada nucleotídeo. Um cabeçalho identificador é atribuído ao *read*, sendo ele organizado em conjunto com a sequência e qualidade para a formação de um *read*.

#### 3.3.1.2. Ambiente de desenvolvimento e teste do detectRF

DetectRF foi desenvolvido em Python 3.9.7 utilizando o ambiente de desenvolvimento integrado sublime, em um *laptop* com sistema Manjaro Linux 21.0.1 “Ornara”, processador Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz com 12 núcleos e 16GB de RAM DDR4. Suas dependências e versões testadas são tRNAscan-SE (v2.0.9), bedtools (v2.30.0), bowtie (v1.3.1), bowtie2 (v2.4.4), samtools (v1.14),

biopython (v1.79), scipy (v1.7.3), estando as versões apresentadas disponíveis pelo gerenciador de *software* conda (v.4.11.0), o qual foi utilizado durante o desenvolvimento da presente *pipeline*.

### 3.3.1.3. Comparação de acurácia

Para todos os testes de acurácia, foi utilizado um servidor com sistema operacional Oracle Linux 8.4x86\_64, com 2 processadores AMD EPYC 7502 e 512GB de memória RAM DDR4, com o gerenciador conda para a manutenção e instalação dos ambientes e dependências dos *software* testados. As diferentes versões do detectRF foram executadas utilizando como entrada a biblioteca de *reads* em formato fastq, de acordo com o teste (dados simulados ou biológicos), e a sequência genômica, juntamente dos parâmetros “-n 100”, “--mit-model mamal” e “-q”. Para a execução do tsRFinder, uma versão modificada que permite sua execução localmente do *script* em perl da ferramenta foi utilizada, sendo executada em ambiente do conda no mesmo computador citado acima contendo as dependências: perl (v5.22.2.1), bowtie (v1.1.1), bedtools (v.2.25.0) e samtools (v1.9). Pacotes do perl necessários foram instalados pelo cpan com o ambiente conda ativo. Para a predição de tRFs pelo tsRFinder foi fornecida a mesma biblioteca de *reads* fornecida ao detectRF, de acordo com o teste, e o *index* próprio do tsRFinder, além dos parâmetros “-t 100” e “-p 0.01”.

Para as análises de dados aleatórios, foi considerado como padrão ouro a não identificação de tRFs, enquanto que para as análises de dados biológicos, foi levado em consideração o número de tRFs identificados com demonstração experimental e o número total de tRFs identificados. O tempo de execução foi monitorado utilizando o comando time de sistemas Linux, levando em consideração o tempo real de execução.

Para a verificação da presença de sequência idênticas aos tRFs utilizados como padrão para as análises de acurácia com dados humanos, foi gerado um alinhamento da sequência desses tRFs no formato fasta contra o genoma GRCh38 após a exclusão das sequências de tRNAs do mesmo, sendo o alinhamento executado com a ferramenta bowtie2, no ambiente conda utilizado pelo detectRF. O arquivo sam resultante foi avaliado quanto à presença da *flag* NM:i:0, que indica a ausência de *gaps*, inserções e *mismatches* no alinhamento entre as sequências.

### 3.3.2. Resultados

#### 3.3.2.1. Fluxo de funcionamento do *pipeline* detectRF

O fluxograma geral de operação deste *pipeline* é apresentado na Figura 19. O primeiro passo no detectRF é a entrada de arquivos e parâmetros por parte do usuário, conforme a lista de argumentos descrita na Tabela 1, sendo essa executada por meio de linha de comando em terminal. Como parâmetros obrigatórios fornecidos pelo usuário (Figura 19A), encontram-se apenas o apontamento dos arquivos de genoma, no formato fasta, e da biblioteca de *reads* de sRNA propriamente processada, no formato fastq. Após a entrada dos arquivos e parâmetros, o detectRF cria seu diretório de arquivos temporários para a execução de análises no local de execução e, caso requerido pelo usuário, verifica a presença de *reads* com cabeçalhos repetidos, que podem levar a erros na execução.

Como segundo passo (Figura 19B), o detectRF executa a busca por tRNAs na sequência genômica utilizando a dependência tRNAscan-SE, sendo essa predição utilizada para a geração das sequências de tRNA maduro, transcrito de tRNA primário, e regiões exclusivas de transcritos de tRNA primário (*trailer*, *leader* e íntrons, separados e sem sequências presentes em tRNAs maduros). Alternativamente ao passo de predição de tRNAs por meio do tRNAscan-SE, o usuário tem a opção de fornecer sua própria anotação de genes de tRNA em formato bed, restrito apenas por uma formatação condizente com arquivos bed obtidos da gtRNAdb . Após a obtenção das sequências das regiões do genoma citadas utilizando a dependência bedtools, as modificações necessárias em tRNAs maduras são adicionadas, como a sequência CCA na extremidade 3' de tRNAs que não apresentem essa sequência em sua extremidade diretamente de sua transcrição e o nucleotídeo G na extremidade 5' de tRNAs de histidina. Após as modificações, são removidos tRNAs maduros redundantes, mantendo apenas uma das sequências idênticas como representativa. Por fim, para o passo de filtragem de *reads* não pertencentes a tRNAs, um genoma livre de sequências de tRNA, produzido pela exclusão das coordenadas de tRNA primário, é gerado.

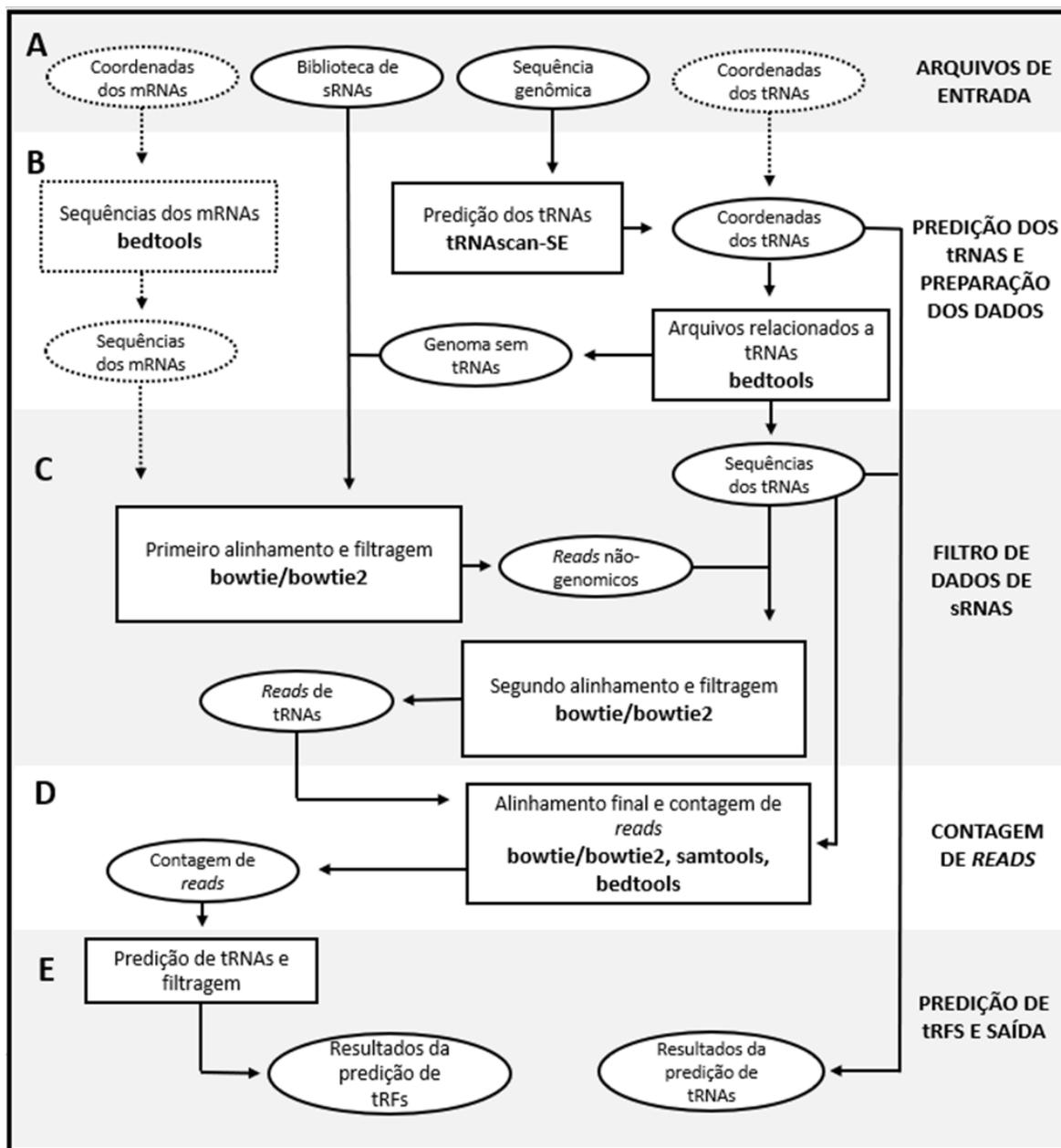
Durante o terceiro passo (Figura 19C), diversas filtrações da biblioteca de *reads* são executadas. Na primeira filtração, os *reads* são alinhados contra o genoma livre de

sequências de tRNA utilizando a dependência bowtie2 com as opções “--end-to-end”, “--very-sensitive” e “--non-deterministic”, conforme utilizado em nosso fluxo de processamento desenvolvido no capítulo 1. Alternativamente, baseado na avaliação de alinhadores de *reads* para dados de miRNAs executada por de Ziemann *et al.* 2016 (ZIEMANN; KASPI; EL-OSTA, 2016), o usuário pode selecionar como método de alinhamento a dependência bowtie com as opções “--best”, “--strata”, “-k 1” e “-M 1”, sendo uma adaptação do método com melhor desempenho para o alinhamento de miRNAs. Os *reads* que apresentarem alinhamentos válidos contra o genoma, sendo considerado válido o alinhamento que não apresente um número de *mismatches* e *gaps* maior que o valor de corte padrão de 0 ou valor fornecido pelo usuário, são excluídos da análise, sendo recuperados os *reads* que apresentaram alinhamentos inválidos contra o genoma. Nesse ponto, pode ser executada uma filtragem opcional contra sequências de mRNA, se fornecidas em formato fasta pelo usuário, sendo aplicado os mesmos alinhadores e racional da filtragem contra o genoma. Os *reads* resultantes dos passos de filtragem anterior (*reads* com alinhamentos inválidos contra genome e, opcionalmente, mRNAs), são então alinhados contra as sequências de tRNA maduro utilizando os mesmos alinhadores utilizados previamente, sendo porém armazenados para a análise de tRFs em tRNAs maduros os *reads* que apresentarem alinhamentos válidos contra essas sequências. Os *reads* com alinhamentos inválidos a tRNAs maduros são, em seguida, alinhados contra as sequências exclusivas de transcrito de tRNA primário, sendo utilizados os mesmos alinhadores e racional que para tRNAs maduros, sendo armazenados os *reads* com alinhamentos válidos contra essas sequências. Durante o armazenamento dos *reads* nas filtrações contra tRNAs maduros e sequências exclusivas de tRNA primário, os *reads* são separados em bibliotecas de acordo com o tipo de sequência (madura ou primária) e de acordo com seu tamanho, sendo essa separação necessária para a execução das análises estatísticas. Caso o intervalo de tamanho de *reads* esperado seja fornecido pelo usuário, durante a separação por tamanho, *reads* que apresentem tamanho fora desse intervalo são removidos por serem possivelmente produtos de um processamento incorreto, enquanto que na ausência dessa delimitação de intervalo, todos os tamanhos de *reads* são aceitos.

No quarto passo (Figura 19D), ocorre a obtenção da contagem de *reads* requerida para a execução da estatística do *pipeline*. Cada biblioteca gerada até então é individualmente alinhada contra as suas respectivas sequências (tRNA maduro ou sequência exclusiva de transcrito de tRNA primário), utilizando a mesma estratégia de alinhamento selecionada para os passos de filtragem. A partir do resultado do alinhamento, o programa extrai a contagem de *reads* total de cada tRNA, também executando as dependências samtools e bedtools genomcov para a obtenção da cobertura de *reads* para cada nucleotídeo de cada tRNA.

No quinto passo (Figura 19E), a partir das contagens de *reads* obtidas em cada alinhamento, o detectRF faz a identificação de tRFs pela busca de regiões cuja contagem de *reads* fuja de uma distribuição aleatória, utilizando para isso uma versão revisada do método estatístico aplicado pelo tsRfun. Inicialmente, para cada tRNA a ser analisado, a chance de alinhamento de um *read* qualquer de tamanho igual ao tamanho dos *reads* da biblioteca analisada contra cada nucleotídeo do tRNA é calculada. Após o cálculo da chance de alinhamento de um *read* a cada nucleotídeo do tRNA, esse valor, somado aos valores de cobertura observados no alinhamento de *reads* final é utilizada para a identificação de nucleotídeos cujas coberturas de *reads* fogem de uma distribuição aleatória, sendo 0.01 o valor de corte da probabilidade de uma contagem de *reads* ser observada em uma distribuição aleatória na execução padrão. Dada a identificação de uma região de nucleotídeos consecutivos de tamanho igual ou maior que o tamanho do *read* analisado, o número de *reads* que correspondem à região de nucleotídeos consecutivos é recuperada, normalizada pelo número de ocorrências dessa sequência dentro do tRNA, e comparada com o valor de corte para a identificação de nucleotídeos significativos. Caso a região seja maior que o tamanho de *read* analisado, a região é subdividida em regiões sobrepostas de tamanho igual ao tamanho do *read* para essa análise. Caso a região apresente um número de *reads* acima do valor de corte, a mesma é considerada um tRF. Após a avaliação de todos os tRNAs em todas as bibliotecas, o conjunto de tRFs identificados é reportado em formato fasta e bed, juntamente de um arquivo tsv contendo as informações sobre os tRFs identificados. Adicionalmente, também são reportadas as sequências de tRNA

identificadas, maduras e primárias, em formato fasta, assim como arquivos gerados durante a predição de tRNAs pelo tRNAscan-SE caso requerido pelo usuário.



**Figura 19: Fluxo de operação do detectRF.** Representação simplificada dos principais passos da predição de tRFs executada pelo detectRF, conforme detalhado no texto. Elipses representam o tipo de dado gerado/analísado enquanto retângulos representam a execução de programas e análises. Linhas pontilhadas representam passos opcionais. Letras em negrito dentro das elipses e retângulos apresentam os programas utilizados para esses passos.

**Tabela 1: Parâmetros disponíveis no detectRF.** Apresentação de todos os parâmetros disponíveis relativos às análises no detectRF, assim como uma descrição simplificada de cada parâmetro.

Parâmetro	Descrição
-g	Arquivo contendo a sequência genômica no formato fasta.
-r	Arquivo contendo a biblioteca de <i>reads</i> a ser analisada em formato fastq, estando a mesma previamente processada.
-o	Padrão de nome para a saída. Caso não seja provido um padrão, a data, hora e nome do arquivo do genoma são utilizados para a geração de um padrão de nome para a saída.
--srna-len	Tamanhos mínimo e máximo de sRNA utilizados para a construção da biblioteca de sRNAs. Caso não sejam providos, o programa utilizará o tamanho mínimo e máximo de <i>read</i> identificado.
--mode	Modo de predição de tRFs do tRNAscan, podendo ser sensível ou específico.
-p	Valor de p a ser utilizado como valor de corte nos testes estatísticos de predição de tRFs. Caso não seja provido, o valor utilizado é de 0.01
-n	Números de núcleos de processamento a serem utilizados na análise. Caso não seja provido, todos os núcleos à excessão de um serão utilizados.
--check-fastq	O programa revisa a biblioteca de <i>reads</i> para o usuário para avaliar a presença de <i>reads</i> com cabeçalhos duplicados, que podem causar problemas na análise.
--trna-pred	Arquivo contendo as coordenadas de genes de tRNA no formato BED.
--trna-search	Modo de predição a ser utilizado pela dependência tRNAscan-SE, sendo as opções eucariótico, bacteriano, arqueal, mitocondrial de mamífero, mitocondrial de vertebrado, outras organelas e geral.
--mit-model	Uso de modelos de tRNA mitocondrial pelo tRNAscan-SE para a diferenciação entre tRNAs mitocondriais e citosólicos, estando disponível apenas para vertebrados ou especificamente mamíferos.
--pred-alg	Algoritmo de predição aplicado pelo tRNAscan-SE, sendo as opções Infernal, máxima sensibilidade, legado e COVE.
--no-pseudo-trna	Remove tRNAs preditos como pseudo-genes da análise.
--no-cca-addition	Não adiciona CCA na extremidade 3' de tRNAs que não apresentem essa sequência a partir da transcrição.
--no-g-addition	Não adiciona G na extremidade 5' de tRNAs de Histidina.
--leader-size	Número de bases à montante do tRNA a serem consideradas a região líder do tRNA. Por padrão, é utilizado um valor 40.
--trailer-size	Número de bases à jusante do tRNA a serem consideradas a região líder do tRNA. Por padrão, é utilizado um valor 40.
--store-files	Armazena os arquivos gerados pelo tRNAscan-SE para serem reportados juntamente aos resultados do detectRF ao fim da corrida.
--aln-mode	Ferramenta de alinhamento de <i>reads</i> e opções a serem utilizadas pelo detectRF, podendo ser utilizado o bowtie ou bowtie2.
--phred	Tipo de pontuação Phred utilizada na biblioteca de <i>reads</i> .
--mrna-filter	Arquivo contendo sequências de RNAs mensageiros a serem utilizados como filtragem opcional.
--mismatch-genome	Número de <i>mismatches</i> permitidos durante a filtragem de <i>reads</i> alinhados contra o genoma. Por padrão, é utilizado o valor 0.
--gap-genome	Número de <i>gaps</i> permitidos durante a filtragem de <i>reads</i> alinhados contra o genoma. Por padrão, é utilizado o valor 0.
--mismatch-trna	Número de <i>mismatches</i> permitidos durante a filtragem de <i>reads</i> alinhados contra os tRNAs. Por padrão, é utilizado o valor 0.
--gap-trna	Número de <i>gaps</i> permitidos durante a filtragem de <i>reads</i> alinhados contra os tRNAs. Por padrão, é utilizado o valor 0.

### 3.3.2.2. Estatísticas de avaliação de contagens aleatórias e algoritmos de aceleração e acurácia

Além de nossa motivação em desenvolver uma ferramenta passível de ser executada para qualquer organismo com facilidade, dois pontos de grande relevância são a acurácia e o tempo de execução. Assim, foi feita uma análise da estatística aplicada pelo tsRFun para a predição de tRFs, utilizada como base para esse programa, a fim de identificar pontos a serem otimizados quanto aos cálculos utilizados para a identificação de regiões com contagens de *reads* fora de uma distribuição aleatória dentro dos tRNAs.

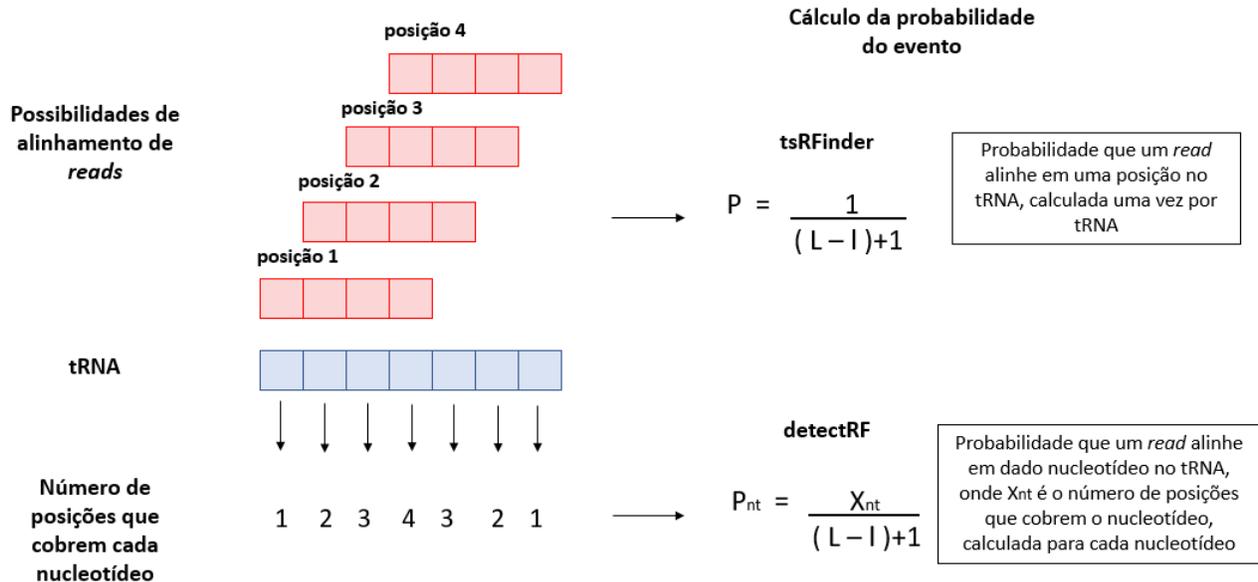
O ponto base para o cálculo estatístico que visa distinguir se a contagem de *reads* observada em um dado nucleotídeo do tRNA é ou não esperada em uma distribuição aleatória é o cálculo da chance de que um *read* cubra um nucleotídeo. A abordagem proposta no tsRFinder calcula essa chance como

$$P = \frac{1}{(L-l)+1} \quad (1)$$

, sendo L o tamanho do tRNA e l o tamanho dos *reads* analisados, atribuindo a todos os nucleotídeos do tRNA a mesma chance de possuir um *read* alinhado a eles. Enquanto esse cálculo retrata com precisão a chance do alinhamento de um *read* em diferentes posições do tRNA, ele não condiz com a chance do alinhamento de um *read* em diferentes nucleotídeos, visto que posições diferentes podem cobrir o mesmo nucleotídeo (Figura 20). Dessa maneira, a atribuição de que apenas 1 das possibilidades de alinhamento possa cobrir um nucleotídeo dentro do universo de possibilidades de alinhamento apresentado na equação (1) gera uma subestimativa da chance de alinhamento de um *read* contra um nucleotídeo para todos os nucleotídeos que não sejam os da extremidade do tRNA, podendo levar a falsos-positivos. Como alternativa, propomos uma extensão desse cálculo (Figura 20), ainda utilizando o mesmo cálculo para o universo e possibilidades de alinhamento

$$P = \frac{Xnt}{(L-l)+1} \quad (2)$$

, levando porém em consideração o número de possibilidades de alinhamento que cobrem cada nucleotídeo ( $X_{nt}$ ) e, conseqüentemente, dando para diferentes nucleotídeos, diferentes chances de alinhamento de um *read* sobre os mesmos.



**Figura 20: Bases do cálculo da chance de alinhamento de um read sobre um nucleotídeo.**

À esquerda, representação das possibilidades de alinhamento de um dado *read* (vermelho, nucleotídeos representados por caixas) em um dado tRNA (azul, nucleotídeos representados por caixas) sem consideração por seqüências. O *read* em questão apresenta 4 possibilidades de alinhamento dado seu tamanho, sendo o número de possibilidade de alinhamento obtido pelo cálculo da subtração do tamanho do tRNA ( $L$ ) pelo tamanho do *read* ( $l$ ), somado à 1. Assim, em uma distribuição aleatória, a probabilidade de que uma posição seja ocupada por um *read* ( $P$ ) para todas as posições é de 1 dividido pelo número de possibilidades de alinhamento, sendo esse cálculo aplicado pelo tsRFinder. Porém, dado que cada posição de alinhamento de *read* se estende por diversos nucleotídeos, cada nucleotídeo apresenta um conjunto diferente de posições que os cobrem. Dessa maneira, para avaliar a probabilidade de que um nucleotídeo seja ocupado por um *read* ( $P_{nt}$ ), detectRF leva em consideração o número de posições de alinhamento que cobrem aquele nucleotídeo ( $X_{nt}$ ), pois mesmo que a probabilidade de alinhamento entre as posições sejam iguais, nucleotídeos com maior número de posições que o contemplem apresentam maior probabilidade de terem *reads* alinhados à si.

Outro ponto avaliado foi o cálculo utilizado para a identificação de nucleotídeos com contagem de *reads* cuja probabilidade de ocorrência em uma distribuição aleatória seja mínima. Para isso, o tsRFinder utiliza um cálculo de probabilidade da observação

da contagem de *reads*  $k$  ou contagens maiores, dada a probabilidade de alinhamento de um *read* contra o nucleotídeo em questão ( $p$ ) e o número total de *reads* alinhados contra o tRNA ( $x$ ).

$$P(X \geq k) = \sum_{x=k}^n \binom{n}{x} p^x (1-p)^{n-x} \quad (3)$$

Enquanto que o cálculo da probabilidade de  $k$  ou mais *reads* remove o problema da identificação de probabilidades baixas relacionadas com contagens de *reads* baixas demais, esse cálculo não apenas avalia um valor não observado (avaliando o valor observado e valores não observados), e envolve a repetição de cálculos do coeficiente binomial entre  $k$  e  $x$  dado o somatório, o que é custoso computacionalmente. Dessa maneira, propusemos o cálculo direto da probabilidade do valor  $k$ , sem a consideração de valores não observados, sendo necessário apenas o cálculo de um coeficiente binomial e reduzindo o custo computacional.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (4)$$

Além disso, dado que múltiplos nucleotídeos dentro do mesmo tRNA apresentam a mesma chance de que um *read* alinhe contra os mesmos, ao invés de calcularmos a probabilidade para cada valor observado em cada nucleotídeo, é calculado o menor valor de  $k$  com probabilidade de ocorrência  $< 0.01$  para cada probabilidade de alinhamento de *read* contra nucleotídeo do tRNA, que corresponde a múltiplos nucleotídeos. Após o cálculo do valor  $k$  de corte para cada chance de alinhamento de *read* contra nucleotídeo, os valores observados de cada nucleotídeo são comparados com seu respectivo valor  $k$  de corte para a identificação de valores  $k$  significativos.

A abordagem da identificação de valores corte de  $k$  também permite um novo passo de filtragem. Enquanto avaliando o número de *reads* alinhados contra um dado nucleotídeo, não é considerada a posição de alinhamento de cada *read*, de maneira que uma alta contagem de *reads* sobre uma região não significa que aquela região seja composta por diversas cópias do mesmo *read*, padrão esperado para tRFs, podendo ser uma cobertura originada a partir de *reads* alinhados em diferentes regiões próximas, o que não condiz com tRFs. O tsRFinder, após a identificação de regiões de nucleotídeos significativos consecutivos avalia se os *reads* correspondendo à região

identificada representam 60% ou mais da contagem observada para a região. A utilização de critérios subjetivos e sem base estatística como um valor de corte em porcentagem, sem consideração ao valor total de *reads* ou os valores associados ao que é considerado aleatório em termos de contagem para o tRNA, reduz consideravelmente a confiança obtida pela análise estatística anterior. Visto que nossa abordagem identifica valores de cobertura durante a identificação dos nucleotídeos significativos, aplicamos esses valores para a filtragem da cobertura das regiões. Dessa forma, dado um conjunto de nucleotídeos com contagens significativas sequenciais, os valores de corte  $k$  utilizados para a identificação de significância nesse conjunto são recuperados, e o maior valor de  $k$  dentro dessa região é utilizado como valor de corte para a contagem de *reads* correspondente à região. Alternativamente, para maior sensibilidade, o usuário pode selecionar a utilização do menor valor de  $k$  da região para essa filtragem.

Visto que o cálculo do valor  $k$  de corte é uma varredura do universo de valores de  $k$  possíveis dentro do número total de *reads* alinhados, o cálculo da probabilidade de todos os valores de  $k$  de  $x$  a 0 até a identificação do menor valor com probabilidade menor que 0.01 é exaustivo e custoso computacionalmente. Assim, implementamos um algoritmo para a aceleração da varredura dos valores de  $k$  que consiste na avaliação da probabilidade associada ao valor de  $k$  entre os dois valores de  $k$  extremos. Inicialmente, a probabilidade de  $k=x$  é calculada, de maneira que  $x$  seja tomado como valor limite superior de  $k$  caso o mesmo apresente uma probabilidade de ocorrer menor que 0.01, ou o nucleotídeo não seja analisado caso a probabilidade seja superior a 0.01. Caso a probabilidade acima seja menor que 0.01,  $x$  é tomado como último valor válido, e é calculada média dos valores de  $k$  da distribuição binomial multiplicando o número total de *reads* pela probabilidade de alinhamento de um *read* contra um nucleotídeo. A média é então tomada como o valor limite inferior de  $k$ , visto que valores abaixo do mesmo nunca podem apresentar uma probabilidade significativa associada com valores de  $k$  acima do esperado. Após isso, a média entre os valores limites é calculada e arredondada para o número inteiro mais próximo (podendo ser acima ou abaixo dependendo se a probabilidade identificada é maior ou menor que a probabilidade de corte definida, respectivamente), se necessário, sendo então avaliada

a probabilidade de ocorrência do valor da média. Caso o valor apresente uma probabilidade de ocorrência maior que 0.01, ele se torna o novo valor limite inferior, enquanto que se apresentar uma probabilidade de ocorrência menor que 0.01, ele se torna o novo valor limite superior e é tomado como o último valor válido. Uma nova média é calculada entre os limites e avaliada, sendo esse processo repetido até que a média entre os limites retorne ao último valor válido, e esse seja tomado como o valor  $k$  de corte.

Por fim, visto o custo computacional do cálculo de probabilidades binomiais, uma segunda versão do detectRF com a substituição desse mesmo cálculo por uma aproximação foi gerada para fins de comparação de acurácia e tempo de execução. Enquanto que aproximações são associadas com a perda de acurácia no cálculo, a necessidade de execução da ferramenta em tempos hábeis para conjuntos de dados grandes é uma extrema necessidade na bioinformática moderna. Assim, utilizamos o cálculo proposto por Arratia e Gordon (ARRATIA; GORDON, 1989) como substituto exclusivamente ao cálculo do valor  $k$  de corte, sem alteração dos demais cálculos e filtragens, visto que essa aproximação apresentou um resultado aproximadamente 2,5% acima do valor calculado pelo cálculo da probabilidade binomial no exemplo utilizado pelos autores.

#### 3.3.2.3. Avaliação de acurácia e tempo de execução

Para a avaliação da acurácia e tempo de execução do detectRF, a versão utilizando o cálculo padrão de distribuição binomial, assim como a versão com a implementação da aproximação acima citada, foram testadas quanto à capacidade de predição de tRFs conhecidos em dados experimentais e à não-predição de tRFs em dados simulados de distribuição aleatória. Os dados experimentais humanos são os mesmos utilizados pelo tsRFun para avaliação de acurácia, dado que a linhagem celular e condição dessa biblioteca condiz com o utilizado para a demonstração experimental de 17 tRFs, sendo o mais próximo de um *benchmark* disponível para esse tipo de dado. Além disso, dada a necessidade de que a ferramenta seja capaz de discernir entre dados com distribuição aleatória ou que desviem desta, também foram gerados dados simulados a partir da sequência dos tRNAs utilizados pelo tsRFun que

mimetizassem a distribuição de *reads* esperada em tRNAs com presença de fragmentos de degradação.

Para os dados humanos (Tabela 2), o detectRF com o cálculo binomial foi capaz de prever apenas 3 dos 17 tRFs com demonstração experimental, sendo no total preditos 671 tRFs. Comparativamente, enquanto a versão *online* do tsRFfinder tem reportada a predição de 10 dos 17 tRFs padrão e um total de 220 tRFs preditos, apresentando tanto maior especificidade quanto sensibilidade teoricamente, a versão local do tsRFfinder foi capaz de prever apenas 8 dos 17 tRFs padrão, e totalizando 1044 tRFs preditos. Apesar disso, quando avaliada a presença de sequência idênticas aos tRFs padrão em regiões do genoma que não tRNAs, 4 apresentaram um alinhamento perfeito contra essas regiões utilizando bowtie2, de maneira que apenas 13 dos 17 tRFs padrão poderiam ser preditos com alta confiança como tRFs baseado exclusivamente nos dados de sRNA-seq, e não configurariam um bom padrão para essa avaliação. Para esse teste, o tempo de execução foi de 106 minutos e 47 segundos para o detectRF em paralelização com 100 núcleos de processamento, enquanto que o tsRFfinder apresentou um tempo de execução de 31 segundos. A fim de verificar a sobreposição das predições, foi feita a remoção da redundância de tRFs com a mesma sequência para cada predição para comparação. Enquanto que o número de tRFs não redundantes entre os programas foi similar, com 425 tRFs únicos preditos para o detectRF e 399 para o tsRFfinder, apenas 83 tRFs com a exata mesma sequência foram identificados entre as duas predições, evidenciando uma grande divergência de resultados. Infelizmente, devido à ausência de um conjunto de dados confirmados mais extensos, é impossível avaliar qual conjunto de tRFs preditos apresenta uma maior acurácia.

**Tabela 2: Resultados da predição de tRFs com detectRF e tsRFfinder.** Comparação dos resultados e tempo de execução do detectRF e tsRFfinder em dados experimentais de *H. sapiens* e dados aleatórios simulados. NA não se aplica. Tempo de execução em minutos (min) e segundos (s).

	Programa	tRFs totais	tRFs únicos	tRFs padrão	Tempo de execução
Dados humanos	detectRF	671	425	3/17	106min47s
	tsRFfinder	1044	399	8/17	31sec
Dados aleatórios	detectRF	210	154	NA	266min28s
	tsRFfinder	668	246	NA	2min45s

Para os dados simulados de distribuição aleatória (Tabela 2), enquanto que o detectRF fez a predição de 210 tRFs falsos-positivo, o tsRFinder identificou 668 tRFs falsos-positivo. Mesmo após a remoção de redundância, detectRF apresentou uma quantidade menor de tRFs falso-positivos preditos, com um total de 154 tRFs não-redundantes contra 246 tRFs não-redundantes do tsRFinder. O tempo de execução do detectRF foi de 266 minutos e 28 segundos, enquanto o tsRFinder teve um tempo de execução de 2 minutos e 45 segundos. Em nenhum dos testes envolvendo a análise do genoma humano a versão do detectRF utilizando a aproximação foi capaz de concluir a execução, devido a problemas gerados durante os cálculos da estatística, ainda a serem analisados quanto ao motivo.

### 3.3.3. Discussão

A predição de tRFs a partir de dados *in silico* com alta confiabilidade é uma tarefa complexa, vista a necessidade da distinção entre os mesmos e produtos de degradação dos tRNAs, assim como, conforme demonstrado na análise das características dos tRFs identificados no capítulo 2, tRFs não necessariamente apresentam grande conservação de características. Além disso, o fato de serem produzidos a partir de outras moléculas de RNA altamente estruturadas e com papel central no funcionamento celular faz com que os tRFs apresentem uma relação mais direta com as características esperadas do seu tRNA parental do que com os demais tRFs produzidos pela célula. Assim, a utilização de uma característica não-associada à sequência ou região de produção, como a identificação de regiões de tRNAs com quantidades de *reads* fora do esperado na ausência de uma produção ativa de tRFs, é a alternativa mais viável para uma predição organismo-independente.

Esse tipo de avaliação já é executada por diferentes ferramentas (LOHER; TELONIS; RIGOUTSOS, 2017; SELITSKY; SETHUPATHY, 2015; SHI et al., 2018), porém apenas o tsRFun (WANG et al., 2022) apresenta um teste robusto para identificar regiões nos quais a quantidade de *reads* se encontra anormalmente elevada, tendo como base a idéia de que os fragmentos de degradação de tRNA apresentem uma produção aleatória. Enquanto que o tsRFun é bem-sucedido na predição de tRFs em *H. sapiens*, a metodologia empregada por essa ferramenta apresenta um potencial

inexplorado para a busca por tRFs em qualquer organismo. Assim, estendendo os cálculos estatísticos propostos pelo tsRFun, foi desenvolvido o detectRF, com o objetivo de permitir a qualquer usuário fazer predições acuradas e com suporte estatístico de tRFs para qualquer organismo.

Dada a proposta de acessibilidade do detectRF, diversos pontos de seu desenvolvimento foram idealizados de maneira a permitir um fácil uso tanto em termos de instalação e parametrização, quanto de arquivos e preparação necessária dos mesmos para a execução. O detectRF foi desenvolvido utilizando ferramentas em versões disponíveis pelo gerenciador de ambientes conda, de maneira que permita ao usuário uma instalação de maneira simples e contida em um ambiente próprio, de maneira a evitar conflitos de versões de ferramentas e dependências. Além disso, todos os arquivos requeridos e opcionais utilizados pelo detectRF são em formatos padrão, como genomas e transcritomas em formato fasta, bibliotecas de *reads* em formato fastq, e predições de genes de tRNA no formato bed obtidos do gtRNADB (CHAN; LOWE, 2016). Quanto à preparação de arquivos, o único processamento requerido do usuário é que as bibliotecas de *reads* estejam limpas quanto à presença de adaptadores ou de bases de baixa qualidade, visto que não está incluso na *pipeline* um passo de processamento da biblioteca. Por fim, a maior parte dos parâmetros do detectRF são opcionais, sendo os obrigatórios extremamente básicos, facilitando a execução por parte de usuários com pouco conhecimento de bioinformática, enquanto que ainda disponibilizando para usuários com maior conhecimento uma gama de opções que permitam otimizar a análise para os seus dados.

Enquanto foram utilizados como ponto de partida para os cálculos do detectRF os testes estatísticos aplicados pelo tsRFun, pontos passíveis de melhoria nos mesmos foram identificados e essas alterações aplicadas na implementação do detectRF, tanto em termos de melhoria na acurácia, como na utilização de cálculos que condizem melhor com o esperado de uma distribuição aleatória ou a aplicação de filtros com relevância estatística, assim como na utilização de metodologias de aceleração na busca de valores significativos. Além disso, também buscamos alternativas ao cálculo da probabilidade binomial (ARRATIA; GORDON, 1989) que permitissem uma redução do tempo de execução, como a utilização de aproximações, de maneira a permitir a

utilização da ferramenta em dados de larga escala, porém sem perda considerável da acurácia.

Uma das propostas do detectRF é o aumento da acurácia da metodologia aplicada pelo tsRFun, porém foi observado que ocorreu a predição de um menor número de tRFs padrão pelo detectRF do que pelo tsRFun em dados humanos, o que indica uma perda na sensibilidade do método. Enquanto que parte dos tRFs utilizados como padrão inerentemente apresentam problemas para essa utilização, visto que apresentam alinhamento perfeito contra regiões do genoma que não tRNAs e portanto não apresentam alto suporte para predição como tRFs a partir de dados *in silico*, uma porção considerável dos demais não foi identificada, sendo alguns não identificados por ambos os programas. Uma explicação para a não identificação de tRFs padrão seria a falta de *reads* representativos dos mesmos na biblioteca avaliada, sendo isso demonstrado pelos autores do tsRFun em sua avaliação desses dados (WANG et al., 2022). Dada a utilização de uma biblioteca de *reads* sem replicatas experimentais ou técnicas, diversas variações durante os experimentos, preparação da amostra e sequenciamento podem interferir na representatividade de uma molécula de RNA na biblioteca de *reads*. Além disso, a avaliação experimental que demonstra a presença desse tRF não necessariamente é traduzida em um agrupamento de *reads* que fuja da distribuição aleatória, sendo essa uma clara limitação da aplicação dessa metodologia. Por fim, ainda nesse ponto, gargalos de qualidade como a filtragem final de tRFs do detectRF que são muito mais restritivos que os aplicados pelo tsRFun podem levar a perda de tRFs que não atendam aos critérios estabelecidos de desvio da distribuição aleatória.

Apesar de uma menor quantidade de tRFs padrão identificados nas análises de dados humanos, o detectRF apresentou uma quantidade menor de predições falso-positivo nos dados simulados de distribuição aleatória de *reads* quando comparado com o tsRFun. Apesar de esperado que ambos os programas não produzissem alta quantidade de predições falsos-positivas, foi intencionalmente permitido que fossem geradas distribuições de *reads* aleatórias com quantidades de *reads* extremamente altas (até 20000 por tRNA), que não são esperadas de serem encontradas em tRNAs sem tRFs, de maneira que regiões comuns a diferentes tRNAs

pudessem apresentar grande variação na quantidade de *reads* ali presentes. Além disso, esse efeito poderia ser ainda mais proeminente no detectRF, visto que a biblioteca de *reads* aleatórios foi gerada a partir dos conjunto de tRNAs utilizado pelo tsRFun, que apresenta redundância, enquanto que o detectRF utiliza um conjunto de tRNAs únicos para a sua predição, de maneira que para o detectRF as contagens de *reads* de tRNAs redundantes são somadas, podendo potencializar o efeito acima. Por fim, outro ponto que dificulta ainda mais análises com o detectRF nesse conjunto de dados é que a *pipeline* faz a remoção de *reads* que tenham alinhamento perfeito contra o genoma antes da predição, o que não é feito na versão local do tsRFun, de maneira a produzir regiões de queda de contagem de *reads* dentro do tRNA quando essas são comuns a tRNAs e outras regiões o genoma. Apesar dessas dificuldades impostas ao detectRF pela seleção dos tRNAs utilizados pelo tsRFun para a geração de bibliotecas de *reads* aleatórios, o detectRF ainda apresentou menos tRFs falsos-positivos sendo produzidos a partir desses dados. Esse resultado indica que, enquanto menos sensível, o detectRF pode apresentar uma maior especificidade que o tsRFinder. Porém, dada a ausência de um conjunto de testes mais extenso para o *benchmark* de preditores de tRFs, a avaliação entre os programas está limitada à análise de 17 tRFs para os quais o tsRFinder já tem demonstrada capacidade preditiva.

Para ambas as análises, o tempo de execução do detectRF foi maior que o tempo utilizado pelo tsRFun. Porém, alguns pontos devem ser levados em consideração: (i) enquanto o tsRFun faz apenas a predição de tRFs diretamente em dados de *H. sapiens* previamente indexados, o detectRF faz a predição e edição de sequências de tRNA a partir do genoma, assim como a filtragem dos *reads* contra sequências não-tRNA, (ii) o genoma humano apresenta uma sequência extremamente longa, o que torna a execução das dependência e manipulação dos arquivos demoradas, especialmente em pontos não-passíveis de paralelização, (iii) ainda que mais longa, a execução do detectRF apresentou uma duração de menos de 2 horas, o que é um tempo extremamente curto de execução, podendo ser reduzido drasticamente para organismos com genomas reduzidos. Assim, ainda que não apresentando tempos de execução comparáveis aos do tsRFun, o detectRF ainda se

apresenta como uma alternativa viável em termos de tempo de execução para a predição de tRFs.

Dados os resultados observados, podemos delimitar alguns pontos claros da *pipeline* a serem avaliados a fim de melhorar seu desempenho. Primeiramente, é necessária a revisão dos cálculos aplicados pela versão que substitui o cálculo da probabilidade binomial por uma aproximação, visto que em testes durante o desenvolvimento nenhum erro foi identificado, de maneira a permitir seu funcionamento correto e avaliação da sua acurácia e tempo de execução. Além disso, é necessário avaliar quais pontos da predição de tRFs podem estar sendo restritivos demais a ponto de impedir a predição de tRFs confirmados experimentalmente, assim como também avaliar na biblioteca a real presença de *reads* que possam confirmar a presença desses tRFs não identificados. Porém, enquanto que ainda existe espaço para melhoria, a *pipeline* aqui descrita não apenas já se apresenta capaz de prever tRFs, incluindo tRFs confirmados experimentalmente, como tem indícios de uma elevada especificidade, com uma execução simples e organismo-independente. Assim sendo, o detectRF apresenta alto potencial para a predição de tRFs, podendo vir a se tornar uma das ferramentas padrões na predição de tRFs em organismos não-modelos.

## 4. CONCLUSÕES

Enquanto que a relevância para a auto-regulação dos processos de expressão gênica por parte dos sRNAs já é bem consolidada, assim como o conhecimento sobre vias canônicas e classes de sRNAs conhecidas há mais tempo, ainda existe uma grande gama de mecanismos envolvendo sRNAs que permanece pouco explorado. A presença de tRFs na ausência de uma via de RNAi funcional indica a presença de possíveis vias alternativas para a utilização desses sRNAs em fungos, sendo que essas vias podem não apenas se apresentar como mecanismos completamente novos para a utilização de sRNAs, como também podem estar associadas com mecanismos ancestrais ainda não explorados, dada a presença ubíqua de sRNAs. Além disso, a presença de sRNAs em vesículas extracelulares de maneira diferencial, sem uma produção e/ou carregamento indiscriminado em diferentes condições e linhagens, assim como a presença de possíveis alvos em mRNAs de hospedeiros indicam que o papel de sRNAs no desenvolvimento da infecção pode ir além da regulação de fatores envolvidos nesse processo. Nesse ponto, é possível que sRNAs possuam um papel relevante na interação patógeno-hospedeiro por meio de sua utilização como moléculas de comunicação trans-reino, com a finalidade de sequestro da maquinaria de regulação relacionada à sRNAs da contra-parte e interferência nas respostas adaptativas da última com relação ao processo infeccioso.

Além da demonstração de evidências relacionadas a processos pouco elucidados em sRNAs fúngicos, também fomos capazes de desenvolver uma ferramenta para a predição de tRFs a partir de dados de sRNA-seq, a fim de facilitar o estudo dessas moléculas. Enquanto que existe a disponibilidade de dados de sRNA-seq para a identificação dessas moléculas, assim como o crescente aumento na acessibilidade a metodologias de sequenciamento de ácidos nucleicos, essas condições são de pouco proveito na ausência de ferramentas que permitam sua análise. Assim, o desenvolvimento de uma ferramenta que permita a análise de maneira simplificada e, especialmente, de dados de organismos não-modelo, é de grande valor para o desenvolvimento de novos estudos e aprofundamento do conhecimento sobre tRFs nos mais diversos organismos.

Por fim, dadas as informações acima citadas, é possível tomar cada capítulo da presente tese como sendo mais um dos passos iniciais em direção à exploração de pontos não totalmente elucidados da biologia de sRNAs, trazendo evidência da existência de mecanismos inexplorados e ferramentas que facilitem seus estudos. Assim, enquanto que muitas questões e teorias são levantadas frente aos resultados identificados, sem a possível confirmação das mesmas no presente estudo, esperamos que essas mesmas questões e teorias incentivem o desenvolvimento de futuros estudos em áreas de até então pouca atenção.

## 5. PERSPECTIVAS

Com relação ao estudo envolvendo RNAs vesiculares de *C. auris* no capítulo 2, ainda se faz necessário tanto a confirmação em *wet lab* das predições de tRFs e miRNAs por meio de *stem-loop* RT-qPCR, conforme o que foi executado no capítulo 1, além de ensaios para avaliar o potencial de ação dos tRFs vesiculares sobre a regulação da expressão gênica de alvos de mRNA com relevância para a resposta imunitária. Para esse último, ensaios como a avaliação da expressão de genes alvo por RT-qPCR após a administração de vesículas extracelulares de *C. auris* ou de tRFs sinteticamente produzidos para uma avaliação direta, além da avaliação de clivagem do alvo por 5'-RACE podem ser alternativas.

Com relação ao desenvolvimento do detectRF para a aplicação na identificação de tRFs a partir de sRNA-seq no capítulo 3, ainda é necessário a revisão da versão do programa que substitui o cálculo de probabilidade binomial por uma aproximação, a fim de identificar pontos que possam levar ao erro de execução, permitindo seu teste e comparação com o detectRF que utiliza a probabilidade binomial e o tsRFinder. Além disso, também é necessária uma avaliação de possíveis medidas a serem tomadas com a finalidade de melhorar a acurácia do detectRF de maneira geral, visto a reduzida identificação dos tRFs utilizados como padrão para a predição. Por fim, seria de interesse para a demonstração das capacidades do detectRF a aplicação do mesmo em dados de organismos de diferentes reinos e filos.

## 6. BIBLIOGRAFIA

- ACUÑA, S. M.; FLOETER-WINTER, L. M.; MUXEL, S. M. MicroRNAs: Biological Regulators in Pathogen-Host Interactions. **Cells**, v. 9, n. 1, 2 Jan. 2020.
- ARRATIA, R.; GORDON, L. Tutorial on large deviations for the binomial distribution. **Bulletin of Mathematical Biology**, v. 51, n. 1, p. 125–131, 1989.
- ATAYDE, V. D.; TSCHUDI, C.; ULLU, E. The emerging world of small silencing RNAs in protozoan parasites. **Trends in Parasitology**, v. 27, n. 7, p. 321–327, Jul. 2011.
- BABSKI, J. et al. Small regulatory RNAs in Archaea. **RNA Biology**, v. 11, n. 5, p. 484–493, 31 Mar. 2014.
- BAJPAI, V. K. et al. Invasive fungal infections and their epidemiology: measures in the clinical scenario. **Biotechnology and Bioprocess Engineering**, v. 24, n. 3, p. 436–444, Jun. 2019.
- BAKOWSKA-ŻYWICKA, K. et al. The widespread occurrence of tRNA-derived fragments in *Saccharomyces cerevisiae*. **FEBS open bio**, v. 6, n. 12, p. 1186–1200, Dec. 2016.
- BARTEL, D. P. Metazoan MicroRNAs. **Cell**, v. 173, n. 1, p. 20–51, 22 Mar. 2018.
- BENOIT, M.; DESNUES, B.; MEGE, J.-L. Macrophage polarization in bacterial infections. **Journal of Immunology**, v. 181, n. 6, p. 3733–3739, 15 Sep. 2008.
- BOND, C. S.; FOX, A. H. Paraspeckles: nuclear bodies built on long noncoding RNA. **The Journal of Cell Biology**, v. 186, n. 5, p. 637–644, 7 Sep. 2009.
- BONGOMIN, F. et al. Global and Multi-National Prevalence of Fungal Diseases-Estimate Precision. **Journal of fungi (Basel)**, v. 3, n. 4, 18 Oct. 2017.
- BOREK, E. et al. High turnover rate of transfer RNA in tumor tissue. **Cancer Research**, v. 37, n. 9, p. 3362–3366, Sep. 1977.
- BROCK, M. Fungal metabolism in host niches. **Current Opinion in Microbiology**, v. 12, n. 4, p. 371–376, Aug. 2009.
- BROWN, G. D. et al. Hidden killers: human fungal infections. **Science Translational Medicine**, v. 4, n. 165, p. 165rv13, 19 Dec. 2012.
- BUCK, A. H. et al. Exosomes secreted by nematode parasites transfer small RNAs to mammalian cells and modulate innate immunity. **Nature Communications**, v. 5, p. 5488, 25 Nov. 2014.
- CAI, Q.; HE, B.; JIN, H. A safe ride in extracellular vesicles - small RNA trafficking

between plant hosts and pathogens. **Current Opinion in Plant Biology**, v. 52, p. 140–148, Dec. 2019.

CÁNOVAS-MÁRQUEZ, J. T. et al. Role of the Non-Canonical RNAi Pathway in the Antifungal Resistance and Virulence of Mucorales. **Genes**, v. 12, n. 4, 17 Apr. 2021.

CARTHEW, R. W.; SONTHEIMER, E. J. Origins and Mechanisms of miRNAs and siRNAs. **Cell**, v. 136, n. 4, p. 642–655, 20 Feb. 2009.

CHABELSKAYA, S.; GAILLOT, O.; FELDEN, B. A *Staphylococcus aureus* small RNA is required for bacterial virulence and regulates the expression of an immune-evasion molecule. **PLoS Pathogens**, v. 6, n. 6, p. e1000927, 3 Jun. 2010.

CHAI, L. Y. A. et al. Fungal strategies for overcoming host innate immune response. **Medical Mycology**, v. 47, n. 3, p. 227–236, May 2009.

CHAN, P. P.; LOWE, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. **Nucleic Acids Research**, v. 44, n. D1, p. D184-9, 4 Jan. 2016.

CHEKULAEVA, M.; FILIPOWICZ, W. Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. **Current Opinion in Cell Biology**, v. 21, n. 3, p. 452–460, Jun. 2009.

CHENG, G. et al. Deep sequencing-based identification of pathogen-specific microRNAs in the plasma of rabbits infected with *Schistosoma japonicum*. **Parasitology**, v. 140, n. 14, p. 1751–1761, Dec. 2013.

COLE, C. et al. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. **RNA (New York)**, v. 15, n. 12, p. 2147–2160, Dec. 2009.

CROOKS, G. E. et al. WebLogo: a sequence logo generator. **Genome Research**, v. 14, n. 6, p. 1188–1190, Jun. 2004.

CROSTON, T. L. et al. MicroRNA Regulation of Host Immune Responses following Fungal Exposure. **Frontiers in immunology**, v. 9, p. 170, 7 Feb. 2018.

CUI, C. et al. A fungal pathogen deploys a small silencing RNA that attenuates mosquito immunity and facilitates infection. **Nature Communications**, v. 10, n. 1, p. 4298, 20 Sep. 2019.

DANECEK, P. et al. Twelve years of SAMtools and BCFtools. **GigaScience**, v. 10, n. 2, 16 Feb. 2021.

DINGER, M. E. et al. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. **PLoS Computational Biology**, v. 4, n. 11, p. e1000176, 28 Nov. 2008.

DI FAZIO, A. et al. Dicer dependent tRNA derived small RNAs promote nascent RNA

silencing. **Nucleic Acids Research**, v. 50, n. 3, p. 1734–1752, 22 Feb. 2022.

DOHERTY, E. A.; DOUDNA, J. A. Ribozyme structures and mechanisms. **Annual review of biophysics and biomolecular structure**, v. 30, p. 457–475, 2001.

DUTTA, T.; SRIVASTAVA, S. Small RNA-mediated regulation in bacteria: A growing palette of diverse mechanisms. **Gene**, v. 656, p. 60–72, 20 May 2018.

DU, H. et al. *Candida auris*: Epidemiology, biology, antifungal resistance, and virulence. **PLoS Pathogens**, v. 16, n. 10, p. e1008921, 22 Oct. 2020.

EISENBART, S. K. et al. A Repeat-Associated Small RNA Controls the Major Virulence Factors of *Helicobacter pylori*. **Molecular Cell**, v. 80, n. 2, p. 210–226.e7, 15 Oct. 2020.

EMRI, T. et al. Echinocandins: production and applications. **Applied Microbiology and Biotechnology**, v. 97, n. 8, p. 3267–3284, Apr. 2013.

FIRACATIVE, C. Invasive fungal disease in humans: are we aware of the real impact? **Memorias do Instituto Oswaldo Cruz**, v. 115, p. e200430, 9 Oct. 2020.

FRANÇA, F. D. et al. Alteration in cellular viability, pro-inflammatory cytokines and nitric oxide production in nephrotoxicity generation by Amphotericin B: involvement of PKA pathway signaling. **Journal of Applied Toxicology**, v. 34, n. 12, p. 1285–1292, Dec. 2014.

FU, Q.; WANG, P. J. Mammalian piRNAs: Biogenesis, function, and mysteries. **Spermatogenesis**, v. 4, p. e27889, 7 Feb. 2014.

GEBETSBERGER, J. et al. tRNA-derived fragments target the ribosome and function as regulatory non-coding RNA in *Haloferax volcanii*. **Archaea**, v. 2012, p. 260909, 27 Dec. 2012.

GIRARD, A.; HANNON, G. J. Conserved themes in small-RNA-mediated transposon control. **Trends in Cell Biology**, v. 18, n. 3, p. 136–148, Mar. 2008.

GOODARZI, H. et al. Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement. **Cell**, v. 161, n. 4, p. 790–802, 7 May 2015.

GULERIA, P. et al. Plant small RNAs: biogenesis, mode of action and their roles in abiotic stresses. **Genomics, proteomics & bioinformatics / Beijing Genomics Institute**, v. 9, n. 6, p. 183–199, Dec. 2011.

GUO, Z.; LI, Y.; DING, S.-W. Small RNA-based antimicrobial immunity. **Nature Reviews. Immunology**, v. 19, n. 1, p. 31–44, Jan. 2019.

HARTMANN, R. K. et al. The making of tRNAs and more - RNase P and tRNase Z. **Progress in molecular biology and translational science**, v. 85, p. 319–368, 2009.

HAUSSECKER, D. et al. Human tRNA-derived small RNAs in the global regulation of RNA silencing. **RNA (New York)**, v. 16, n. 4, p. 673–695, Apr. 2010.

HOKKEN, M. W. J. et al. Facilitators of adaptation and antifungal resistance mechanisms in clinically relevant fungi. **Fungal Genetics and Biology**, v. 132, p. 103254, Nov. 2019.

HUANG, V.; LI, L.-C. miRNA goes nuclear. **RNA Biology**, v. 9, n. 3, p. 269–273, 1 Mar. 2012.

JEANG, K.-T. RNAi in the regulation of mammalian viral infections. **BMC Biology**, v. 10, p. 58, 26 Jun. 2012.

JOHN, B. et al. Human MicroRNA targets. **PLoS Biology**, v. 2, n. 11, p. e363, Nov. 2004.

KARGINOV, F. V.; HANNON, G. J. The CRISPR system: small RNA-guided defense in bacteria and archaea. **Molecular Cell**, v. 37, n. 1, p. 7–19, 15 Jan. 2010.

KEAM, S. P. et al. tRNA-Derived RNA Fragments Associate with Human Multisynthetase Complex (MSC) and Modulate Ribosomal Protein Translation. **Journal of Proteome Research**, v. 16, n. 2, p. 413–420, 3 Feb. 2017.

KIM, H. K. et al. A transfer-RNA-derived small RNA regulates ribosome biogenesis. **Nature**, v. 552, n. 7683, p. 57–62, 7 Dec. 2017.

KNEALE, M. et al. Global access to antifungal therapy and its variable cost. **The Journal of Antimicrobial Chemotherapy**, v. 71, n. 12, p. 3599–3606, Dec. 2016.

KOEPPE, K. et al. A Novel Mechanism of Host-Pathogen Interaction through sRNA in Bacterial Outer Membrane Vesicles. **PLoS Pathogens**, v. 12, n. 6, p. e1005672, 13 Jun. 2016.

KRISHNAN, J. et al. Toll-like receptor signal transduction. **Experimental & Molecular Medicine**, v. 39, n. 4, p. 421–438, 31 Aug. 2007.

KUMAR, P.; KUSCU, C.; DUTTA, A. Biogenesis and Function of Transfer RNA-Related Fragments (tRFs). **Trends in Biochemical Sciences**, v. 41, n. 8, p. 679–689, Aug. 2016.

KUMAR, P. et al. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. **BMC Biology**, v. 12, p. 78, 1 Oct. 2014.

KUSCU, C. et al. tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. **RNA (New York)**, v. 24, n. 8, p. 1093–1105, Aug. 2018.

KWON-CHUNG, K. J. et al. Cryptococcus neoformans and Cryptococcus gattii, the etiologic agents of cryptococcosis. **Cold Spring Harbor perspectives in medicine**, v. 4, n. 7, p. a019760, 1 Jul. 2014.

LANGENBERGER, D. et al. Dicer-processed small RNAs: rules and exceptions. **Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution**, v. 320, n. 1, p. 35–46, Jan. 2013.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, v. 9, n. 4, p. 357–359, 4 Mar. 2012.

LEE, R. C.; FEINBAUM, R. L.; AMBROS, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. **Cell**, v. 75, n. 5, p. 843–854, 3 Dec. 1993.

LEE, Y. S. et al. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). **Genes & Development**, v. 23, n. 22, p. 2639–2649, 15 Nov. 2009.

LINXWEILER, J.; JUNKER, K. Extracellular vesicles in urological malignancies: an update. **Nature Reviews. Urology**, v. 17, n. 1, p. 11–27, Jan. 2020.

LI, S.; XU, Z.; SHENG, J. tRNA-Derived Small RNA: A Novel Regulatory Small Non-Coding RNA. **Genes**, v. 9, n. 5, 10 May 2018.

LI, Z. et al. Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs. **Nucleic Acids Research**, v. 40, n. 14, p. 6787–6799, Aug. 2012.

LOHER, P.; TELONIS, A. G.; RIGOUTSOS, I. MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. **Scientific Reports**, v. 7, p. 41184, 21 Feb. 2017.

LOSS-MORAIS, G.; WATERHOUSE, P. M.; MARGIS, R. Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets. **Biology Direct**, v. 8, p. 6, 12 Feb. 2013.

LUO, S. et al. *Drosophila* tsRNAs preferentially suppress general translation machinery via antisense pairing and participate in cellular starvation response. **Nucleic Acids Research**, v. 46, n. 10, p. 5250–5268, 1 Jun. 2018.

MAUTE, R. L. et al. tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in B cell lymphoma. **Proceedings of the National Academy of Sciences of the United States of America**, v. 110, n. 4, p. 1404–1409, 22 Jan. 2013.

MA, L.; BAJIC, V. B.; ZHANG, Z. On the classification of long non-coding RNAs. **RNA Biology**, v. 10, n. 6, p. 925–933, 1 Jun. 2013.

MEGEL, C. et al. Plant RNases T2, but not Dicer-like proteins, are major players of tRNA-derived fragments biogenesis. **Nucleic Acids Research**, v. 47, n. 2, p. 941–952, 25 Jan. 2019.

MENG, H. et al. Dicer and Argonaute Genes Involved in RNA Interference in the

Entomopathogenic Fungus *Metarhizium robertsii*. **Applied and Environmental Microbiology**, v. 83, n. 7, 1 Apr. 2017.

MI, H. et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. **Nucleic Acids Research**, v. 49, n. D1, p. D394–D403, 8 Jan. 2021.

MLECZKO, A. M.; CELICHOWSKI, P.; BAŃKOWSKA-ŻYWICKA, K. Transfer RNA-derived fragments target and regulate ribosome-associated aminoacyl-transfer RNA synthetases. **Biochimica et biophysica acta. Gene regulatory mechanisms**, 6 Jun. 2018.

MOAZED, D. Small RNAs in transcriptional gene silencing and genome defence. **Nature**, v. 457, n. 7228, p. 413–420, 22 Jan. 2009.

MUNHOZ DA ROCHA, I. F. et al. Cross-Kingdom Extracellular Vesicles EV-RNA Communication as a Mechanism for Host-Pathogen Interaction. **Frontiers in cellular and infection microbiology**, v. 10, p. 593160, 18 Nov. 2020.

MUNHOZ DA ROCHA, I. F. et al. Cellular and Extracellular Vesicle RNA Analysis in the Global Threat Fungus *Candida auris*. **Microbiology spectrum**, v. 9, n. 3, p. e0153821, 22 Dec. 2021.

MUÑOZ, J. F. et al. Coordinated host-pathogen transcriptional dynamics revealed using sorted subpopulations and single macrophages infected with *Candida albicans*. **Nature Communications**, v. 10, n. 1, p. 1607, 8 Apr. 2019.

MURPHY, E. R.; PAYNE, S. M. RyhB, an iron-responsive small RNA molecule, regulates *Shigella dysenteriae* virulence. **Infection and Immunity**, v. 75, n. 7, p. 3470–3477, Jul. 2007.

NICOLÁS, F. E.; GARRE, V. RNA interference in fungi: retention and loss. **Microbiology spectrum**, v. 4, n. 6, Dec. 2016.

O'CONNOR, C. M.; VANICEK, J.; MURPHY, E. A. Host microRNA regulation of human cytomegalovirus immediate early protein translation promotes viral latency. **Journal of Virology**, v. 88, n. 10, p. 5524–5532, May 2014.

OBERBAUER, V.; SCHAEFER, M. R. tRNA-Derived Small RNAs: Biogenesis, Modification, Function and Potential Impact on Human Disease Development. **Genes**, v. 9, n. 12, 5 Dec. 2018.

PASQUIER, E. et al. Long-term Mortality and Disability in Cryptococcal Meningitis: A Systematic Literature Review. **Clinical Infectious Diseases**, v. 66, n. 7, p. 1122–1132, 19 Mar. 2018.

PATEL, N.; NAIR, M. The small RNA RsaF regulates the expression of secreted virulence factors in *Staphylococcus aureus* Newman. **Journal of Microbiology**, v. 59,

n. 10, p. 920–930, Oct. 2021.

PELLON, A. et al. Role of Cellular Metabolism during Candida-Host Interactions. **Pathogens (Basel, Switzerland)**, v. 11, n. 2, 28 Jan. 2022.

PÉREZ-ARQUES, C. et al. A non-canonical RNAi pathway controls virulence and genome stability in Mucorales. **PLoS Genetics**, v. 16, n. 7, p. e1008611, 13 Jul. 2020.

PERLIN, D. S.; RAUTEMAA-RICHARDSON, R.; ALASTRUEY-IZQUIERDO, A. The global problem of antifungal resistance: prevalence, mechanisms, and management. **The Lancet Infectious Diseases**, v. 17, n. 12, p. e383–e392, Dec. 2017.

PU, M. et al. Regulatory network of miRNA on its target: coordination between transcriptional and post-transcriptional regulation of gene expression. **Cellular and Molecular Life Sciences**, v. 76, n. 3, p. 441–451, Feb. 2019.

RAJASINGHAM, R. et al. Global burden of disease of HIV-associated cryptococcal meningitis: an updated analysis. **The Lancet Infectious Diseases**, v. 17, n. 8, p. 873–881, Aug. 2017.

RAMAN, V. et al. Small RNA Functions Are Required for Growth and Development of *Magnaporthe oryzae*. **Molecular Plant-Microbe Interactions**, v. 30, n. 7, p. 517–530, Jul. 2017.

ROBBINS, P. D.; MORELLI, A. E. Regulation of immune responses by extracellular vesicles. **Nature Reviews. Immunology**, v. 14, n. 3, p. 195–208, Mar. 2014.

RYVKIN, P. et al. Using machine learning and high-throughput RNA sequencing to classify the precursors of small non-coding RNAs. **Methods**, v. 67, n. 1, p. 28–35, 1 May 2014.

SELITSKY, S. R.; SETHUPATHY, P. tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. **BMC Bioinformatics**, v. 16, p. 354, 4 Nov. 2015.

SHI, J. et al. SPORTS1.0: A Tool for Annotating and Profiling Non-coding RNAs Optimized for rRNA- and tRNA-derived Small RNAs. **Genomics, proteomics & bioinformatics / Beijing Genomics Institute**, v. 16, n. 2, p. 144–151, Apr. 2018.

SHOHAM, S.; LEVITZ, S. M. The immune response to fungal infections. **British Journal of Haematology**, v. 129, n. 5, p. 569–582, Jun. 2005.

SPEER, J. et al. tRNA breakdown products as markers for cancer. **Cancer**, v. 44, n. 6, p. 2120–2123, Dec. 1979.

STREIT, R. S. A. et al. Analysis of tRNA-derived RNA fragments (tRFs) in *Cryptococcus* spp.: RNAi-independent generation and possible compensatory effects in a RNAi-deficient genotype. **Fungal biology**, v. 125, n. 5, p. 389–399, May 2021.

SUPEK, F. et al. REVIGO summarizes and visualizes long lists of gene ontology terms. **Plos One**, v. 6, n. 7, p. e21800, 18 Jul. 2011.

TALAPKO, J. et al. *Candida albicans*-The Virulence Factors and Clinical Manifestations of Infection. **Journal of fungi (Basel)**, v. 7, n. 2, 22 Jan. 2021.

THOMAS, P. D. et al. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. **Nucleic Acids Research**, v. 34, n. Web Server issue, p. W645-50, 1 Jul. 2006.

TORRES-MARTÍNEZ, S.; RUIZ-VÁZQUEZ, R. M. The rnaI universe in fungi: A varied landscape of small rnas and biological functions. **Annual Review of Microbiology**, v. 71, p. 371–391, 8 Sep. 2017.

TRISTÁN-RAMOS, P. et al. The tumor suppressor microRNA let-7 inhibits human LINE-1 retrotransposition. **Nature Communications**, v. 11, n. 1, p. 5712, 11 Nov. 2020.

TSAI, S.-T. et al. Three-year mortality in cryptococcal meningitis: Hyperglycemia predict unfavorable outcome. **Plos One**, v. 16, n. 5, p. e0251749, 28 May 2021.

VAN DER POL, E. et al. Classification, functions, and clinical relevance of extracellular vesicles. **Pharmacological Reviews**, v. 64, n. 3, p. 676–705, Jul. 2012.

VITSIOS, D. M. et al. Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. **Nucleic Acids Research**, v. 45, n. 21, p. e177, 1 Dec. 2017.

WALZER, K. A.; CHI, J.-T. Trans-kingdom small RNA transfer during host-pathogen interactions: The case of *P. falciparum* and erythrocytes. **RNA Biology**, v. 14, n. 4, p. 442–449, 3 Apr. 2017.

WANG, J.-H. et al. tsRFun: a comprehensive platform for decoding human tsRNA expression, functions and prognostic value by high-throughput small RNA-Seq and CLIP-Seq data. **Nucleic Acids Research**, v. 50, n. D1, p. D421–D431, 7 Jan. 2022.

WANG, Q. et al. The tRNA-Derived Small RNAs Regulate Gene Expression through Triggering Sequence-Specific Degradation of Target Transcripts in the Oomycete Pathogen *Phytophthora sojae*. **Frontiers in plant science**, v. 7, p. 1938, 22 Dec. 2016.

WANG, Y. et al. Insects defend against fungal infection by employing microRNAs to silence virulence-related genes. **Proceedings of the National Academy of Sciences of the United States of America**, v. 118, n. 19, 11 May 2021.

WANG, Z. et al. Genome-wide identification and characterization of transfer RNA-derived small RNAs in *Plasmodium falciparum*. **Parasites & vectors**, v. 12, n. 1, p. 36, 15 Jan. 2019.

WEICK, E.-M.; MISKA, E. A. piRNAs: from biogenesis to function. **Development**, v. 141, n. 18, p. 3458–3471, Sep. 2014.

WIEDERHOLD, N. P. Antifungal resistance: current trends and future strategies to combat. **Infection and drug resistance**, v. 10, p. 249–259, 29 Aug. 2017.

WILSON, R. C.; DOUDNA, J. A. Molecular mechanisms of RNA interference. **Annual review of biophysics**, v. 42, p. 217–239, 2013.

XIE, Y. et al. Action mechanisms and research methods of tRNA-derived small RNAs. **Signal transduction and targeted therapy**, v. 5, n. 1, p. 109, 30 Jun. 2020.

XU, W.-L. et al. Computational Approaches to tRNA-Derived Small RNAs. **Non-coding RNA**, v. 3, n. 1, 4 Jan. 2017.

ZHANG, C. Novel functions for small RNA molecules. **Current Opinion in Molecular Therapeutics**, v. 11, n. 6, p. 641–651, Dec. 2009.

ZHENG, L.-L. et al. tRF2Cancer: A web server to detect tRNA-derived small RNA fragments (tRFs) and their expression in multiple cancers. **Nucleic Acids Research**, v. 44, n. W1, p. W185-93, 8 Jul. 2016.

ZIEMANN, M.; KASPI, A.; EL-OSTA, A. Evaluation of microRNA alignment techniques. **RNA (New York)**, v. 22, n. 8, p. 1120–1138, Aug. 2016.