Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática

João Paulo Pereira de Almeida

EXPLORANDO O VIROMA DE MOSQUITOS VETORES:
uma abordagem metagenômica utilizando pequenos RNAs da resposta imune do hospedeiro

Belo Horizonte
2023

João Paulo Pereira de Almeida

EXPLORANDO O VIROMA DE MOSQUITOS VETORES: uma abordagem metagenômica utilizando pequenos RNAs da resposta imune do hospedeiro

Tese de doutorado apresentada ao Programa de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais para obtenção do título de Doutor em Ciências.

Orientador: Prof. Dr. João Trindade Marques

Coorientador: Prof. Dr. Eric Roberto Guimarães Rocha Aguiar

Belo Horizonte
2023

043 Almeida, João Paulo Pereira de.
Explorando o Viroma de mosquitos vetores: uma abordagem metagenômica utilizando pequenos RNAs da resposta imune do hospedeiro [manuscrito] / João Paulo Pereira de Almeida. – 2023.
184 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. João Trindade Marques. Coorientador: Prof. Dr. Eric Roberto Guimarães Rocha Aguiar.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Metagenômica. 3. MicroRNAs. 4. Culicidae. 5. Arbovirus. 6. Aprendizado de Máquina. I. Marques, João Trindade. II. Aguiar, Eric Roberto Guimarães Rocha. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

ATA DA DEFESA DE TESE

JOÃO PAULO PEREIRA DE ALMEIDA

Às oito horas do dia **27 de outubro de 2023**, reuniu-se, através de videoconferência, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho de **João Paulo Pereira de Almeida** intitulado: **"Explorando o Viroma de mosquitos vetores: uma abordagem metagenômica utilizando pequenos RNAs da resposta imune do hospedeiro"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. João Trindade Marques**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor/Pesquisador	**Instituição**	**Indicação**
Dr. João Trindade Marques - Orientador	Universidade Federal de Minas Gerais	Aprovado
Dr. Eric Roberto Guimarães Rocha Aguiar - Coorientador	Universidade Estadual de Santa Cruz	Aprovado
Dr. Francisco Pereira Lobo	Universidade Federal de Minas Gerais	Aprovado
Dr. Gabriel da Luz Wallau	Instituto Aggeu Magalhães/Fundação Oswaldo Cruz	Aprovado
Dr. Helder Takashi Imoto Nakaya	Hospital Israelita Albert Einstein	Aprovado
Dr. Sávio Torres de Farias	Universidade Federal da Paraíba	Aprovado

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 27 de outubro de 2023.

Documento assinado eletronicamente por **Savio Torres de Farias**, **Usuário Externo**, em 27/10/2023, às 16:09, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **João Trindade Marques**, **Professor do Magistério Superior**, em 30/10/2023, às 06:31, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eric Roberto Guimaraes Rocha Aguiar**, **Usuário Externo**, em 30/10/2023, às 22:30, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gabriel da Luz Wallau**, **Usuário Externo**, em 31/10/2023, às 10:30, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Francisco Pereira Lobo**, **Professor do Magistério Superior**, em 31/10/2023, às 11:05, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Helder Takashi Imoto Nakaya**, **Usuário Externo**, em 10/11/2023, às 09:54, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2739018** e o código CRC **E42A1D0F**.

Dedico esse trabalho aos meus pais, Paulo e Vilma, pelo apoio incondicional aos meus estudos e carreira acadêmica.

AGRADECIMENTOS

Ao meu orientador, Prof. João Marques. Por todas oportunidades e ensinamentos, sua excelência, competência e dedicação na carreira acadêmica são exemplos para mim. Terei sempre orgulho de ter feito parte do seu laboratório nessa fase da minha carreira acadêmica.

Ao meu coorientador, Prof. Eric Aguiar. Por todo conhecimento de bioinformática incondicionalmente transmitido, oportunidades proporcionadas e por ter se tornado um grande amigo. Sua dedicação e afinco com a carreira acadêmica são exemplos para mim.

Ao Dr. Roenick Olmo, pessoa central no sequenciamento das bibliotecas de RNA-seq sem as quais esse trabalho de doutorado não existiria. Pelo companheirismo e ensinamentos proporcionados. Seu bom humor e gentileza para com os colegas são exemplos de comportamento na vida acadêmica para mim.

Ao Dr. Mathias Todjro. Pelas inúmeras discussões sobre experimentos com mosquitos que me ajudaram a entender melhor a biologia dos vírus que estudamos no laboratório. Pelo companheirismo e respeito em todos os projetos que trabalhamos juntos.

Ao Dr. Isaque Faria. Pela disposição em discutir ciência e sempre ter sugestões de referências que me ajudaram na execução dessa tese. Por toda sua dedicação e cuidados com o Laboratório RNAi.

A todos os integrantes atuais e passados do Laboratório RNAi que tive o prazer de conviver durante o doutorado, Dra. Flávia Ferreira, Dr. Álvaro Ferreira, Dra. Emanuele Silva, Raianna Boni, Siad Amadou, Thiago Jiran, Ezequiel Salvador, Carlos Estevez, Elisa Gonçalves, Ellen Caroline e Daniele Almeida. Por terem me feito sentir em casa e por todo apoio na execução dessa tese.

Aos alunos de iniciação cientifica que orientei ou coorientarei. Juliana Armache, Lucas Coutinho, Hebert Costa, Victor Abdallah e Tamires Franco. Foi um prazer fazer parte da formação de pessoas tão talentosas quanto vocês que me ensinaram muito e tem colaborações significativas no meu doutorado.

A Profa. Fabiola Ribeiro e a todos integrantes do laboratório de Neurobioquimica que convivi. Pelo bom convívio e empenho em fazer do nosso laboratório um ambiente agradável e funcional.

A Professora Betânia Drumond e a doutoranda Ana Luiza Cruz do LabVirus do ICB, por sempre atenderem prontamente aos nossos pedidos de informações sobre os mosquitos coletados pelo seu grupo.

Ao Dr. Luciano Moreira (Fiocruz-MG), coordenador do Projeto *Wolbachia* em *A. aegypti,* e aos demais membros do projeto que tive o prazer de interagir, pela colaboração e informações compartilhadas na execução do projeto de detecção de vírus em Niterói- RJ.

Ao meu colega e amigo Bruno Silva. Por todas as colaborações cientificas e ajuda técnica com a servidora do nosso laboratório. Pela incansável disposição em discutir bioinformática e companheirismo.

Ao Prof. Ricardo Solar (Bob). Por fazer o ICB se parecer um pouco mais com a nossa antiga casa na ecologia da UFV. Pelos conselhos acadêmicos e conversas sobre ciência. Pelas oportunidades de ministrar as disciplinas de R. Você será sempre uma referência de honestidade intelectual e competência acadêmica para mim.

A Dra. Sandra Abbo da Universidade de Wageningen. Pelo rigor científico, comprometimento e companheirismo na execução do nosso projeto com *Aedes japonicus.*

Ao nosso grupo “nota 100” da disciplina EGTP, Renato Oliveira, Bruno Silva e Marcos Viana. Pelo companheirismo e amizade durante todo o doutorado.

Aos colegas do grupo da disciplina de Algoritmos, Hemanoel Passarelli, Lucas Pontes e Thaís Rodrigues. Pela companhia nas longas noites de estudo para as provas e trabalhos, foi mais fácil e divertido com vocês.

Aos secretários da PPG Bioinformática, Tiago e Sheila. Pela prontidão e gentileza que tiveram comigo em todas as situações que precisei de vocês.

A Profa. Mariana Quezado. Por permitir e apoiar nossa colaboração quando as minhas conversas com o Bruno Silva sobre o mestrado dele se transformaram em análises de dados, foi um prazer colaborar com vocês.

Aos meus amigos do laboratório do mestrado, Vicente Gomes, Alan Lorenzetti e Felipe Caten. Por continuarem grandes companheiros acadêmicos mesmo a distância durante a pandemia e estarem sempre a disposição para discutir bioinformática.

A Profa. Andrea Macedo. Por ter me recebido na UFMG e me apresentado o ICB. Por toda atenção e apoio dentro e fora da UFMG.

Aos meus pais, Paulo e Vilma, e a minha irmã Mariana. Pelo apoio incondicional aos meus estudos e carreira acadêmica. Vocês são o alicerce sólido que me permite construir um futuro pessoal e profissional.

A minha companheira, Gabriella. Pelos muitos anos de jornada juntos. Pela cumplicidade e apoio em todos os momentos durante meu doutorado.

A toda a família que ganhei em BH, Gabriella, Ricardo, Giovanna, Arthur, Victor e Manuela. Por todo apoio e bons momentos que me proporcionaram, foi tudo muito mais fácil com vocês por perto.

A todas as pessoas que algum dia se dispuseram, sem qualquer pretensão de recompensa, a compartilhar conhecimentos e soluções com rigor técnico-científico nos mais diversos fóruns e ambientes da internet. Essa tese teve a ajuda de muitos de vocês que fazem do conhecimento técnico-científico algo acessível e público. Para mim, esse é um dos maiores exemplos de Humanismo.

"*Eu sou uma coleção de moléculas orgânicas chamada Carl Sagan. Você é uma coleção de moléculas quase idêntica que recebe um nome diferente. Mas é apenas isso? Não há nada além de moléculas aqui? Algumas pessoas acham essa ideia depreciativa para a dignidade humana. Para mim, é enaltecedor o fato de que as leis da física do nosso universo permitem a evolução de máquinas moleculares tão complexas e sutis como nós*."

Carl Sagan. Cosmos, 1980.

AGÊNCIAS DE FOMENTO

Essa tese de doutorado teve o suporte de instituições e órgãos de fomento públicos e privados, nacionais e internacionais que permitiram a execução do projeto com apoio financeiro e de infraestrutura.

CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico.

FAPEMIG – Fundação de Amparo à Pesquisa do Estado de Minas Gerais.

ICB UFMG – Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais.

Programa de Pós-Graduação em Bioinformática UFMG.

Google Latin American Research Award (2019-2020).

ZikaAlliance Consortium.

Université de Strasbourg.

Instituto René Rachou – FIOCRUZ MINAS

AWS – Amazon Web Services.

RESUMO

Mosquitos são os principais vetores de arbovírus do planeta. O recente emprego das técnicas de sequenciamento de ácidos nucleicos em larga escala associadas a metagenômica permitiram a descoberta de centenas de vírus em mosquitos. A maioria desses vírus são classificados como vírus específicos de insetos, não sendo capazes de infectar hospedeiros vertebrados. Diversos estudos têm mostrado a capacidade desses vírus em afetar a competência vetorial de mosquitos para a transmissão de arbovírus de importância médica. Apesar dos avanços proporcionados pela metagenômica na descoberta de novos vírus em eucariotos, existem três grandes desafios computacionais associados a essa abordagem aplicada em larga: associação de sequências de vírus com genomas segmentados ao mesmo vírus; diferenciação entre sequências virais exógenas e elementos virais endógenos; e a falta de sensibilidade para a detecção de sequências virais altamente divergentes devido a dependência de técnicas baseadas em similaridade de sequências. Nosso grupo propõe uma estratégia metagenômica utilizando pequenos RNAs da resposta imune dos hospedeiros para abordar esses desafios da metagenômica viral. Nesse estudo, analisamos 122 bibliotecas inéditas de pequenos RNAs geradas com amostras de mosquitos vetores de 10 espécies coletados em quatro continentes. Foram identificados 28 vírus, 17 potencialmente novos, pertencentes a mais de 17 famílias virais; e 1726 sequências virais endógenas oriundas de 115 espécies virais pertencentes a 24 famílias virais. Identificamos e caracterizamos um segmento viral não detectado por métodos de similaridade de sequência que pertence a um Narnavírus, sendo esse um exemplo claro da matéria escura da matagenômica viral. Desenvolvemos a ferramenta *Small RNA Metavir,* um pipeline automatizado para a análises metagenômicas com pequenos RNAs que permite a distinção de sequências virais exógenas de endógenas com classificadores baseados em aprendizado de máquina treinados nesse estudo. Além da precisa detecção de sequências virais, mostramos que nossa estratégia pode detectar RNAs de *Wolbachia,* um importante endossimbionte que afeta a competência vetorial de mosquitos. Mostramos que a carga de pequenos RNAs de *Wolbachia* possui uma correlação negativa com a carga de vírus específicos de insetos em mosquitos de Niterói-RJ onde essa bactéria é usada para o controle da transmissão de arbovírus por *Aedes aegypti*. Nessa tese, foram obtidos resultados e desenvolvidas abordagens de bioinformática que permitem caracterizar de forma precisa o viroma de mosquitos vetores e analisá-lo a luz de componentes biológicos relevantes para a competência vetorial de mosquitos. Os resultados obtidos contribuem para avanços no conhecimento das intrincadas relações entre mosquitos e vírus, algo essencial para a fundamentação de novas estratégias de controle biológico desses vetores de patógenos.

Palavras-chave: Pequenos RNAs; Metagenômica; Mosquitos; Arbovírus; Aprendizado de máquina; Simbiontes.

ABSTRACT

Mosquitoes are the primary vectors of arboviruses worldwide. The recent application of high-throughput nucleic acid sequencing techniques coupled with metagenomics has enabled the discovery of hundreds of viruses in mosquitoes. Most of these viruses are classified as insect-specific, incapable of infecting vertebrate hosts. Several studies have demonstrated the ability of these viruses to impact the vector competence of mosquitoes for the transmission of medically important arboviruses. Despite the advances provided by metagenomics in uncovering new viruses in eukaryotes, three major computational challenges are associated with this approach when applied in large-scale: associating viral sequences with segmented genomes of the same virus; distinguishing between exogenous viral sequences and endogenous viral elements; and the lack of sensitivity in detecting highly divergent viral sequences due to the dependence on sequence similarity-based techniques. Our group proposes a metagenomic strategy using small RNAs from host immune responses to address these challenges in viral metagenomics. In this study, we analyzed 122 novel small RNA libraries generated from samples of vector mosquitoes from 10 species collected on four continents. We identified 28 viruses belonging to more than nine viral families and 1726 endogenous viral sequences derived from 115 viral species across 24 viral families. We identified and characterized a viral segment not detected by sequence similarity methods, belonging to a Narnavirus, providing a clear example of the dark matter of viral metagenomics. We developed the *Small RNA Metavir* tool, an automated pipeline for metagenomic analysis with small RNAs that enables the distinction between exogenous and endogenous viral sequences using machine learning classifiers trained in this study. In addition to the precise detection of viral sequences, we demonstrated that our strategy can detect RNAs from Wolbachia, a crucial endosymbiont that affects the vector competence of mosquitoes. We showed that the load of Wolbachia small RNAs has a negative correlation with the load of insect-specific viruses in mosquitoes from Niterói-RJ, where this bacterium is used for the control of arbovirus transmission by *Aedes aegypti*. In this thesis, we obtained results and developed bioinformatics approaches that allow for the precise characterization of the virome of vector mosquitoes and its analysis in light of biological components relevant to mosquito vector competence. The results contribute to advancements in understanding the intricate relationships between mosquitoes and viruses, essential for the development of new strategies for the biological control of these pathogen vectors.

Keywords: Small RNAs; Metagenomics; Mosquitoes; Arbovirus; Machine learning; Symbionts.

LISTA DE ABREVIAÇÕES

A	Adenina
AUC	*Area Under the Curve*
Ago	Proteina Argonauta
BC	*Baltimore Class*
BLAST	*Basic Local Alignment Search Tool*
C	Citosina
CDC	Centro de Controle e Prevenção de Doenças
CPAN	*Comprehensive Perl Archive Network*
CRISPR	*Clustered Regularly Interspaced Short Palindromic Repeats*
Ct	*Cycle threshold*
DET	*Detection Error Tradeoff*
DNA	Ácido desoxirribonucleico
E-value	*Expectation value*
EVE	Elementos virais endógenos
G	Guanina
GFF	*General Feature Format*
HLC	*Human landing catch*
ICTV	*International Committee on Taxonomy of Viruses*
ISV	*Insect specific viruses*
KNN	*k-Nearest Neighbors*
LECA	*Last eukayotic common ancestor*
LUCA	Último ancestral comum universal
MSA	*Multiple Sequence Alignment*
NCBI	*National Center for Biotechnology Information*
OAS	Oligoadenilato sintetases
ORF	*Open Read Frame*
PAE	*Predicted Aligned Error*
PAGE	*Polyacrylamide Gel Electrophoresis*
PCA	*Principal Component Analysis*
PCR	*Polymerase Chain Reaction*
PIWI	*P-element Induced Wimpy Testis*
RISC	*RNA-Induced Silencing Complex*
RNA	Ácido ribonucleico
RNA-seq	Sequenciamento de RNA
RNAi	RNA de interferência
ROC	*Receiver Operating Characteristic*
RPKM	*Reads Per Kilobase Million*
RPM	*Reads Per Million*
RT-qPCR	*Reverse Transcription Quantitative Polymerase Chain Reaction*

Refseq	*Reference Sequence database*
S1	Segmento genômico 1 de Narnavirus
S2	Segmento genômico 2 de Narnavirus
SL	*Stem-loop*
SVM	*Support Vector Machine*
Sam	*Sequence alignment map*
T	Timina
U	Uracila
UTR	*Untranslated Region*
UV	Ultravioleta
Aa	Aminoácidos
cDNA	DNA complementar
d.p.e	Dias pós emergência
dsRNA	RNA de fita dupla
fORF	*forward ORF*
ftp	*File Transfer Protocol*
mRNA	RNA mensageiro
miRNA	Micro RNA
ncRNA	RNA não codificante
nt	Nucleotideos
pLDDT	*Per-residue log-likelihood gain*
piRNA	Pequenos RNAs associados a proteína PIWI
rORF	*reverse ORF*
rRNA	RNA ribossomal
sfRNA	RNA subgenômico de flavivírus
siRNA	Pequeno RNA de interferência
ssRNA	RNA de fita simples
t-SNE	*t-distributed Stochastic Neighbor Embedding*
tRNA	RNA transportador

Abreviações das espécies de mosquitos

Aaeg	*Aedes aegypti*
Aaeg2	Linhagem celular
Aalb	*Aedes albopictus*
Afur	*Aedes furcifer*
Ajap	*Aedes japonicus*
Alut	*Aedes luteocephalus*
Atay	*Aedes taylori*
Avex	*Aedes vexans*
Avit	*Aedes vittattus*

Haem	*Haemagogus sp.*
Sabe	*Sabethes sp.*

Abreviações de locais de coleta de mosquitos

BR	Brasil
FR	França
Lab	Laboratório
MClaros	Montes Claros
MG	Minas Gerais
NL	Netherlands
RJ	Rio de Janeiro
SE	Senegal
SI	Singapura
SJRioPreto	São José do Rio Preto
SU	Suriname

Abreviações de nomes de vírus

AAV	*Aedes anphevirus*
ANV	*Aslam narnavieus*
APV	*Aedes phasmavirus*
AejapAV1	*Aedes japonicus Anphevirus* 1
AejapBV1	*Aedes japonicus Bunyavirus* 1
AejapBV2	*Aedes japonicus Bunyavirus* 2
AejapNV1	*Aedes japonicus Narnavirus* 1
AejapRV1	*Aedes japonicus Rhabdovirus* 1
AejapTV1	*Aedes japonicus Totivirus* 1
AelutIFV1	*Aedes luteocephalus iflavirus* 1
AetayTV1	*Aedes taylori totivirus* 1
AevexBV1	*Aedes vexans Bunyavirus* 1
AevexCV1	*Aedes vexans chrysovirus* 1
AevitAV1	*Aedes vittattus anphevirus* 1
AevitPV1	*Aedes vittattus phenuivirus* 1
BRV	*Bahianus rhabdovirus*
CFAV	*Cell Fusing Agent virus*
CHIKV	*Chikungunya virus*
DENV	*Dengue virus*
GMV	*Guadeloupe mosquito virus*

GMV2	*Guadeloupe mosquito virus* 2
GMqlV	*Guadeloupe mosquito quaranja-like virus*
GSLV	*Guangzhou sobemo-like vírus*
HSRV	*Haemagogus sedoreovirus*
HTV	*Humaita-Tubiacanga virus*
IPLV	*Illomantsi partiti-like virus*
LTV	*Lactea totivirus*
NPLV	*Nyamuk partiti-like virus*
OVV	*Orbis virgavirus*
PCLV	*Phasi Charoen-like vírus*
RCV	*Rio Chico virus*
SFLV	*Sabethes flavi-related virus*
SGHLV	*Salivary Gland hypertrophy like-virus*
USUV	*Usutu virus*
WNV	*West Nile virus*
YFV	*Yellow fever virus*
ZIKV	*Zika virus*

LISTA DE SÍMBOLOS E UNIDADES

CH_3	Metil
Gb	Gigabase
Kb	Kilobase
Mb	Megabase
mm	Milímetro
$NaIO_4$	Periodato de sódio
nm	Nanômetro
O	Oxigênio
OH	Hidroxila
P	Monofosfato
pH	Potencial hidrogeniônico
Q	Valor *phred*
Å	Ångström
µg	Micrograma
µm	Micrômetro
2'	Carbono 2' da ribose na molécula de RNA
5'	Carbono 5' no extremo de fosfato da ribose na molécula de RNA
3’	Carbono 3' no extremo de hidroxila da ribose na molécula de RNA

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE APÊNDICES

SUMÀRIO

1 - Introdução

1.1 - Os vírus e a vida

Essa tese tem como objeto de estudo alguns dos agentes biológicos mais diversos e ubíquos que conhecemos em nosso planeta, os vírus. Apesar da relação intricada com a vida, os vírus não possuem todas as propriedades físico-químicas atribuídas a um organismo vivo. Definir o que é “vida” é uma tarefa que provoca cientistas e filósofos materialistas comprometidos em entender o nosso universo com base em observações, experimentos e evidências cientificas. Quando nossa espécie se vislumbra com grandes desafios como compreender a origem da vida; estender o tempo da vida humana; catalogar toda vida em nosso planeta; ou buscar vida em outros planetas, delimitações conceituais de vida se fazem necessárias para sistematizar os esforços que poderão nos levar a tais conquistas homéricas. Um comitê de pesquisadores organizado pela NASA propôs a seguinte definição operacional: “A vida é um sistema químico autossustentável passível de evolução Darwiniana” (NATIONAL AERONAUTICS AND SPACE ADMINISTRATION, 1994). A elegância dessa definição está em abranger dois aspectos fundamentais da questão imposta: o termodinâmico e o da transmissão de informação ao longo de gerações.

Organismos vivos, em sua unidade fundamental, as células, podem ser modelados como sistemas termodinâmicos abertos, ou seja, trocam matéria e energia com seu meio. Uma simplificação de uma das predições impostas pela segunda lei da termodinâmica é a de que todos os fenômenos naturais tendem a um aumento de entropia, em outras palavras, o aumento da desordem é uma tendência natural de todos os sistemas. Para uma célula, a consequência final do aumento constante da entropia é o estado de equilíbrio com o seu meio, no contexto biológico, a morte. Na didática explicação do prêmio Nobel em física, Erwin Schrödinger, os organismos vivos se mantêm longe do estado de equilíbrio com o meio se alimentando de “entropia negativa” (SCHRODINGER, 1951). A entropia negativa é obtida a partir da assimilação de moléculas organizadas e com alto teor energético. A manutenção da entropia positiva com entropia negativa para evitar o equilíbrio com o meio, em outras palavras, a manutenção da homeostase, ocorre através do metabolismo. Os vírus são metabolicamente inertes sem a maquinaria celular de seus hospedeiros, sendo essa uma das principais características que muitos utilizam para não classifica-los como vivos (MOREIRA; LÓPEZ-GARCÍA, 2009).

A replicação espontânea de estruturas moleculares a partir de um molde químico não é uma propriedade exclusiva da vida. Em nosso planeta, são conhecidas diversas estruturas químicas que, em determinadas condições, geram réplicas moleculares. Muitos cristais possuem essa propriedade. Na definição de vida em questão, o trecho: "...passível de evolução Darwiniana", se refere a um sistema genético, no qual a replicação das moléculas permite o acumulo de pequenos erros que podem ser replicados originando uma variedade de réplicas com diferentes níveis de estabilidade e capacidade de produzir mais cópias, isto é, com distintas "capacidades de adaptação" (*fitness)* em função da produção de novas cópias (BENNER, 2010). No livro "O gene egoísta", Richard Dawkins enumera três propriedades fundamentais dos replicadores que provavelmente iniciaram o processo de evolução biológica na Terra: longevidade, fecundidade e fidelidade de cópia (DAWKINS, 1990). Os vírus compartilham as propriedades de replicadores com os organismos vivos, garantindo a esses agentes infecciosos protagonismo na história da evolução da vida como conhecemos. Porém, todo seu processo de replicação é dependente das células hospedeiras, "vírus são produzidos, mas não auto-reproduzidos" (MOREIRA; LÓPEZ-GARCÍA, 2009).

Uma perspectiva relevante para sintetizar as comparações entre vírus e células em relação ao conceito de vida é a voltada para a autonomia das estratégias evolutivas de perpetuação do material genético. Vírus possuem uma estratégia evolutiva não autônoma, diferente das células (PROSDOCIMI et al., 2023). Nessa perspectiva, células não são sinônimo de vida. Vírus e células adotam estratégias diferentes de perpetuação do material genético, ambos considerados sistemas biológicos pelo fato de serem portadores e transmissores de informação biológica codificada, o que os distingue do mundo abiótico (DE FARIAS; JOSE; PROSDOCIMI, 2021) .

1.2 - Os vírus

Os vírus são parasitas intracelulares obrigatórios dependentes da maquinaria molecular de seus hospedeiros para replicação do próprio material genético, sendo capazes de infectar organismos pertencentes aos três domínios da vida. Estima-se que existam cerca de 10^{31} partículas virais habitando nosso planeta, superando numericamente a quantidade de formas de vida celulares conhecidas ("Microbiology by numbers", 2011; WIGINGTON et al., 2016). O termo "virosfera" é comumente usado

como referência a esse diverso e ubíquo conjunto de agentes biológicos em nosso planeta (KOONIN et al., 2021).

A relevância das pesquisas sobre vírus está relacionada ao entendimento da evolução e diversidade da vida, desenvolvimento de aplicações biotecnológicas e a prevenção de impactos socioeconômicos em áreas críticas para a manutenção da nossa sociedade como agricultura e saúde pública humana e animal. São conhecidos cerca de 200 vírus capazes de causar doenças a nossa espécie (FORNI et al., 2022; VIRALZONE, 2023). A recente pandemia causada pelo vírus SARS-CoV-2 (*Severe acute respiratory syndrome coronavirus* 2) (*Coronaviridae*) explicita as trágicas consequência de um surto viral em uma sociedade globalizada, evidenciado a importância da pesquisa básica sobre esses agentes infeciosos para a prevenção de novos surtos.

A palavra "vírus" tem sua origem do Latim e se refere a uma substância venenosa ou toxina, tendo sido utilizada para descrever patógenos. Apesar de até hoje, principalmente após uma recente pandemia, a palavra ter uma conotação negativa no senso comum, sendo associada a doenças, as funções e significância dos vírus vão muito além de causadores de doenças (PROSDOCIMI et al., 2023; ROOSSINCK, 2011).

Ainda que definidos como parasitas, uma categoria de interação ecológica interespecífica que acarreta ônus para o hospedeiro em termos de sobrevivência, há evidências de que certos vírus podem operar como comensais, isto é, não causando prejuízos significativos aos seus hospedeiros. Na espécie humana, os vírus circulares pertencentes às famílias *Anelloviridae* e *Redondoviridae* são exemplos de potenciais vírus comensais (LIANG; BUSHMAN, 2021). Há também possibilidade de relações simbióticas, nas quais existem benefícios mútuos como reforço da defesa imune dos hospedeiros contra novas infecções ou como armas biológicas que facilitam a obtenção de alimentos como na relação de vespas parasitoides e polidnavirus (ROOSSINCK, 2011).

As pesquisas sobre os vírus têm catalisado significativos avanços biotecnológicos. Uma das contribuições mais relevantes foi a descoberta da transcriptase reversa em retrovírus, realizada de forma simultânea e independente pelos pesquisadores David Baltimore e Howard Temin em 1970 (COFFIN, 2021). Tal

descoberta revolucionou a biologia molecular, os estudos de retrovírus e biologia do câncer. A capacidade de gerar um polímero de ácido desoxirribonucleico (DNA) a partir de um molde de ácido ribonucleico (RNA) *in vitro* é um processo essencial em inúmeros experimentos de biologia molecular usado em aplicações que vão desde a quantificação de RNAs alvos de células e de vírus por RT-qPCR, até etapas de complexos protocolos para construção de bibliotecas para sequenciamento de nova geração por RNA-seq. Dessa maneira, a descoberta da transcriptase reversa não apenas revolucionou os campos da biologia molecular e virologia, mas também serviu como alicerce para metodologias essenciais na pesquisa biotecnológica contemporânea.

1.2.1 - Origem dos vírus

Três hipóteses que explicam a origem dos vírus tem sido consideradas historicamente (KRUPOVIC; DOLJA; KOONIN, 2019). Na primeira, a hipótese do “"mundo primordial dos vírus", os vírus seriam descendentes diretos dos primeiros replicadores que surgiram antes das células. Na segunda, a da “origem redutiva” ou ”regressão”, os vírus teriam surgido a partir da degeneração de células ancestrais que perderam sua autonomia e se tornaram parasitas intracelulares obrigatórias. Na terceira, a hipótese dos “genes fugitivos” os vírus teriam múltiplas origens independentes a partir de genes de hospedeiros celulares que adquiriram a capacidade de replicação autônoma egoísta e infectividade.

A complexidade da origem dos vírus, provavelmente, extrapola os limites de explicação dos três modelos hipotéticos isolados, pois se trata de um grupo polifilético no qual diferentes táxons virais podem ter surgido de formas independentes (KRUPOVIC; DOLJA; KOONIN, 2019; PROSDOCIMI et al., 2023). O fato de que os genes que codificam proteínas envolvidas na replicação viral comumente não possuem homólogos próximos em organismos celulares conhecidos é uma evidência que corrobora a hipótese do “mundo primordial dos vírus" (KOONIN; SENKEVICH; DOLJA, 2006). Já a descoberta de vírus gigantes codificando quase todos os genes da maquinaria de tradução (ABRAHÃO et al., 2018) é uma evidência que pode corroborar a hipótese da “regressão”. Há uma hipótese de que a evolução dos vírus pode ter sido quimérica, com a maquinaria de replicação oriunda dos replicadores pré-celulares e as proteínas estruturais adquiridas em diferentes estágios evolutivos (KRUPOVIC; DOLJA; KOONIN, 2019).

1.2.2 - Dimensões virais

Os vírus apresentam uma variedade de formas e tamanhos. São menores do que limite visível a olho nu para os seres humanos, que é aproximadamente 100 µm (0,1 mm), e a maioria desses agentes biológicos também é menor do que o limite de tamanho para visualização com microscópio de luz, que é de cerca de 0,3 µm. Alguns dos menores vírus conhecidos, pertencentes a família *Circoviridae*, possuem entre 17 a 22 nm (0.017 a 0.022 µm) de diâmetro, representantes da família *Flaviviridae* possuem aproximadamente 50 nm, já os Poxvirus, os maiores vírus de vertebrados, possuem cerca de 0,3 µm, podendo ser observados em microscopia de luz (BURRELL; HOWARD; MURPHY, 2017). Tais dimensões microscópicas fizeram do emprego de técnicas como a microscopia eletrônica de transmissão, que permite a visualização de objetos de até 1nm (0,001 µm) algo essencial para avanços do conhecimento da forma e estrutura dos vírus. No que se refere às dimensões dos vírus, um grupo particularmente notável é o dos vírus gigantes que apresentam tamanhos excepcionais em relação aos padrões convencionais de tamanho viral. Vírus da família *Mimiviridae,* originalmente isolados de amebas*,* podem chegar a cerca de 0,75 µm de diâmetro e são maiores que algumas bactérias (XIAO et al., 2005).

1.2.3 - Partículas virais

Durante os diferentes estágios de seu ciclo de replicação, um vírus pode ser observado em diferentes formas com diferentes estruturas moleculares. As partículas virais infecciosas completas recebem o nome de vírion (BURRELL; HOWARD; MURPHY, 2017). O virion (**Figura 1**) é a forma extracelular de um vírus e tem como função proteger o material genético, garantindo a disseminação dos vírus entre hospedeiros. Essas estruturas são compostas pelo material genético viral (DNA ou RNA) envolvido por uma camada proteica chamada de capsídeo. O capsídeo é composto por centenas de capsômeros, subunidades proteicas idênticas interagindo não-covalentemente e organizadas regularmente. A utilização de uma grande quantidade de subunidades proteicas idênticas permite aos vírus a codificação

genética de macromoléculas com poucos genes. A ocorrência repetida das interações proteína-proteína dos capsídeos leva à montagem de estruturas simétricas.

Apesar da grande diversidade de forma dos vírus, dois padrões simétricos de estruturas virais são bem estabelecidos na literatura. O formato icosaédrico, um dos "sólidos platônicos" ou também, "poliedro regular", com 12 vértices, 30 arestas e 20

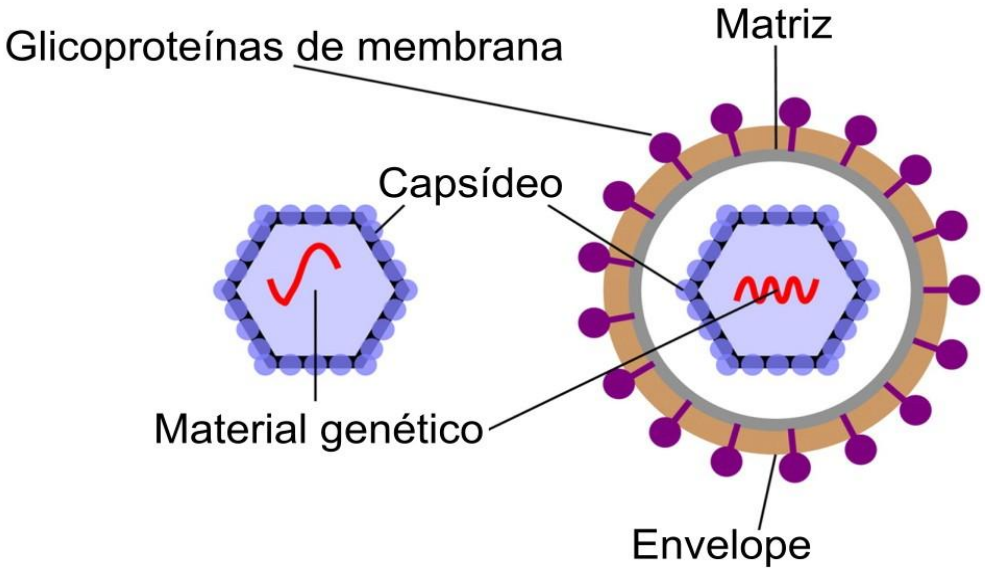


Figura 1 - Esquema representativo de uma partícula viral. Os virions podem ser não-envelopados (esquerda) ou envelopados (direita).

faces, sendo solução ótima para o problema de construir, a partir de subunidades repetidas, uma estrutura resistente que englobe um volume máximo com a quantidade mínima de energia (BURRELL; HOWARD; MURPHY, 2017). O formato icosaédrico é a estrutura tridimensional termodinamicamente mais estável de capsídeo viral, cerca de 60% dos táxons virais conhecidas possuem essa topologia (KRUPOVIC; DOLJA; KOONIN, 2019). O segundo padrão frequente é o de simetria helicoidal que dá origem a estruturas cilíndricas. Nas estruturas helicoidais, os genomas de RNA de alguns vírus podem formar uma espiral dentro do capsídeo, tal conformação é observada nos famosos Vírus do mosaico de plantas (BURRELL; HOWARD; MURPHY, 2017).

A estrutura formada pelo capsídeo e o material genético viral encapsulado é chamada de nucleocapsídeo. Em alguns vírus, o nucleocapsídeo pode ser revestido por uma membrana lipídica, chamada envelope (**Figura 1**), que contém proteínas virais que podem estar glicosiladas. O envelope viral é adquirido em um processo chamado brotamento, quando o nucleocapsideo é expelido através de uma membrana celular, podendo ser oriundo das membranas celulares externas, ou de membranas

internas de organelas e do núcleo celular de eucariotos. A maioria das proteínas localizadas na superfície externa do envelope viral são glicoproteínas com cadeias de carboidratos ligadas a aminoácidos específicos. As glicoproteínas podem ser encontradas como peplômeros ancorados à membrana ou espículas (*spike*), frequentemente organizadas em forma de dímeros ou trímeros (BURRELL; HOWARD; MURPHY, 2017). As glicoproteínas virais desempenham um papel crucial na interação dos vírus com as células hospedeiras e na patogênese das infecções virais, permitindo que os vírus se liguem a receptores nas células hospedeiras, facilitando a entrada do vírus na célula. Além de seu papel na entrada viral, as glicoproteínas também desempenham um papel importante na resposta imunológica do hospedeiro. Elas podem ser alvos de anticorpos neutralizantes produzidos pelo sistema imunológico, limitando a replicação viral e a disseminação da infecção, constituindo importantes alvos de estudos para o desenvolvimento de vacinas e terapias antivirais.

As diferenças das propriedades físico-químicas entre os vírus envelopados e não-envelopados são relevantes para o estabelecimento de estratégias de inativação das partículas virais como medidas sanitárias (LIN et al., 2020), manutenção de estoques virais; e extração de moléculas virais para fins de pesquisa. A integridade da camada lipídica dos vírus envelopados é crítica para a capacidade de infecção, logo a desestabilização ou dissolução do envelope lipídico com solventes apolares e detergentes levará a inativação das partículas virais. Os vírus não-envelopados são geralmente mais resistentes às condições ambientais e a agentes de inativação. A inativação de vírus não-envelopados é geralmente alcançada por meio de métodos que afetam diretamente o capsídeo viral, como alteração do pH, calor, radiação ultravioleta (UV) ou o uso de agentes químicos específicos que visam danificar ou desestabilizar a estrutura do capsídeo (SADRAEIAN et al., 2022).

A grande diversidade e complexidade de vírus se estende não só a topologia dos capsídeos, mas também a sua quantidade e organização em diferentes vírus. Vírus da família *Sedoreoviridae,* que possuem como representantes de importância médica os Rotavirus e de importância veterinária o arbovírus *Bluetongue virus*, não possuem envelope e podem apresentar um capsídeo de duas camadas. Os vírus, também não envelopados, das famílias *Partitiviridae* e *Chrysoviridae* possuem genomas segmentados e cada segmento é encapsidado separadamente. Esses vírus

são classificados como “multipartidos”, organização que resulta em partículas virais transmissíveis que não contêm toda a informação genética, fazendo da transmissão simultânea de diversas partículas virais para uma nova célula ou hospedeiro um requisito essencial para manter a integridade e replicação do genoma viral (SICARD et al., 2016).

1.2.4 - Capsídeos e a definição de vírus

Os vírus das famílias *Narnaviridae*, *Mitoviridae*, *Endornaviridae*, *Deltaflexiviridae, Hypoviridae* e do gênero *Umbravirus* são encontrados em fungos, plantas e invertebrados, não possuem capsideo, não formando virion e consequentemente, não possuem forma extracelular. Há evidências de que os Narnavirus se replicam no citoplasma e os Mitovirus no interior das mitocôndrias. Em fungos, esses vírus são transmitidos através de diferentes processos biológicos, como mistura citoplasmática e fusão de hifas, relacionados a reprodução e divisão desses organismos. Essas duas famílias virais possuem proximidade filogenética entre si, e apesar de infectarem eucariotos, possuem proximidade filogenética com a família de bacteriófagos *Leviviridae* (HILLMAN; ESTEBAN, 2009).

Os vírus sem capsídeo consistem em um desafio conceitual para a definição de vírus. Raoult e Forterre propuseram que os vírus são organismos que codificam virions, enquanto as formas de vida celulares são organismos que codificam ribossomos (FORTERRE, 2016; FORTERRE; KRUPOVIC; PRANGISHVILI, 2014; RAOULT; FORTERRE, 2008). Essa definição excluiria conceitualmente os vírus das famílias sem capsídeos da virosfera. Esses vírus constituem apenas uma das muitas exceções que frustram tentativas de definição de vírus. Koonin et al., 2021 propõe uma definição multidimensional na qual existe um “espaço virtual” composto pelos elementos genéticos replicadores e dentro desse espaço se encontra a virosfera, podendo ser essa dividida em “Ortovirosfera”, que engloba os vírus que possuem virions, e a “Perivirosfera “, um subespaço contendo os replicadores que não possuem todas características dos vírus “tradicionais” codificadores de virions (KOONIN et al., 2021).

1.2.5 - Material genético viral

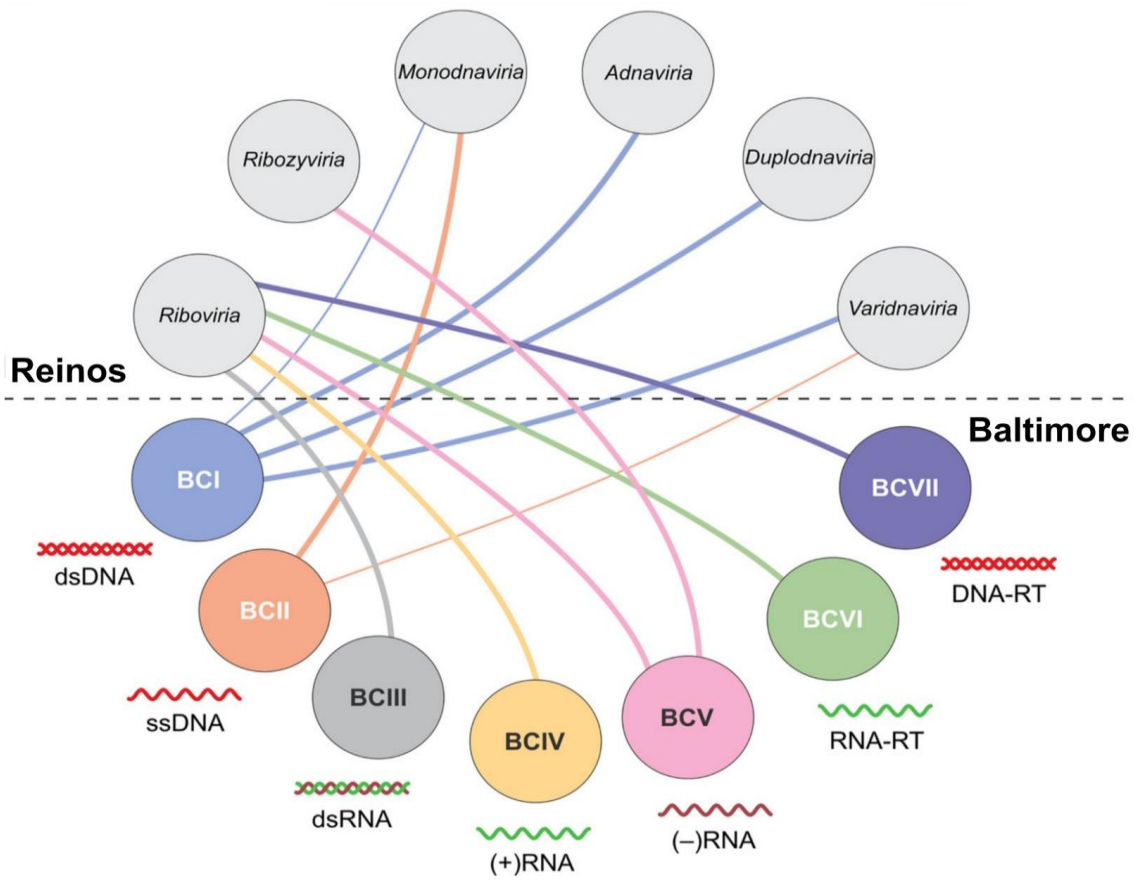


Figura 2 - Classificação de Baltimore e Reinos Virais. *As linhas coloridas representam as conexões entre as classes e os reinos virais monofiléticos. Linhas espessas representam relações fundamentais e linhas finas ocorrências excepcionais. Modificada de KOONIN et., al 2021.*

Os genomas virais podem ser armazenados na forma de DNA ou RNA, sendo encontrados na forma de fita dupla ou simples, podendo ser contínuos, segmentados ou circulares. A classificação viral de Baltimore divide os vírus em sete classes considerando a composição de ácidos nucleicos dos genomas que determinam as formas de transmissão e expressão da informação genética viral. Essa classificação continua relevante após mais de 50 anos de sua proposição, possuindo conexão com as atuais classificações filogenéticas dos grandes domínios virais monofiléticos (**Figura 2**). A relevância da classificação viral de Baltimore cruza a fronteira da virologia e fundamenta conceitualmente o que entendemos como processos de transferência de informação biológica na natureza (BALTIMORE, 1971; KOONIN; KRUPOVIC; AGOL, 2021).

A diferença de composição de ácidos nucleicos entre os vírus que infectam organismos pertencentes a cada um dos três grandes domínios da vida é evidente. Há uma predominância de vírus de RNA em eucariotos e de vírus de DNA, mais especificamente de dupla-fita, em procariotos (**Figura 3**). Não há registros de vírus de RNA em arqueias. Tal padrão de composição de viromas é uma das evidências para suportar a hipótese sintrófica de origem dos eucariotos na qual uma Deltaproteobactéria teria englobado um Argard arqueon que deu origem ao núcleo e, em um segundo evento, englobou uma Alphaproteobacteria que deu origem as mitocôndrias (KRUPOVIC; DOLJA; KOONIN, 2023).

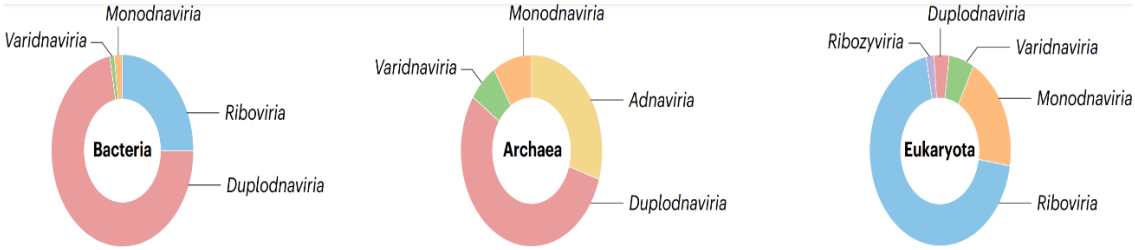


Figura 3 - Representação dos Reinos Virais nos três Domínios da Vida. Modificada de KUPROVIC, DOLJA & KOONIN, 2023.

Os genes virais podem ser divididos em codificantes de proteínas estruturais e não-estruturais. As proteínas estruturais são as utilizadas na construção do capsídeo, envelope e outros componentes do virion. As proteínas não estruturais estão envolvidas em processos como replicação do material genético viral, modificações químicas para construção do virion e evasão do sistema imune do hospedeiro.

A maior parte dos genomas virais é composta por genes codificadores de proteínas (MAHMOUDABADI; PHILLIPS, 2018). Virus possuem diferentes estratégias de condensação da informação genética que pode estar associadas a pressões de seleção natural relacionadas a velocidade e precisão da replicação e estabilidade dos genomas virais (DOMS, 2016; MAHMOUDABADI; PHILLIPS, 2018). Poliproteínas: uso de um único promotor para replicação de várias proteínas no mesmo mRNA que é traduzido e posteriormente as proteínas são individualizadas pela ação de proteases virais ou do hospedeiro. Essa estratégia é encontrada nos vírus da família *Flaviviridade,* caracterizando os genomas virais dos arbovírus de importância médica Dengue, Zika e Yellow fever (SIMMONDS et al., 2017). Alguns vírus utilizam a maquinaria da célula para realizar o splicing alternativo do próprio genoma de modo que um único promotor é utilizado para transcrever diferentes genes virais. Diversos

vírus de importância médica como membros das famílias *Orthomyxoviridae* e *Retroviridae* utilizam essa estratégia (https://viralzone.expasy.org/1943). Sobreposição de janelas de leitura (*reading frames*): Ao possuir códons de iniciação em diferentes posições do genoma, os genes virais podem se sobrepor com diferentes genes compartilhando mesmas porções de ácidos nucleicos, porém em janelas de leitura distintas. Tal sobreposição pode ocorrer na fita codificadora oposta, como no caso de alguns vírus ambigrámaticos da família *Narnaviridae* que possuem genes codificados nas fitas senso e antisenso (ABBO et al., 2023; DERISI et al., 2019; DINAN et al., 2020). *Ribosomal frameshifting*: Em alguns vírus, um ribossomo começa a traduzir uma proteína e, em seguida, "desliza" de volta uma base antes de começar novamente. Quando isso acontece, ele está agora em um novo frame de leitura e começa a traduzir uma proteína viral diferente. Essa estratégia é empregada por diversos vírus, incluindo vírus da família *Totiviridae* encontrados em insetos (ABBO et al., 2023), o arbovírus West nile vírus (*Flaviviridae*) e membros da família *Coronaviridae* (https://viralzone.expasy.org/860).

Alguns vírus possuem genes de RNAs não-codificantes com funções associadas a evasão do sistema imune e manutenção da latência viral (MAHMOUDABADI; PHILLIPS, 2018; TYCOWSKI et al., 2015). miRNAs codificados *pelo Ovine herpesvirus*-2 estão envolvidos na regulação da latência viral desse patógeno que pode ser fatal para ovinos infectados (RIAZ et al., 2014).

Virus do gênero Flavivirus (familia *Flaviviridae)* produzem um RNA não codificante conservado chamado de sfRNA (flaviviral subgenomic RNA). Esses ncRNAs são produtos da degradação incompleta do RNA genômico viral pela exorribonuclease celular 5'-3' XRN1, que não consegue clivar regiões altamente estruturadas e conservadas presentes no 3' UTR de flavivírus (SLONCHAK; KHROMYKH, 2018). Há evidências de que esses ncRNAs possuem grande influência na patogenicidade viral, adaptação ao hospedeiro e surgimento de novas cepas patogênicas. A transmissão do vírus da Zika por *A. aegypti* foi atenuada quando foram inseridas mutações na região do sfRNA (GÖERTZ et al., 2019). Além disso, os autores mostraram uma maior produção de siRNAs (21nt de dupla-fita) derivados do vírus com sfRNA mutado, evidenciando o papel desse ncRNA na evasão da resposta antiviral mediada por RNAi,

1.2.6 - Replicação viral

O ciclo de replicação dos vírus que infectam animais e produzem virion pode ser resumido em sete fases (LOUTEN, 2016):

1- Adsorção: O vírus se liga à superfície da célula hospedeira através de interações específicas entre proteínas virais e receptores celulares;

2- Penetração: O vírus entra na célula hospedeira, muitas vezes por endocitose, permitindo que seu material genético alcance o interior celular;

3- Desnudamento: O capsideo é desmontada e o material genético do vírus é liberado no citoplasma da célula hospedeira;

4- Replicação: O material genético viral é replicado dentro da célula hospedeira, produzindo cópias do genoma viral e iniciando a síntese de novas proteínas virais.

5- Montagem: As novas proteínas virais e as cópias do genoma são reunidas para formar novas partículas virais;

6- Maturação: As partículas virais recém-formadas são modificadas e organizadas, muitas vezes através de clivagem proteolítica, para se tornarem infecciosas;

7- Liberação: As partículas virais maduras são liberadas da célula hospedeira, muitas vezes destruindo a célula para que possam infectar outras células e continuar o ciclo de replicação, similar ao ciclo lítico dos fagos, ou as partículas virais podem ser exocitados pelo processo de brotamento, sem romper a membrana do hospedeiro, adquirindo um envelope oriundo de porções de membranas do hospedeiro.

A dinâmica de replicação dos vírus que infectam procariotos, chamados de fagos, pode ser resumida em ciclos lítico e ciclo lisogênico. Uma diferença marcante entre os vírus que infectam animais e os que infectam procariotos está na fase de fixação, pois os vírus de animas não possuem as caudas fibrilares que os fagos utilizam para se ligar porção externa das células. Outra diferença é que os vírus de animais não liberam seu material genético direto no citoplasma logo após a fixação.

Como uma das consequências da ausência de mecanismos de correção 3’ exonuclease das RNA-replicases da maioria dos vírus de RNA (MENÉNDEZ-ARIAS, 2009; STEINHAUER; DOMINGO; HOLLAND, 1992), esses vírus apresentam taxas de substituição de nucleotídeos maiores que as dos vírus de DNA e aproximadamente seis ordens de magnitude maiores do que aquelas em seus hospedeiros (SANJUÁN; DOMINGO-CALAP, 2016). Tais taxas de mutação constituem um fator determinante na diversidade genética, patogenicidade e adaptação a novos hospedeiros. Em uma

perspectiva de aplicação a saúde humana e animal, as altas taxas de mutação dos vírus de RNA constituem um desafio para o desenvolvimento de vacinas.

O ciclo de replicação dos retrovírus possui um intima relação com os genomas de seus hospedeiros. Esses vírus codificam em seus genomas e carregam em seus virions a enzima transcriptase reversa (RT), que é uma DNA polimerase dependente de RNA. A transcriptase reversa é capaz de transcrever reversamente o genoma de ssRNA em uma cadeia linear de cDNA complementar de dupla fita, que é então integrada a um cromossomo hospedeiro (LOUTEN, 2016). O retrovírus melhor estudado é o vírus da AIDS (acquired immunodeficiency syndrome), o HIV, human immunodeficiency vírus (*Retroviridae*).

Os retrovírus evidenciam a intrincada relação entre os vírus e a evolução dos genomas eucariotos. Os genes que codificam as proteínas Sincitina 1 e 2 envolvidas no processo de placentação em símios são oriundos de genes de envelopes virais de retrovírus (CORNELIS et al., 2012). Os genomas eucarióticos são permeados por elementos provavelmente oriundos de integrações virais como LINES, SINES, transposons, retrotransposons que podem chegar a compor 42% do genoma como na nossa espécie (LANDER et al., 2001). Esses elementos constituem matéria bruta para a ação de processos evolutivos que moldam os genomas eucarióticos, configurando exaptações que podem dar origem a novas funções biológicas.

1.2.7 - Elementos Virais Endógenos

Elemento Virais Endógenos (EVEs) são sequências de DNA derivadas de fragmentos ou genomas completos de vírus exógenos integradas no genoma de um hospedeiro. Uma vez que essas inserções acontecem em células germinativas, os novos *loci* podem ser herdados pelas gerações seguintes e fixados na população.

A integração genômica faz parte do ciclo de replicação dos retrovírus. O DNA viral integrado ao genoma hospedeiro durante o ciclo de replicação é chamado de provirus. A integração é autônoma, sendo mediada pela transcriptase reversa e proteínas integrases (WICKER et al., 2007) codificadas pelo próprio genoma retroviral. Tal fenômeno explica e existência das EVEs retrovirais. Porém, são encontradas EVEs oriundas de vírus pertencentes a todas as classes de Baltimore, sendo conhecidas também as EVEs não-retrovirais (HORIE et al., 2010; KATZOURAKIS; GIFFORD, 2010). A integração das EVEs não-retrovirais ainda se trata de um fenômeno pouco

compreendido, é provavelmente mediada por recombinação não-homóloga (ARBUCKLE et al., 2010) ou interação com outros retroelementos, como transposons (HORIE et al., 2010). O fato de que as EVEs não-retrovirais são comumente encontradas em loci genômicos associadas a elementos moveis é uma evidência de que a integração é mediada por tais elementos (AGUIAR et al., 2020).

Integrações de sequências virais podem ter efeitos deletérios, neutros ou benéficos aos hospedeiros. EVEs domesticadas podem adquirir novas funções e serem mantidas nos genomas dos hospedeiros por seleção natural. EVEs participam de eventos que vão desde a regulação da expressão genica (SOFUKU et al., 2018) até a plasticidade sináptica (MORTELMANS; WANG-JOHANNING; JOHANNING, 2016) em mamíferos. Há autores que argumentam que as EVEs podem fazer parte do sistema imune antiviral configurando um sistema imune adaptativo no combate a infecção por vírus circulantes. Os mecanismos subjacentes à imunidade derivada de EVE podem envolver interferência nos receptores celulares, reconhecimento de sequências de ácido nucleico (por exemplo, RNAi) ou até mesmo sabotagem da replicação por meio da produção de proteínas virais defeituosas a partir de EVEs (ASWAD; KATZOURAKIS, 2012). Essa hipótese é atrativa no contexto do estudo de importantes vetores de zoonoses virais, como os mosquitos. A remoção por CRISPR-Cas9 de uma EVE não-retroviral do genoma de *Aedes aegypti* levou ao aumento da replicação do vírus cognato, CFAV, nos ovários do mosquito (SUZUKI et al., 2020).

O estudo das EVEs pode levar a avanços na compreensão das intricadas relações entre vírus e hospedeiros, com implicações diretas em estudos sobre aspectos genômicos de vetores animais que ameaçam a saúde humana. Além disso, nos permite inferir o histórico de infecções virais que impactaram a evolução do genoma de hospedeiros. Como fosseis genômicos, as EVEs tem permitido que paleovirologistas tracem a origem de vírus, incluindo vírus de grande relevância atual, como os da família *Flaviviridae* (LI et al., 2022).

1.2.8 - Taxonomia Viral

Em 1977, Woose e Fox dividiram os procariotos em dois grandes grupos baseados em comparações de sequências de RNA ribossomal da subunidade menor (WOESE; FOX, 1977). Esse estudo levou ao estabelecimento do que conhecemos

hoje como os três domínios da vida*, Eukaryota*, *Archaea* e *Bacteria.* O modelo de árvore filogenética que suporta esses grandes grupos é corroborado por analises mais recentes e em larga escala (HUG et al., 2016). A existência de *uma "árvore da vida"* estabelecida com sequências de genes ribossomais nos permite inferir a existência de um ancestral comum universal (LUCA) e podemos considerar os genes ribossomais "marcadores universais" da vida em nosso planeta. Por se tratar de um grupo parafilético, os vírus não possuem um marcador universal similar, apresentando um cenário mais complexo de classificações filogenéticas e taxonômicas.

A taxonomia dos vírus é desenvolvida e supervisionada pelo Comitê Internacional de Taxonomia de Vírus (ICTV). A classificação e nomenclatura taxonômica dos vírus possui papel fundamenta na área clínica, agricultura, ensino e para agências reguladoras de diversos setores (SIDDELL et al., 2023). Atualmente, as 11273 espécies virais reconhecidas pelo ICTV são classificadas em um sistema taxonômico de 15 categorias (*ranks*). O princípio guia para a recente organização taxonômica é a congruência com as histórias evolutivas inferidas para cada vírus, respeitando o princípio de monofiletismo e características genômicas apropriadas (SIMMONDS et al., 2023). Aspectos ecológicos e fenotípicos também são usados para refinamento das fronteiras das categorias.

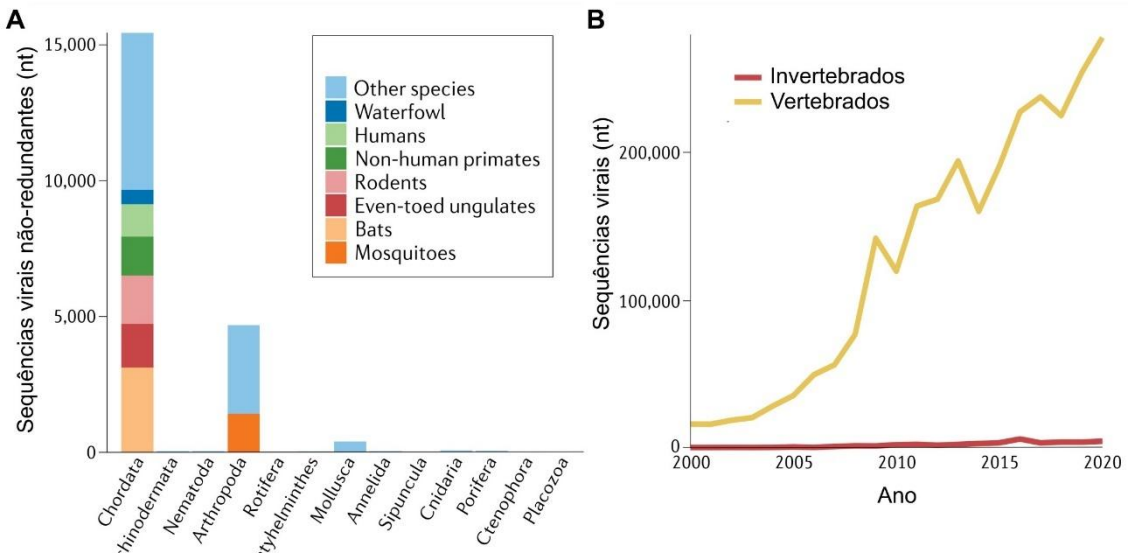


Figura 4 - Representatividade de sequências virais por filo animal. A. Representação gráfica do número de entradas únicas de nucleotídeos de vírus no banco de dados nt do GenBank (NCBI), ordenadas por espécie de hospedeiro, demonstrando que os vírus associados a cordados superam em número os advindos de todos os outros filos animais. **B**. Representação gráfica do rápido aumento nas entradas de vírus associados a vertebrados no banco de dados de sequências do GenBank nas últimas duas décadas e os números comparativamente baixos de vírus associados a invertebrados identificados no mesmo período**.** Modificada de HARVEY & HOLMES, 2022.

A taxonomia é uma ciência dinâmica, no caso dos vírus a fluidez pode ser ainda maior. Revisões baseadas em novos conjuntos de dados são a regra na taxonomia viral. Tal fenômeno é impulsionado pela aplicação do sequenciamento de ácidos nucleicos em larga escala em estudos metagenômicos aliado a avanços constantes na construção de ferramentas de bioinformática que permitem a mineração de sequências virais em dados públicos (EDGAR et al., 2022; ROUX; MATTHIJNSSENS; DUTILH, 2021). Uma evidência do descompasso gerado por tais aplicações é o fato de que a base de dados Taxonomy do NCBI (https://www.ncbi.nlm.nih.gov/taxonomy acessado em 20/08/2023) registra mais de 239 mil vírus taxonomicamente distintos, mais de 20 vezes o número de espécies virais aceitas pelo ICTV. Esforços para sistematizar e padronizar a classificação de sequências advindas de estudos genômicos e mineração de dados tem sido tomados e são críticos para a produção de bases de dados robustas e precisas que permitirão avanços na virologia (ADRIAENSSENS et al., 2023) .

Apesar do aumento exponencial de sequências virais disponíveis em bancos de dados que tem expandido nossa compreensão da virosfera, um viés de amostragem para vírus oriundos de hospedeiros vertebrados é evidente (**Figura 4**) (HARVEY; HOLMES, 2022). Harvey & Holmes, 2022 ressaltam que esse viés restringe nosso conhecimento da diversidade viral e limita avanços na compreensão da evolução viral, incluindo a compreensão do fenômeno de saltos entre hospedeiros que podem levar a emergência de novas zoonoses.

1.3 - Os arbovírus

Zoonoses são doenças infecciosas transmitidas de animais não humanos para humanos (https://www.who.int/news-room/fact-sheets/detail/zoonoses). Estima-se que 60% das doenças infecciosas conhecidas e até 75% das doenças infecciosas novas ou emergentes tenham origem zoonótica (SALYER et al., 2017). Arboviroses são zoonoses causadas especificamente por arbovírus e transmitidas por artrópodes.

Os arbovírus (do inglês “*arthropod borne-viruses*”) constituem um grupo de vírus relevante para pesquisas em saúde. Esses vírus patogênicos se replicam tanto em vertebrados quanto em invertebrados. Os arbovírus podem ser patogênicos para humanos e outros animais, como é o caso do vírus da dengue transmitido por mosquitos, o vírus da febre hemorrágica da Crimeia-Congo transmitido por carrapatos e o vírus da língua azul transmitido por midges. A transmissão de arbovírus ocorre,

principalmente, quando um artrópode vetor hematófago, como mosquitos e carrapatos, se alimentam do sangue de vertebrados suscetíveis à infecção. Sob uma perspectiva médica e veterinária, a expansão geográfica das arboviroses é preocupante. Essas doenças são predominantemente transmitidas por mosquitos, incluindo a dengue, zika, febre amarela, febre do Nilo Ocidental, encefalite japonesa, febre chikungunya e febre do Vale do Rift (WILDER-SMITH et al., 2017) .

O *Arbocat*, uma base de dados mantida pelo *Centers for Disease Control and Prevention* (CDC), registra 537 arbovírus (https://wwwn.cdc.gov/arbocat/). Recentemente, essa contagem foi revisada com dados recentes da literatura e sequências genômicas virais resultando em um total de 615 arbovírus, distribuídos em 460 espécies virais distintas (HUANG et al., 2023). Cerca de 100 desse arbovírus infectam humanos (ROSENBERG et al., 2013).

Arbovírus são um grupo ecológico e polifilético de vírus que pertencem a diversas famílias e gêneros virais. Com exceção do *African swine fever vírus (Asfaviridae),* todos arbovírus são vírus de RNA, sugerindo que as altas taxas de mutação dos genomas de RNA são um forte pré-requisito para estabelecimento de um ciclo de replicação alternante nos distintos ambientes representados por hospedeiros vertebrados e invertebrados (HANLEY; WEAVER, 2008).

1.4 - Mosquitos

Insetos (Reino Metazoa; Filo Arthropoda; subfilo Hexapoda; classe Insecta) são animais invertebrados caracterizados por um corpo segmentado (cabeça, tórax e abdômen), três pares de pernas articuladas, um par de antenas e, geralmente, um ou dois pares de asas. Constituem o maior e mais diverso grupo de animais, contendo mais de 70% de todas as espécies conhecidas (MISOF et al., 2014). Essa vasta quantidade de potenciais hospedeiros é acompanhada por uma grande diversidade viral (LI et al., 2015). Análises da diversidade viral em insetos podem fornecer importantes dados sobre a evolução dos vírus, pois, em circulação nesse grupo, são encontrados representantes da maioria de famílias virais conhecidas em animais (SHI et al., 2016).

Mosquitos são insetos da ordem Díptera (*di* = duas, *pteron* = asas), caracterizados, principalmente, por um par de asas funcionais membranosas no mesonoto; asas posteriores reduzidas a halteres; protórax e metatórax reduzidos;

adultos com peças bucais sugadoras, frequentemente adaptadas para perfurar (CARVALHO et al., 2012). Os dípteros constituem uma ordem megadiversa de insetos holometábolos (THOMPSON, 2008). Essa ordem é dividida em duas subordens, Brachycera (*brachy* = curto, *cera* = antena, chifre), na qual são incluídos os dípteros comumente chamados de “moscas”, e Nematocera (*nemato* = fio, alongado), caracterizada, principalmente, pelas antenas longas de seus integrantes.

1.4.1 - Família *Culicidae*

A família *Culicidae,* pertencente a subordem Nematocera, inclui os insetos comumente chamados de mosquitos, pernilongos, muriçocas. Estão descritas aproximadamente 3610 espécies de culicídeos, divididos em 178 gêneros. Esses insetos possuem comprimento variando entre 3 a 9 mm; escamas revestindo a maior parte do corpo e pernas dos adultos; probóscide fina e mais longa que a cabeça, com peças bucais alongadas alocadas em um estojo formado pelo lábio (CARVALHO et al., 2012). As larvas e pupas dessa família são aquáticas, sendo as larvas nadadoras ativas em águas lentas. As larvas de mosquitos do gênero *Toxorhynchites* (subfamília Toxorhynchitinae), também conhecidos como “mosquitos elefante”, podem ser predadoras de larvas de outros mosquitos, podendo ser usadas para controle biológico de mosquitos vetores de arbovírus (MALLA et al., 2023). Os adultos dessa família são comumente mais ativos durante o crepúsculo ou à noite, sendo as fêmeas hematófagas, geralmente necessitando de sangue para amadurecimento dos ovos. Os machos se alimentam de néctar e outros líquidos de origem vegetal. Os gêneros *Anopheles* (subfamília Anophelinae), *Aedes* (subfamília Culicinae; tribo Aedini) e *Culex* (subfamília Culicinae; tribo Culicini) recebem destaque por incluírem mosquitos vetores de agentes etiológicos causadores de doenças a espécie humana e a outros animais. Os mosquitos mantêm uma relação evolutiva intrincada com a nossa espécie, desempenhando um papel crucial como agentes de seleção natural para a resistência a parasitas. Além disso, eles exerceram uma influência marcante na trajetória da nossa história, emergindo como elementos adversos que definiram batalhas históricas e empregados como armas biológicas (SNYDER, 2020).

1.4.2 - Gênero *Aedes sp.*

O gênero *Aedes sp.* (Meigen, 1818) é o maior gênero da tribo Aedini, compreendendo 932 espécies (“Aedes Meigen, 1818 WRBU”, [s.d.]). A taxonomia do grupo é complexa, tendo passado por amplas revisões (WILKERSON et al., 2015).

Estudos utilizando marcadores moleculares tem contribuído para a compreensão das relações filogenéticas entre as espécies e para traçar a história da dispersão global desses mosquitos que tiveram sua origem na África (COOK et al., 2005; GLORIA-SORIA et al., 2016; ZADRA; RIZZOLI; ROTA-STABELLI, 2021).

Apenas os genomas de *A. aegypti* (MATTHEWS et al., 2018) e *A. albopictus* (PALATINI et al., 2020) estão sequenciados. Recentemente, uma versão “rascunho” do genoma de *A. koreicus* (KURUCZ et al., 2022) foi disponibilizada no NCBI. Os genomas de *A. aegypti* e *A. albopictus* possuem tamanhos estimados similares de ~1.3Gb, ambos organizados em três cromossomos. Dentre as muitas características conservadas ou apenas similares entre os dois genomas, a grande porcentagem de elementos repetitivos e transposons é uma marca de ambas as espécies, com 65% para *A. aegypti* e 55% para *A. albopictus*.

1.4.3 - Principais transmissores de arbovírus do planeta

A palavra Aedes tem origem grega, podendo ser traduzida como odioso ou desagradável. O nome do gênero oriundo de tal adjetivo é explicado pela relação desses insetos com a espécie humana. Os mosquitos do gênero *Aedes* são os vetores de arbovírus mais importantes devido a sua ampla distribuição geográfica e adaptação a ambientes antropomorfizados (POWELL; TABACHNICK, 2013). São transmissores de arbovírus de relevância global como *Dengue virus* (DENV), *Zika virus* (ZIKV) (*Flaviviridae*) e chikungunya virus, além de outros que infectam milhares de pessoas em regiões tropicais e subtropicais todos os anos causando grandes prejuízos socioeconômicos. Apenas nas últimas duas décadas, testemunhamos a emergência de três arbovírus transmitidos por mosquitos nas Américas: a chegada do *West Nile Virus* (WNV) na América do Norte em 1999 (WEAVER; REISEN, 2010); a emergência e espalhamento do *Chikungunya virus* (CHIKV) (*Togaviridae*) nas Américas em 2013 (ZELLER; VAN BORTEL; SUDRE, 2016); e o surto de ZIKV no Brasil em 2015, cuja associação com casos de microcefalia (MLAKAR et al., 2016) e de Síndrome Guillain-Barré (ARAUJO; FERREIRA; NASCIMENTO, 2016) desencadeou uma crise de saúde internacional.

O vírus da Dengue é o arbovírus de distribuição geográfica mais ampla, causando aproximadamente 390 milhões de infecções e 20 mil mortes por ano (BHATT et al., 2013). Levantamentos de anos anteriores mostram que a soma dos

prejuízos diretos e indiretos causados por arboviroses, apenas no Brasil, podem chegar à 2,3 bilhões de reais por ano (TEICH; ARINELLI; FAHHAM, 2017). Em um cenário no qual o número de casos tem aumentado anualmente, os gastos reais, provavelmente chegam a valores ainda maiores.

O aumento dos casos de doenças causadas por vírus transmitidos por mosquitos em áreas endêmicas está intrinsecamente ligado ao aumento da densidade populacional humana, acelerada urbanização, a ausência de programas de vigilância eficientes e à emergência de resistência na população de mosquitos aos inseticidas químicos frequentemente empregados. A implementação restrita de estratégias de controle vetorial adequadas, somada à carência geral de conscientização da população quanto à adoção de medidas profiláticas pertinentes, também contribui para tal cenário (LIGSAY; TELLE; PAUL, 2021) .

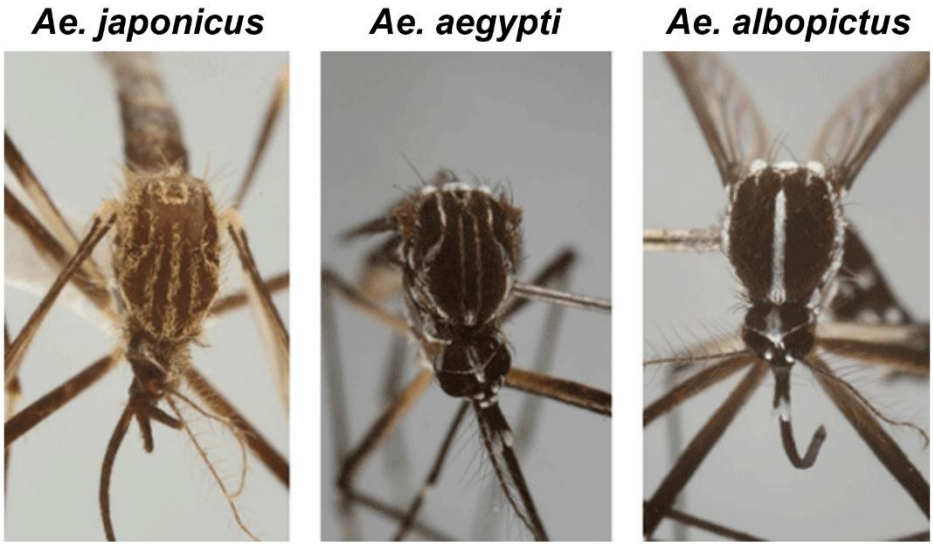


Figura 5 - Fotos ilustrativas dos mosquitos invasores das espécies, *Ae. japonicus*, *Ae. aegypti* e *Ae. albopictus*, transmissores de arbovírus. Padrões e formatos de listas na porção dorsal do tórax distinguem essas três espécies. Fotos de Lyle Buss, Universidade da Florida (fonte: https://entnemdept.ufl.edu/creatures/AQUATIC/aedes_japonicus.html).

Devido à globalização e mudanças climáticas causadas por nossa espécie, a distribuição geográfica de mosquito vetores como *A. aegypti* e *A. albopictus* (**Figura 5**), os dois vetores de arbovírus mais importantes do gênero, está se expandindo continuamente, expondo cada vez mais pessoas à ameaça de arbovírus ao longo dos anos (KRAEMER et al., 2019). Atualmente, cinco espécies de mosquitos invasores, potenciais transmissores de arbovírus, do gênero *Aedes* estão estabelecidas na

Europa: *A. albopictus*, *A. japonicus*, *A. koreicus*, *A. atropalpus* e *A. aegypti*, sendo as duas últimas raramente detectadas (MEDLOCK et al., 2015). Dentre esses, *A. albopictus* é o vetor de arbovírus mais disseminado na Europa, com casos autóctones de Chikungunya e dengue recentemente documentados na Itália e França atribuídos a transmissão por mosquitos dessa espécie (COCHET et al., 2022; LAZZARINI et al., 2020). Outras duas espécies invasoras potenciais transmissoras de arbovírus, incluindo Zika e dengue, com ampla distribuição global são *A. vexans* e *A. vittattus* (OUTAMMASSINE; ZOUHAIR; LOQMAN, 2022). Em 2021, *A. vittatus* foi manchete nos noticiários brasileiros devido a recente ocorrência desse vetor de arbovírus que representa uma nova ameaça à saúde pública humana nas américas (ALARCÓN-ELBAL et al., 2020).

As espécies invasoras não são as únicas responsáveis pela manutenção dos ciclos de transmissão de arbovírus. Na África, as espécies *A. furcifer*, *A. taylori,e A. luteocephalus* são componentes chaves dos ciclos selváticos de transmissão de arbovírus (DIALLO et al., 2022; VALENTINE; MURDOCK; KELLY, 2019). Nas américas, as espécies nativas dos gêneros *Haemagogus* (tribo Aedini) e *Sabethes* (tribo Sabethini) são transmissores do YFV para humanos e primatas não humanos, mantendo o ciclo silvático desse vírus (VALENTINE; MURDOCK; KELLY, 2019).

É importante ressaltar que, para a maioria das arboviroses, não há disponibilidade de vacinas ou tratamentos antivirais eficazes, fazendo do controle vetorial a medida mais eficaz para combater doenças transmitidas por mosquitos vetores (WORLD HEALTH ORGANIZATION, 2017). Diante de um cenário de rápida expansão global e aumento da resistência a inseticidas desses vetores, métodos de controle inovadores são necessários. Recentes avanços em estudos sobre a fisiologia e resposta imune antiviral desses mosquitos tem mostrado que bactérias e outros vírus podem alterar a competência vetorial desses mosquitos para arbovírus de importância médica. A identificação de agentes biológicos com esse potencial é um pré-requisito fundamental para o estabelecimento de novas estratégias de controle vetorial.

1.5 - A bactéria *Wolbachia*

*Wolbachia*s são bactérias gram-negativas, endossimbiontes, obrigatoriamente intracelulares, pertencentes ao filo Proteobactéria, subfilo alfaproteobactéria, ordem

Rickettsiales e família Anaplasmataceae (O'NEILL et al., 1992). *Wolbachia*s foram descobertas por Hertig e Wolbach em 1924 analisando tecidos reprodutivos de *Culex pipiens*. Estima-se que 40% dos artrópodes e 65% dos insetos sejam naturalmente infectados por *Wolbachia* (ZUG; HAMMERSTEIN, 2012), fazendo dessas bactérias os endossimbiontes mais abundantes de invertebrados. Assim como as mitocôndrias, esses organismos são transmitidos verticalmente por herança materna. *Wolbachia*s causam fenótipos que tendem a favorecer o "*fitness*" das fêmeas infectadas, consequentemente aumentando a transmissão das *Wolbachia* (O'NEILL et al., 1992). Os fenótipos mais conhecidos causados nos hospedeiros são: indução de partenogênese (ARAKAKI; MIYOSHI; NODA, 2001); feminização de machos genéticos (NARITA et al., 2007); morte dos embriões machos (DYER; JAENIKE, 2004); e incompatibilidade citoplasmática que reduz a prole de machos infectados e fêmeas não infectadas (BECKMANN; RONAU; HOCHSTRASSER, 2017).

A proteção contra a infecção por vírus nos hospedeiros é um fenótipo causado por *Wolbachia*s amplamente explorado, principalmente pela possibilidade de controlar insetos vetores de patógenos humanos. Estudos com *Drosophila melanogaster* mostram que algumas *Wolbachia*s podem proteger a mosca de infecções por vírus de RNA (TEIXEIRA; FERREIRA; ASHBURNER, 2008). O caso de maior destaque de resistência por vírus causado por *Wolbachia* é com *A. aegypti*. Esse mosquito não parecem ser naturalmente infectado por *Wolbachia,* apesar de alguns relatos inconsistentes de detecção (DA SILVA et al., 2021; ROSS et al., 2020). Estudos prévios tinham a intenção de usar linhagens de *Wolbachia* de *D. melanogaste, wMelPop-CLA,* para diminuir o tempo de vida do mosquito (MCMENIMAN et al., 2009). A infecção artificial teve sucesso, reduzindo em até 50% o tempo de vida dos mosquitos e também gerou um fenótipo de resistência aos arbovírus Dengue e Chikungunya (MOREIRA et al., 2009).

A. aegypti infectados com *Wolbachia* já são utilizados em nível de campo com apoio de iniciativas internacionais (https://www.worldmosquitoprogram.org/) para estabelecimento da infecção artificial dessa bactéria como controle biológico global desse vetor. O estudo realizado na cidade de Niterói, RJ, Brasil, mostrou que que a *Wolbachia* pode ser introduzida em populações de *Ae. aegypti* em áreas urbanas, resultando na redução da incidência de doenças transmitidas por este vetor com o caso da Dengue, Zika e Chikungunya (PINTO et al., 2021) .

1.6 - Os vírus específicos de insetos

Insetos também hospedam vírus específicos (ISVs, do inglês “insect specfic viruses”) (ROUNDY et al., 2017), que são capazes de infectar vertebrados. Diferente dos arbovírus que são mantidos na natureza principalmente por transmissões horizontais que estabelecem seus ciclos de transmissão, a manutenção dos ISVs na natureza ocorre de outras formas. Há evidências de que a principal seja a transmissão vertical, da fêmea para a prole (BOLLING et al., 2012). Assim como observado para alguns arbovírus, a transmissão de ISVs, também pode ocorrer de forma horizontal venérea (entre machos e fêmeas) (MAVALE et al., 2005). Além dessas possibilidades, há uma hipótese de que durante a alimentação de néctar, mosquitos podem adquirir ou transmitir vírus (ROUNDY et al., 2017), porém ISVs nunca foram reportados infectando plantas. O fato de não se replicarem em células de vertebrados, faz desses vírus possíveis alvos para estratégias de controle biológico de mosquitos vetores de arbovírus. Somado a isso, as evidências de que ISVs podem modular a competência vetorial dos insetos, diminuindo ou aumentando a replicação de arbovírus de importância médica, tem aumentado ao longo dos anos (BAIDALIUK et al., 2019; BOLLING et al., 2012; OLMO et al., 2023; ZHANG et al., 2017).

O primeiro ISV foi caracterizado em 1975. O vírus chamado de Cell-fusing agent vírus (CFAV) (*Flaviviridae*) foi isolado de uma cultura de células de *A. aegypti* e recebeu esse nome devido ao efeito citopático que induz nas células infectadas (STOLLAR; THOMAS, 1975). Quase 30 anos se passaram até a caracterização do próximo ISV, Kamiti River Virus (KRV) (*Flaviviridae*), isolado de mosquitos de campo (CRABTREE et al., 2003). Nos últimos anos, o sequenciamento em larga escala de RNA e DNA tem sido comumente utilizado em estudos metagenômicos para avaliar a diversidade genética de vírus em uma amostra biológica, referida como o Viroma, permitindo estudos de vírus e outros organismos sem a necessidade de cultivo em laboratório. A descoberta de novos ISVs foi impactada por esses avanços, expandindo consideravelmente o conhecimento sobre sua diversidade (DE ALMEIDA et al., 2021). O aumento de sequências de ISVs nas bases de dados tem permitido análises filogenéticas que evidenciam grande proximidade evolutiva entre ISVs e arbovírus de importância clínica e veterinária. Há uma hipótese de que muitos arbovírus evoluíram a partir de ISVs que se adaptaram aos hospedeiros vertebrados dos quais seus hospedeiros insetos se alimentavam do sangue (ROUNDY et al., 2017). Considerando

as evidências filogenéticas que suportam essa hipótese e as altas taxas de mutação do vírus (SANJUÁN, 2012), é possível inferir que esse fenômeno ainda ocorra, originando novos patógenos a partir de ISVs circulantes. Nesse contexto, a descoberta e caracterização de ISVs pode contribuir para a compreensão de mecanismos que determinam a gama de hospedeiros que um vírus pode vir a infectar, estabelecendo bases de conhecimento para estratégias de contenção de novos surtos virais.

1.7 - Metagenômica viral

Metagenômica pode ser definida como a análise da diversidade biológica baseada em sequências dos genomas contidos em uma amostra (HANDELSMAN et al., 1998). As novas tecnologias de sequenciamento em larga escala de ácidos nucleicos somadas ao desenvolvimento de ferramentas em bioinformática impulsionaram estudos metagenômicos que vem elucidando viromas em diversas amostras biológicas e permitindo maior sensibilidade nas análises, pois devido à alta profundidade dos sequenciamentos, vírus presentes em baixas quantidades em uma amostra podem ser detectados. As etapas mais comuns de um estudo de viroma estão ilustrados na (**Figura 6)**.

A metagenômica em larga escala utilizando sequenciamento de nova geração aplicada a organismos celulares tem como possibilidade o uso de marcadores universais como os genes das subunidades menores 16S (procariotos) e 18S (eucariotos) que permitem concentrar os esforços e gastos de sequenciamento em amplicons contendo regiões hiper variáveis dessas sequências especificas que permitirão inferências sobre a diversidade de organismos em uma amostra (HINO; MARUYAMA; KIKUCHI, 2016; LANGILLE et al., 2013). Os vírus não possuem tais marcadores universais. Apesar de todos os vírus de RNA (Riboviria) possuírem como homólogos os genes das replicases RdRP e transcriptase reversa (KRUPOVIC; DOLJA; KOONIN, 2019) , esses marcadores seriam limitados a vírus de RNA e a divergência desses genes a nível de sequência de nucleotídeos impossibilitaria o desenho de primers generalistas para geração de amplicons. Dessa forma, a metagenômica viral é embasada em variações da técnica de “shotgun”, que consiste

no sequenciamento não direcionado de todos os genomas presentes em uma amostra (ROUX; MATTHIJNSSENS; DUTILH, 2021).

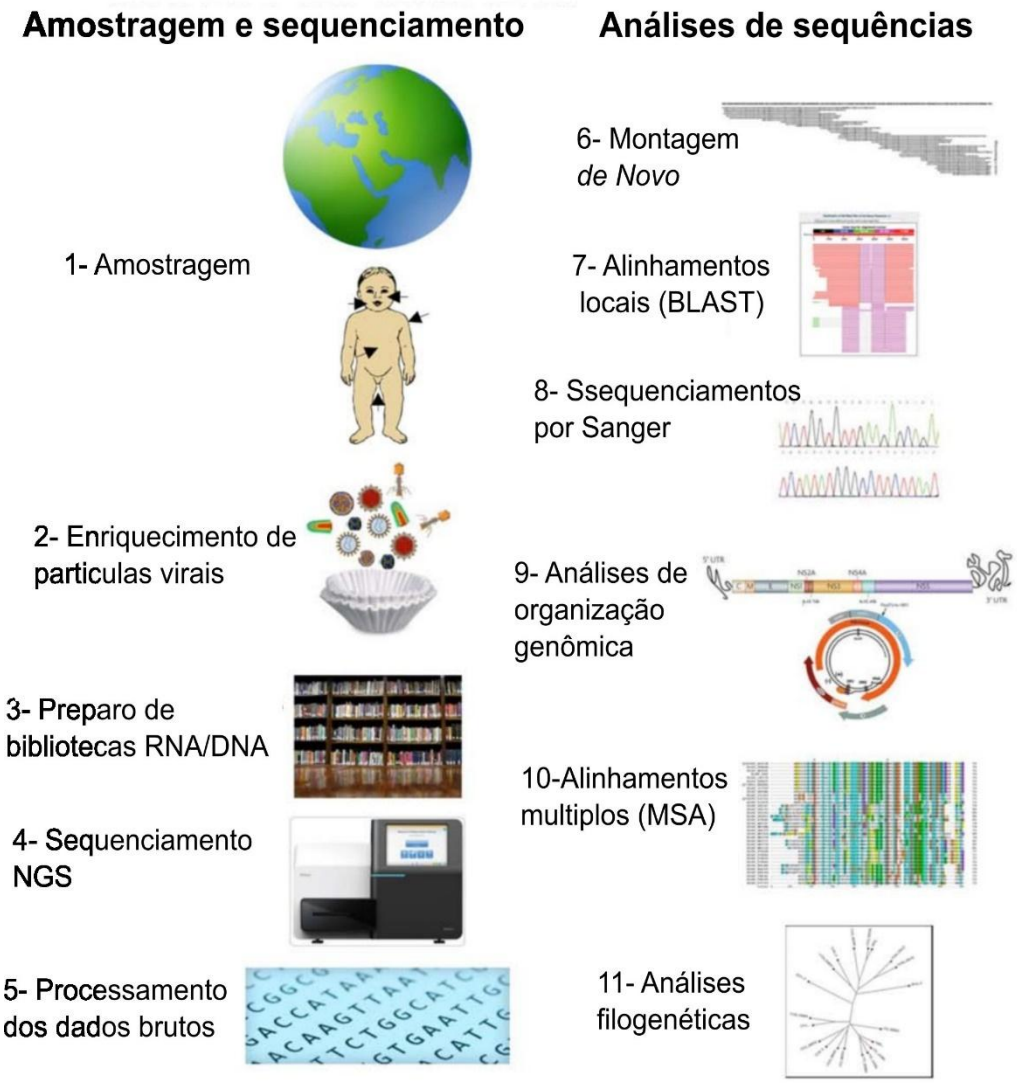


Figura 6 - Passos comuns em estudos de metagenômica viral utilizando sequenciamento de Nova Geração. Os experimentos começam com o preparo das amostras e geração de dados de sequenciamento (coluna à esquerda) e seguem com as etapas de bioinformática envolvendo caracterização genética das sequências virais identificadas (coluna à direta). Modificado de DELWART, 2013.

Os sequenciadores de nova geração sequenciam milhões de pequenos fragmentos de ácidos nucleicos produzindo arquivos contendo leituras (*reads*) correspondentes as sequências desses fragmentos. As *reads* são sobrepostas utilizando algoritmos de montagem de sequências e quando há sobreposição de fragmentos de diferentes *reads,* essas podem compor fragmentos contínuos maiores oriundos de uma mesma sequência original de ácidos nucleicos (*contigs*). Com os *contigs* montados, é possível inferir a origem viral da sequência. Para isso, são realizadas comparações por similaridade de sequências com sequências virais

conhecidas depositadas em bancos de dados de referência como o GenBank (FANCELLO; RAOULT; DESNUES, 2012).

Essas comparações são realizadas por meio de alinhamentos locais. Comumente, utiliza-se o programa BLAST (*Basic local alignment search tool*) (ALTSCHUL et al., 1990), cuja abordagem heurística não garante o melhor alinhamento, mas encontra alinhamentos estatisticamente significativos em tempo viável. Para avaliar a significância estatística do alinhamento local, usa se o parâmetro *Expect value* (E-value), que representa o número de alinhamentos que se espera encontrar ao acaso quando se busca por alinhamentos em uma base de determinado tamanho. Os *contigs* montados podem ser alinhados como nucleotídeos (nt) ou aminoácidos (aa). Os *contigs* com alinhamento significativo com uma sequência viral, são identificados como virais, podendo ser associados a um vírus específico e/ou gene viral. Os *contigs* sem alinhamentos significativos com sequências de nucleotídeos podem ser oriundos de vírus distantes evolutivamente dos contidos nos bancos de dados, podendo apresentar similaridade de sequência apenas em nível de aa. Para tal avaliação, comumente, usa se o programa BLASTx, que realiza a tradução dos *contigs* nas 6 possíveis janelas de leitura (*frames*) e realiza alinhamento local de cada uma das 6 sequências de aa geradas para cada *contig* contra um banco de dados de sequências de aa. O programa BLASTx possui alto custo computacional, principalmente em relação ao seu tempo de execução, muitas vezes consistindo em uma barreira para laboratórios sem estruturas computacionais para essa etapa de análise metagenômica. Recentemente, o programa DIAMOND (BUCHFINK; REUTER; DROST, 2021) tem sido uma alternativa viável, mantendo alta sensibilidade dos alinhamentos locais e recursos taxonômicos similares ao BLAST.

Genomas virais são geralmente pequenos e evoluem rápido, principalmente os de vírus de RNA (SANJUÁN; DOMINGO-CALAP, 2016). Estima se que o programa BLAST seja uma ferramenta eficiente na busca de sequências proteicas homólogas quando os alinhamentos possuem identidade maior que 30% sobre o tamanho total de sequência proteicas ou domínio proteico contidos em uma base de dados. A capacidade da ferramenta de encontrar homólogos verdadeiros reduz significativamente quando esse limiar de 30% não é alcançado (BRENNER; CHOTHIA; HUBBARD, 1998). No contexto da metagenômica, esse problema é ainda maior, pois nem sempre se consegue montar genomas virais completos a partir de

pequenas *reads*. Dessa forma, mesmo após alinhamentos em nível de aa, ainda podem restar *contigs* oriundos de vírus tão divergentes dos presentes nos bancos de dados que não é possível detectar sua origem viral por similaridade de sequências, levando a uma subestimativa da real diversidade viral em uma amostra. Além dessa limitação, alguns vírus, como os integrantes das famílias *Peribunyaviridae* e *Orthomyxoviridae,* possuem genomas segmentados, e vírus altamente divergentes dessas famílias podem não possuir similaridade de sequência de todos segmentos genômicos com a mesma referência viral contida no banco de dados, dificultando a caracterização desses vírus em dados metagenômicos (KRISHNAMURTHY; WANG, 2017).

Uma grande proporção das *reads* e *contigs* montados a partir de bibliotecas de metagenômica não possuem similaridade de sequência com sequências depositadas em bases de dados. Alguns autores se referem a esses *contigs* como “matéria escura” da metagenômica (KRISHNAMURTHY; WANG, 2017). No contexto da metagênomica viral, essa matéria escura pode representar vírus novos ou porções de genomas virais altamente divergentes.

1.8 - Análise de viromas utilizando pequenos RNAs da resposta imune do hospedeiro

A análise de viromas utilizando a técnica RNA-seq, também chamada de metatranscritômica, tem se mostrado uma abordagem eficiente na descoberta de novos vírus e na avaliação da abundância de vírus conhecidos em insetos e outros hospedeiros (DE ALMEIDA et al., 2021; HARVEY; HOLMES, 2022; SHI et al., 2016). Em uma célula hospedeira, todos os vírus produzem moléculas de RNA em algum momento de sua replicação, fazendo o uso dessa técnica propício aos estudos de viromas. O sequenciamento de pequenos RNAs oriundos da resposta imune do hospedeiro permite a montagem de genomas virais (KREUZE et al., 2009; MARQUES et al., 2010). AGUIAR et al., 2015 desenvolveram uma estratégia de análises de viroma utilizando pequenos RNAs virais produzidos pela resposta imune do hospedeiro.

1.8.1 - Origem dos pequenos RNAs

Pequenos RNAs são cadeias curtas de ácidos ribonucleicos, não codificantes que desempenham funções regulatórias e defensivas nos três domínios da vida. A maioria desses RNAs desempenha sua função através de pareamento de bases com

sequências alvos nas células. Tais pareamentos engatilham ciclos de reações envolvendo degradações por RNAses ou modificações estruturais como o bloqueio do acesso da maquinaria de tradução ao mRNA de um gene regulado.

Em eucariotos, as vias mediadas por pequenos RNAs constituem uma ampla rede de sistemas regulatórios conhecida como RNA de interferência (RNAi) (SHABALINA; KOONIN, 2008). Essas vias envolvem a participação de ribonucleases do tipo III que reconhecem e clivam RNAs de fita dupla (dsRNAs) (AGUADO; TENOEVER, 2018). Para desempenhar o papel de reguladores, os pequenos RNAs interagem com proteínas Argonautas que compõe os complexos chamados de RISC (*RNA-induced silencing complex*). Esses complexos reconhecem os RNAs alvos por pareamentos do tipo Watson-Crick e a clivagem ocorre por ação endoribonucleolitca da Argonauta (CARTHEW; SONTHEIMER, 2009). Os pequenos RNAs produtos da clivagem do complexo RISC serão rapidamente degradados pelas vias de degradação de manutenção as células. Os pequenos RNAs gerados tanto pela Dicer quanto pela Argonauta apresentam um grupamento monofosfatos (P) na extremidade 5' e hidroxila (OH) na extremidade 3'. No entanto, os pequenos RNAs gerados durante a fase de iniciação que se associam às proteínas Argonauta podem ser adicionalmente estabilizados por modificações secundárias, como a metilação (CH_3) em 2'O. RNAi pode ser dividido em três vias principais: miRNAs (micro RNAs), siRNAs (*small interference* RNAs) e piRNA (*P-element-induced wimpy testis-interacting* RNAs).

miRNAs são pequenos RNAs endógenos, codificados por seus próprios loci gênicos. Em animais, a via de miRNAs é ativada pelo reconhecimento de estruturas secundárias de RNAs em forma de grampo (~65nt), presentes em longos transcritos de fita simples chamadas de pri-miRNAs. Os pri-miRNAs são processados pela RNase Drosha, clivando os *hairpins* chamados de pre-miRNAs que são exportados para o citoplasma. A enzima Dicer cliva os pre-miRNAs dando origem aos duplexes de miRNAs com cerca de ~20-23 nt (KIM, 2005). Os miRNAs maduros são carregados em uma proteína Argonauta para formar o miRISC. Este complexo irá direcionar regiões complementares nos RNAs alvos. comumente interagindo com 3’ UTR de mRNAs levando à inibição da tradução ou à desestabilização do alvo. O pareamento dos miRNAs não possui complementariedade perfeita com seus alvos.

siRNAs não possuem loci gênicos, são gerados a partir de longos substratos de dsRNA processados progressivamente pela RNAse Dicer gerando pequeno RNAs

de fita duplo faseados de ~20–23 nt (CENIK et al., 2011). O termo faseado (*phased*) refere-se a um padrão específico de pequenas RNAs geradas de forma que tenham sequências complementares com um deslocamento (*offset*) ou fase consistente entre elas. Isso significa que, ao alinhar esses pequenos RNAs, eles terão regiões sobrepostas com uma distância fixa entre os pontos de partida de cada molécula. Os duplexos de siRNA são carregados em proteínas Argonauta especializadas que formam o siRISC que irá catalisar a clivagem endonucleolítica de alvos de RNA (RAND et al., 2005). Diferente dos miRNAs, o pareamento de bases dos siRNAs com seus RNAs alvos é perfeito (SHABALINA; KOONIN, 2008).

Diferentemente dos miRNA e dos siRNAs, a via de piRNAs é ativada por precursores de RNA de fita simples. De forma resumida, longos transcritos não codificantes oriundos de loci genômicos chamados de ”clusters de piRNAs” são clivados dando origem aos piRNAs primários de ~24-30 nt fita simples e faseados. Essa clivagem é realizada pela endoribonuclease mitocondrial, Zucchini/PLD6, que preferencialmente cliva as sequências imediatamente antes de uma base uracila, o que leva a um enriquecimento de 5’ U nos piRNAs primários (MOHN; HANDLER; BRENNECKE, 2015; NISHIMASU et al., 2012). Esses piRNAs primários são carregados nas proteínas PIWI, um subgrupo de proteínas Argonautas em animais (HAN et al., 2015a). Os piRNAs primários também podem desencadear a produção de piRNAs secundárias por meio de um ciclo de auto amplificação dependente de proteínas PIWI, conhecido como mecanismo de pingue-pongue (CZECH; HANNON, 2016). Os piRNAs secundários possuem 10 nt de complementariedade com o piRNAs primário que o gerou e tem por característica a base adenina na décima posição (WANG et al., 2014). O complexo formado por piRNAs e proteínas PIWI, conhecido como piRISC, pode mediar a silenciamento transcricional, bem como a degradação do RNA alvo (CZECH; HANNON, 2016). Os piRNAs são enriquecidos em linhagens celulares germinativas e desempenham um papel crítico na manutenção da estabilidade genômica, principalmente na regulação de elementos transponíveis (DI GIACOMO et al., 2013; ROOVERS et al., 2015) .

Provavelmente, as vias que compões o sistema de RNAi evoluíram a partir de sistemas de defesa antivirais e de regulação de mobilização de transposons, e que ao menos protótipos dessas vias envolvendo as proteínas ancestrais das RNAses do tipo

III Dicer e Argonautas-PIWI, já podiam ser encontradas no LECA (*Last eukayotic common ancestor*) (SHABALINA; KOONIN, 2008).

1.8.2 - Pequenos RNAs virais derivados da resposta imune dos hospedeiros

A clivagem específica de sequências mediada por pequenos RNAs é utilizada tanto em procariotos quanto em eucariotos para combater infecções virais (AGUADO; TENOEVER, 2018). Em eucariotos, a evolução dos sistemas de RNAi como mecanismos antivirais ocorreu, provavelmente, pela adaptação de funções de RNAses do tipo III herdadas das arqueias (SHABALINA; KOONIN, 2008). Essa inovação foi mantida e as vias de RNAi continuam sendo a principal resposta antiviral em plantas, artrópodes e nematoides (CERUTTI; CASAS-MOLLANO, 2006; NAYAK et al., 2013) .

As vias de RNAi podem atuar de forma direta ou indireta durante infecções virais. Na atividade antiviral da via de siRNAs, dsRNAs virais são processados pela enzima Dicer em siRNAs que são carregados na enzima Argonauta guiando a degradação de RNAs virais complementares. A via de siRNA está diretamente envolvida na imunidade antiviral na maioria dos eucariotos, incluindo fungos, animais e plantas; embora em mamíferos, pareça estar restrito a células indiferenciadas (LI et al., 2013; MAILLARD et al., 2013). Alguns vírus possuem estratégias de inibição da via de siRNAs. Um exemplo é a proteína B2 codificada pelo *Flock house vírus* (CHAO et al., 2005). Nessa situação de inibição da via, pequenos RNAs virais ainda são produzidos, porém sem as características típicas de dsRNAs gerados pela Dicer sugerindo a degradação do RNA viral por outros mecanismos de controle (AGUIAR; OLMO; MARQUES, 2016).

Alguns vírus de insetos ativam a via de piRNAs durante a infecção. A função antiviral dessa via ainda é incerta. A análise de pequenos RNAs e infecção viral em mutantes de *D. melanogaster* para diferentes genes da via de piRNA evidenciou que essa via não possui atividade antiviral nessa espécie (PETIT et al., 2016). Porém há grandes diferenças na regulação e componentes da via de piRNA entre o modelo *D. melanogaster* e mosquitos *Aedes* (MIESEN; JOOSTEN; RIJ, 2016). A expressão das proteínas PIWI em Drosófila é majoritariamente restrita as gônadas (LI et al., 2009a), enquanto em *Aedes* algumas proteínas PIWI são expressas em tecidos somáticos (AKBARI et al., 2013). Mosquitos Aedes passaram por uma expansão do repertório

de proteínas PIWI codificadas em seus genomas, com oito proteínas: Piwi 1–7 e Ago3; enquanto Drosophila possui três proteínas principais da via: Piwi, Aubergine e Ago3. Alguns autores discutem que essa expansão de proteínas PIWI pode ter levado ao ganho de função antiviral, explicando a produção de piRNAs virais observada em diferentes infecções em mosquitos (MIESEN; JOOSTEN; RIJ, 2016; VARJAK; LEGGEWIE; SCHNETTLER, 2018).

EVEs estão amplamente distribuídas nos genomas dos artrópodes e comumente geram piRNAs (TER HORST et al., 2019). Especificamente em *A. aegypti* e *A. albopictus*, as EVEs são uma abundante fonte de piRNAs (AGUIAR et al., 2020; PALATINI et al., 2020). Esses piRNAs são majoritariamente complementares à orientação das ORFs virais putativas, sugerindo um papel de direcionamento a alvos de RNA codificantes de proteínas virais cognatas para degradação. A maquinaria de piRNA é capaz de se adaptar a novas integrações genômicas de elementos transponíveis quando nos clusters de piRNA, o que pode gerar a uma adaptação herdável quando esse fenômeno ocorre em células germinativas (KHURANA et al., 2011). Há uma hipótese de que as EVEs poderiam compor uma espécie de “memória imune” de infecções virais que teria a via de piRNAs como efetora da resposta antiviral (JOOSTEN et al., 2021) . A simplicidade de tal modelo de mecanismo antiviral e similaridade com os sistemas CRISPR-Cas9 de procariotos (BARRANGOU et al., 2007) torna essa hipótese atrativa, O estudo mostrando que a remoção da EVE do ISV CFAV aumenta a replicação do vírus cognato em ovários de *A. aegypti* e que a interação EVE-vírus ocorre por meio dos piRNAs é uma evidência na direção dessa hipótese (SUZUKI et al., 2020). Considerando o fato de que o flavivírus CFAV inibe a replicação de arbovírus na mesma espécies (BAIDALIUK et al., 2019), esse potencial mecanismo de imunidade antiviral por EVEs se torna ainda mais relevante como um componente biológico que influência a competência vetorial de mosquitos para patógenos. Uma hipótese alternativa é a de que pequenos RNAs derivados de EVE são um subproduto de sua associação com elementos transponíveis não funcionais e não têm necessariamente uma função (AGUIAR et al., 2020). As integrações de sequências virais podem ser um subproduto de sua associação com elementos transponíveis. Devido a interação com tais elementos, as integrações ocorrem em loci nos quais piRNAs já são gerados para regulação de transposons e as recém integradas EVEs passarão a compor os RNAs transcritos que dão origem a esses

pequenos RNAs. Analisar quais EVEs geram piRNAs ativamente em diferentes populações de mosquitos é fundamental para elucidar se esses elementos genômicos possuem ou não papel antiviral.

A produção de miRNAs virais aparenta ser um fenômeno restrito a alguns vírus de animais. Alguns *Herpesvirus* codificam miRNAs que possuem características de pri-miRNAs eucarióticos que são processados até miRNAs maduros e desempenham importante função no processo de latência desse vírus (RIAZ et al., 2014; UMBACH et al., 2008). Importante ressaltar que apesar da raridade de miRNAs oriundos de loci genômicos virais, miRNAs do hospedeiro são importantes reguladores envolvidos indiretamente na resposta imune antiviral (THAKUR; KUMAR, 2022).

1.8.3 - Pequenos RNAs virais em mamíferos

A história evolutiva dos cordados apresenta um ponto de transição em que as principais defesas antivirais foram das vias de RNAi para um sistema baseado em proteínas, chamado Interferon (AGUADO; TENOEVER, 2018). As defesa antivirais baseadas em RNAi parecem ter sido restringidas a células indiferenciadas com ocorrência mutuamente exclusiva aos sistemas de interferon (MAILLARD et al., 2013, 2016).

Em mamíferos, pequenos RNAs virais podem ser gerados como produto da degradação pela RNAse L (MALATHI et al., 2007). Nessa via, dsRNAs oriundos da replicação viral são reconhecidos por 2’-5’-oligoadenilato sintetases (OAS) que se ligam a esses dsRNAs e catalisam a formação de 2’-5’ oligoadenilatos (2-5A). 2-5A agem como segundo mensageiros, se ligando a RNAse L e induzindo modificações conformacionais que levam essa ribonuclease a sua forma ativa (ZHOU; HASSEL; SILVERMAN, 1993). A RNAse L ativa irá degradar RNAs de fita simples de diferentes tamanhos. As RNAses L também degradam RNAs endógenos, porém a evidência da capacidade seletiva dessas enzimas para moléculas virais (LI; BLACKFORD; HASSEL, 1998). A produção de pequenos RNAs virais pela RNAse L parece ativiar um ciclo positivo mediado pelo reconhecimento desses produtos pela helicase RIG-I (MALATHI et al., 2007). A ativação do RIG-I desencadeia a produção de OAS. Essa via antiviral é um importante componente do sistema imune inato e é relevante na defesa contra diferentes vírus, incluindo SARS-CoV-2 (LI et al., 2021).

Os RNAs virais também são alvos das RNAses que fazem o controle de qualidade e reciclagem dos RNAs celulares. Esse controle é comumente realizado por exoribonucleases como a XRN1 e complexos proteicos como os exossomos (AGUIAR; OLMO; MARQUES, 2016). Os mecanismos de decaimento de RNA desempenham um papel importante na resposta antiviral em mamíferos. RNAs virais comumente não possuem modificações como cap 5’ e cauda poli A, além disso a replicação viral leva a transcritos não convencionais na célula, e.g. transcritos com ORFs pequenas e longos 3’UTR. Tais características fazem dos RNAs virais alvos distinguíveis para as RNAses das vias de decaimento de RNA (GAGLIA; GLAUNSINGER, 2010).

A **Figura 7** compila as vias que processam RNAs virais gerando pequenos RNAs virais em eucariotos previamente descritas. Cada um desses conjuntos de pequenos RNAs virais possui características distinguíveis como: distribuição de tamanho e preferências de nucleotídeo em determinadas posições da sequência. Essas características podem ser usadas para inferir a origem dos pequenos RNA virais quando analisados em larga escala em bibliotecas de RNA-seq.

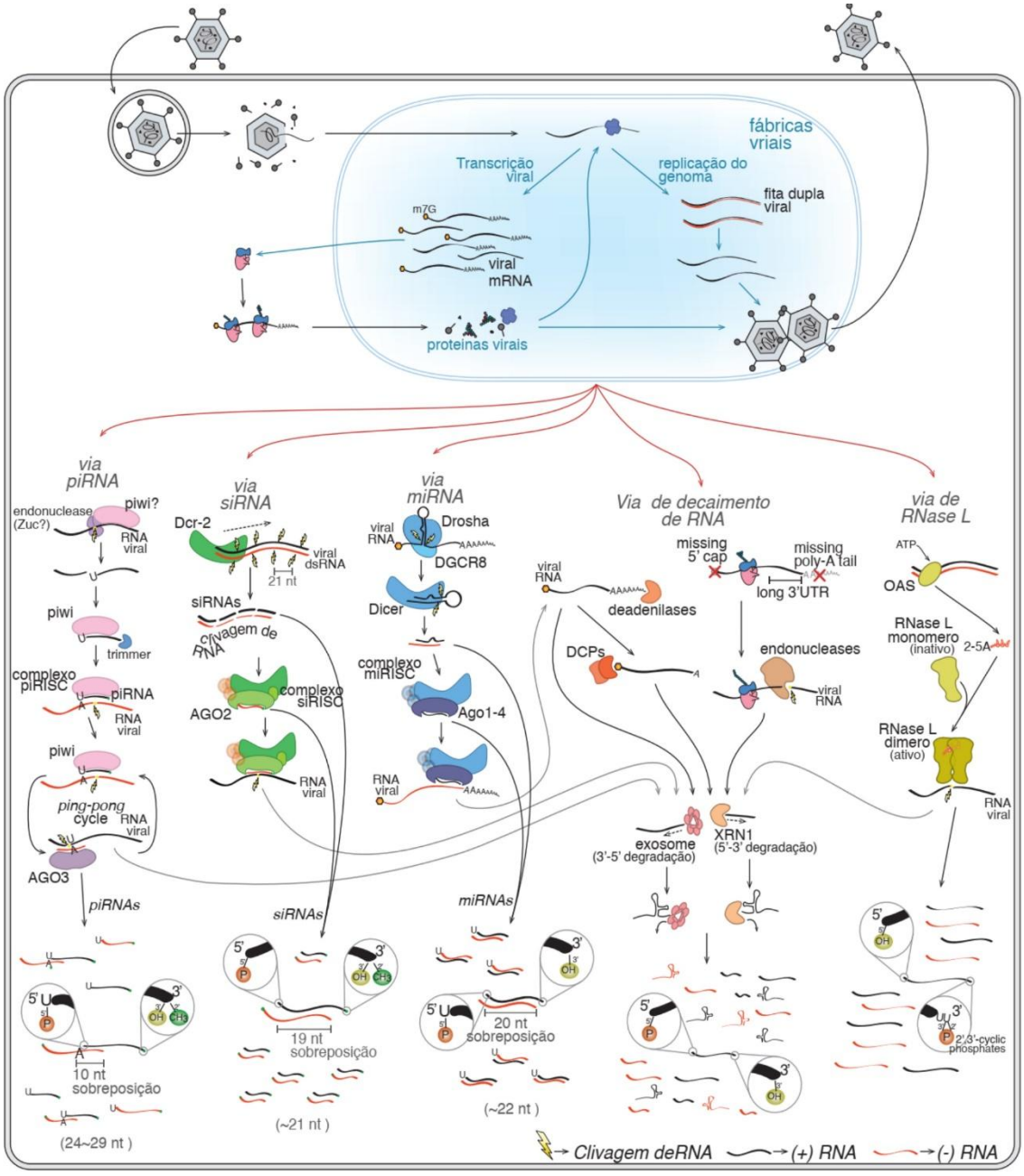


Figura 7 - Geração de pequenos RNAs virais derivados da resposta imune do hospedeiro. Durante a infecção, os genomas e transcritos virais podem ficar expostos e serem reconhecidos por diferentes mecanismos de defase das células hospedeiras. Esses pequenos RNAs possuem características moleculares únicas, como modificações terminais, tamanho, polaridade de fita e preferências de nucleotídeos. Modificada de AGUIAR, OLMO & MARQUES, 2016.

1.8.4 - Abordagem metagenômica utilizando pequenos RNAs da resposta imune do hospedeiro

Enquanto os longos RNAs são produtos diretos da replicação e transcrição viral, a biogênese dos pequenos RNAs envolve processamento adicional de produtos de RNA viral por vias antivirais do hospedeiro, tais como RNA de interferência (RNAi)

(AGUIAR; OLMO; MARQUES, 2016) **(Figura 7)**. Os autores evidenciaram que a fração de pequenos RNAs em mosquitos é naturalmente enriquecida para fragmentos de RNAs virais quando comparada a fração de longos RNAs, favorecendo a reconstituição de genomas virais (AGUIAR et al., 2015). Do ponto de vista metodológico do preparo de bibliotecas para metagenômica viral, tal enriquecimento natural da porção de pequenos RNAs consiste em uma vantagem, pois reduz a necessidade de procedimentos de enriquecimento de partículas virais utilizando etapas de centrifugações e filtragens que elevam as chances de contaminação durante o preparo de bibliotecas para sequenciamento (NACCACHE et al., 2013). Além disso, a porção de pequenos RNAs possui pequena quantidade de RNAs ribossomais, evitando custosos processos de depleção ribossomal e enriquecimento por caudas poli-A que pode influenciar na diversidade viral sequenciada (AGUIAR; OLMO; MARQUES, 2016).

A análise de viromas com pequenos RNAs permiti inferir mecanismos da defesa imune de insetos de acordo com o perfil de distribuição de pequenos RNAs virais. As características moleculares dos pequenos RNAs também podem ser utilizadas na distinção de elementos endógenos virais (EVEs) das sequências virais exógenas, o que tem sido um desafio em estudos de metagenômica viral, as sequências virais exógenas comumente são distinguíveis por um perfil claro de siRNAs (AGUIAR et al., 2015). O perfil de pequenos RNAs associado a cobertura e contagem das *reads* também permite a associação de *contigs* de segmentos genômicos distintos do mesmo vírus (ABBO et al., 2023; AGUIAR et al., 2015; OLMO et al., 2023), o que muitas vezes é um problema utilizando apenas similaridade de sequência (KRISHNAMURTHY; WANG, 2017).

A estratégia de preparo e análise de bibliotecas de pequenos RNAs já foi empregada com sucesso na análise dos viromas e resposta imune antiviral de *A. aegypti* (AGUIAR et al., 2015; OLMO et al., 2023), *A. japonicus* (ABBO et al., 2023), *Culex quinquefasciatus* (LOURENÇO-DE-OLIVEIRA et al., 2018), *Drosophila melanogaster* (AGUIAR et al., 2015), *Lutzomyia longipalpis* (FERREIRA et al., 2018), *Toxorhynchites amboinensis* (DONALD et al., 2018) e *Hevea brasiliensis* (FONSECA et al., 2018), permitindo a caracterização dos viromas e obtenção de informações sobre mecanismos de defesa antiviral desses insetos.

1.9 - Métodos de classificação de sequências virais independentes de alinhamento

1.9.1 - Atributos representativos de sequências virais

Sequências virais altamente divergentes das depositadas em bases de dados podem compor uma parte da “matéria escura” da metagênomica. Para testar essa hipótese, abordagens não baseadas unicamente em similaridade de sequência são necessárias para aumentar o potencial de descoberta de novas sequências virais. Uma alternativa aos métodos dependentes de alinhamento de similaridade de sequência são os métodos “livres de alinhamento” para a classificação de sequências virais (AUSLANDER; GUSSOW; KOONIN, 2021). Tais técnicas exigem a extração de atributos (em inglês, *features*) de sequências que possibilitarão representações vetoriais numéricas das mesmas, permitindo cálculos de distâncias entre sequências em espaços multidimensionais (VINGA, 2014).

Atributos de sequências biológicas baseados em frequências de *k-mers* (*e.g.* dinucleotídeos, trinucleotídeos) constituem um exemplo de representação vetorial amplamente utilizado na bioinformática e em trabalhos com sequências virais (SIMMONDS et al., 2013). Representações vetoriais de “pseudocomposição” semelhantes à de *k-mers*, porém mantendo correlações de posições da sequência também são uma alternativa de representação vetorial (LIU et al., 2015). Também podem ser extraídos atributos relacionados as regiões codificadoras das sequências virais como preferência de códons (BZHALAVA et al., 2018) e densidade de *ORFs* (janelas abertas de leitura, do inglês *Open Reading Frames*) por fita (AMGARTEN et al., 2018).

O uso de pequenos RNAs na análise de viromas proposto por Aguiar et al. 2015, permite a representação das sequências virais baseadas em vetores com os valores das quantidades de pequenos RNAs de diferentes tamanhos. Por exemplo, utilizando os valores de expressão de pequenos RNAs de 15 a 35 nt da fita senso e antisenso, teremos 42 atributos que poderão representar uma sequência viral.

1.9.2 - Aprendizado de máquina

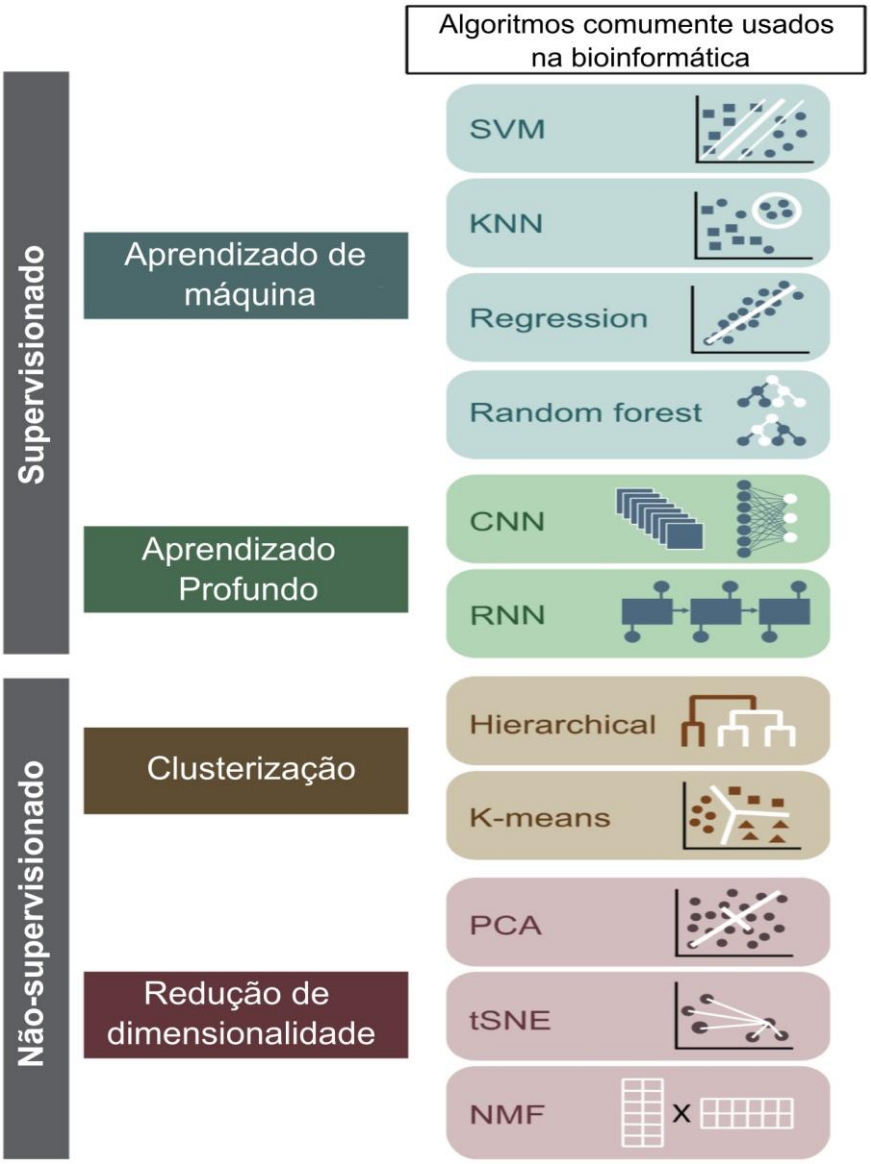


Figura 8 – Algoritmos de Aprendizado de máquina frequentemente usados na Bioinformática. Modificada de AUSLANDER, GUSSOW & KOONIN, 2021.

Muitas das técnicas livres de alinhamento são baseadas em algoritmos de "Aprendizado de máquina" (*Machine Learning)*. O aprendizado de máquina é um campo da ciência da computação que estuda o uso de computadores para simular o aprendizado humano explorando padrões em dados e utilizando funções matemáticas para contínuo aperfeiçoamento de tarefas (AUSLANDER; GUSSOW; KOONIN, 2021). Os algoritmos de aprendizado de máquina podem ser divididos dois paradigmas, supervisionados e não-supervisionados (**Figura 8**). Os algoritmos de aprendizado supervisionados mapeiam a relação entre instâncias (vetores *X*) representadas um conjunto de atributos descritivos e um rótulo ou classe (*y*, *e.g. "*viral" e "não-viral"),

detectando padrões e estabelecendo relações matemáticas que serão utilizadas para classificar novas entradas. O paradigma supervisionado pode ser definido pelas expressões:

Conjunto de treino: pares (X,y), onde $X \in R^d$ e $y \in \{0,1\}$

Objetivo: Obter um modelo/função $f:X \rightarrow y$ para predizer corretamente novas entradas X

Os algoritmos não supervisionados são utilizados para inferir padrões nos dados de entrada sem rotulação prévia (modelos utilizados não são expostos a y) e são amplamente utilizados para visualização de dados aplicando reduções dimensionais (BECHT et al., 2018; MAATEN; HINTON, 2008). Para aplicação de algoritmos supervisionados e não-supervisionados, a definição e extração de atributos que representarão os dados é um passo crítico para o sucesso da classificação, muitas vezes exigindo conhecimento de um especialista sobre os dados que serão utilizados.

2 - Justificativa

Mosquitos são os principais vetores de arbovírus do planeta. Mosquitos do gênero *Aedes* são capazes de transmitir diferentes arbovírus como Dengue, Zika, Chikungunya e Febre Amarela que causam milhões de infecções por ano com profundos impactos socioeconômicos e na saúde pública em regiões tropicais. Fatores como globalização e aquecimento global tem levado a uma expansão da distribuição geográfica desses mosquitos, colocando cada vez mais pessoas sob riscos de contrair arboviroses ao redor planeta. Não há vacinas ou drogas antivirais eficientes e amplamente disponíveis contra a maioria das arboviroses, fazendo do controle dos vetores a principal abordagem no combate dessas doenças. Estudos metagenômicos tem revelado que mosquitos possuem um complexo e diverso viroma, composto por vírus específicos de insetos (ISVs) que podem afetar a competência vetorial para arbovírus. A abordagem metagenômica para a descoberta de vírus possui desafios computacionais como: distinção de vírus circulantes de elementos virais endógenos (EVEs); associação de fragmentos de vírus de genoma segmentados ao mesmo vírus; e identificação de sequências virais altamente divergentes por similaridade de sequência. Além dos ISVs, outros componentes biológicos como as EVEs e bactérias endossimbiontes impactam a fisiologia e a resposta imune dos mosquitos. O desenvolvimento de novas abordagens e ferramentas computacionais que permitam a identificação precisa desses componentes biológicos em larga escala é fundamental para o estabelecimento de novas estratégias de controle biológico e avanços do conhecimento sobre as intrincadas interações dos vírus com mosquitos vetores.

3 - Objetivos

3.1 - Objetivo geral

Caracterizar o viroma de mosquitos vetores utilizando uma abordagem metagenômica baseada em pequenos RNAs da resposta imune do hospedeiro.

3.2 – Objetivos específicos

1- Desenvolver e aplicar estratégias computacionais independentes de similaridade de sequência para determinação de vírus únicos em bibliotecas de pequenos RNAs.
2- Distinguir sequências virais exógenas de EVEs em bibliotecas de pequenos RNAs de forma computacionalmente automatizada;
3- Caracterizar sequências virais altamente divergentes em meio a “matéria escura da metagenômica viral”;
4- Avaliar a detecção de RNAs de *Wolbachia* em bibliotecas de pequenos RNAs de mosquitos e correlacionar a carga de RNAs do endossimbionte com a de pequenos RNAs de vírus circulantes.

4 - Materiais e métodos

4.1 - Análise do viroma de mosquitos vetores utilizando bibliotecas de pequenos RNAs

4.1.1 - Obtenção e sequenciamento das amostras de mosquitos

Mosquitos adultos foram coletados utilizando armadilhas de campo e, em alguns casos, coleta ativa HLC (*human landing catch*), em parceria com colaboradores do consórcio *ZIKAlliance* (https://zikalliance.tghn.org/). Os mosquitos foram previamente identificados por características morfológicas distintivas de cada espécie. RNA total de mosquitos individuais e de grupos de mosquitos foi extraído após maceramento das amostras com *beads* de vidro em solução contendo Trizol (Invitrogen, Carlsbad, USA). O processamento das amostras foi realizado no Brasil no laboratório RNAi (ICB-UFMG) pelo Dr. Yaovi Mathias Todjro e na França pelo Dr. Roenick Olmo (CNRS-Universite´ de Strasbourg). Mais detalhes sobre o processamento das amostras podem ser obtidos nas publicações: (ABBO et al., 2023; OLMO et al., 2023)

Foram implementadas diferentes estratégias para a construção das bibliotecas de pequenos RNAs (**Tabela Suplementar 1)**. A estratégia foi determinada de acordo com a qualidade do RNA avaliada com sistema Bioanalyzer 2100 (Agilent). As bibliotecas foram construídas utilizando RNA total ou pequenos RNAs selecionados por tamanho (~18–30 nt), dependendo da qualidade e rendimento de RNA de cada amostra. No caso de rendimento baixo de RNA, especialmente em amostras de mosquitos individuais, o RNA total foi diretamente para a fase de preparação da biblioteca. Para amostras com mais de 20 µg de RNA, pequenos RNAs foram purificados via PAGE (*Polyacrylamide gel electrophoresis)* em condições desnaturastes e a região do gel contendo bandas correspondentes aos fragmentos com tamanho de aproximadamente 18 a 35nt foram excisados e o RNA foi purificado do gel. Para amostras com mais de 20 µg, porém com perfil de degradação detectado (ausência de picos nítidos de RNA ribossomal), o RNA total foi submetido à um processo de oxidação usando periodato de sódio ($NaIO_4$) (ALEFELDER; PATEL; ECKSTEIN, 1998; MARQUES et al., 2010; YANG et al., 2007) antes da seleção por tamanho. Todas as amostras de RNA que passaram por seleção de tamanho em gel (oxidadas ou não) foram utilizados para a construção de bibliotecas. as bibliotecas foram preparadas utilizando o kit TruSeq Small RNA Library prep (Illumina) ou o kit

NEBNext Multiplex Small RNA Library prep para Illumina (New England BioLabs), seguindo os protocolos recomendados pelos fabricantes, exceto por uma modificação: o adaptador 5' foi substituído por um análogo que contém seis nucleotídeos extras na extremidade 3' para melhorar a precisão da sequência identificadora (*barcoding*) que é sequenciada com o pequenos RNA clonados. As bibliotecas foram sequenciadas na plataforma de sequenciamento GenomEast no Institut de Génétique et de Biologie Moléculaire et Cellulaire em Estrasburgo, França, em modo *single-end* com 50 ciclos nos equipamentos HiSeq2000 ou 4000. O preparo das bibliotecas foi executado pelo Dr. Roenick Olmo (CNRS- Universite´ de Strasbourg).

4.1.2 - Pré-processamento dos dados

A qualidade das bibliotecas sequenciadas foi inspecionada utilizado o programa *Fastqc* (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). As *reads* foram submetidas a filtros de qualidade e retirada de adaptadores utilizando o programa Trim Galore (KRUEGER, 2015). Após a trimagem dos adaptadores, as *reads* contendo qualidade média das bases com valor *phred* (Q) menor que 20, bases ambíguas e menores que 15 nt foram descartadas.

4.1.3 - Alinhamentos das *reads* para filtragem de sequências virais

As *reads* processados foram alinhadas contra uma referência concatenada contendo os genomas de referência dos mosquitos das espécies *A. aegypti* (Refseq: GCA_002204515.1) (MATTHEWS et al., 2018) e *A. albopictus* (Refseq: GCA_006496715.1) (PALATINI et al., 2020) e também contra uma referência contendo genomas bacterianos disponíveis no Genbank utilizando o software Bowtie (v1.3) (LANGMEAD et al., 2009). As *reads* não alinhadas foram filtradas com o software SAMtools (LI et al., 2009b) e utilizadas para os processamentos posteriores.

4.1.4 - Montagem de *contigs*

Os *contigs* foram montados com os softwares SPAdes (BANKEVICH et al., 2012) e Velvet (v. 1.0.13) (ZERBINO; BIRNEY, 2008) combinando montagens de conjunto de *reads* de intervalos de tamanho 20-22nt, 24-30nt e 18-35nt. O programa VelvetOptimser http://bioinformatics.net.au/software.velvetoptimiser.shtml) foi utilizado para estabelecimento de valores de *k-mer* para melhor aproveitamento das montagens combinado ao já fixado *k-mer* de 15 (AGUIAR et al., 2015). Os *contigs* montados com os diferentes intervalos de tamanhos de *reads, k-mers* e montadores

foram combinados usando o programa CAP3 (HUANG; MADAN, 1999). Apenas *contigs* maiores que 200 nt foram utilizados nas análises posteriores.

4.1.5 - Classificação dos *contigs* por similaridade de sequências

Os contigs maiores que 200 nt foram classificados por similaridade de sequência via alinhamentos locais a nível de nucleotídeos com o programa Blast+ (CAMACHO et al., 2009) contra a base de dados *nt* e a nível de aminoácidos com o programa DIAMOND (modo *--very sensitive*) (BUCHFINK; REUTER; DROST, 2021) contra a base de dados *nr.* Ambas as bases de dados foram obtidas do repositório do NCBI (*National Center for Biotechnology Information*) via ftp e mantidas localmente nos servidores do laboratório RNAi. Foram considerados estatisticamente significativos os alinhamentos com *E-value* < 1e-5 a nível de nucleotídeos e *E-value* < 1e-3 a nível de aminoácidos.

4.1.6 - Conferência de domínios proteicos virais

A verificação da presença de domínios proteicos conservados de proteínas virais nas sequências montadas foi realizada utilizando os programas disponíveis em servidores online: Conserved Domain Searcht do NCBI (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) (LU et al., 2020) e pHMMER (https://www.ebi.ac.uk/Tools/hmmer/search/phmmer) (POTTER et al., 2018). A base de dados do Pfam (MISTRY et al., 2021) também foi obtida via ftp (em janeiro de 2021) e mantida localmente em nossos servidores para identificação de domínios conservados c om a função *hmmsearch* do programa HMMER.

4.1.7 - Análises dos perfis de pequenos RNAs virais

Para avaliação da cobertura dos contigs e dos perfis de tamanho dos pequenos RNAs, as *reads* processadas foram mapeadas contra os *contigs* montados e inicialmente classificados como virais por similaridade de sequência utilizando o software Bowtie permitindo um *mismatch* (*-v 1*) e contabilizando apenas o melhor alinhamento de cada read (*-k 1*). O perfil de tamanho dos pequenos RNAs e frequência de nucleotídeos da primeira base da extremidade 5’ das *reads* foram calculadas considerando *reads* de 15-35 nt mapeadas nos *contigs* montados avaliando cada fita (senso e antisenso) separadamente. O processamento dos alinhamentos; quantificações das reads separadas por tamanho e polaridade de fita; e frequência de nucleotídeos nas extremidades 5’ foram realizados utilizando *scripts* escritos nas

linguagens *Perl* e *R* e funções da ferramenta *SAMtools*. Os gráficos de perfis de pequenos RNAs e de cobertura dos contigs virais foram gerados com scripts utilizando o pacote *ggplot2* (WICKHAM, 2011) na linguagem R.

4.1.8 - Curadoria dos *contigs* classificados como virais por similaridade de sequência

A curadoria manual dos *contigs* virais para a distinção de vírus exógenos de EVEs consistiu na inspeção dos cinco melhores resultados do programa *BLAST* (*blastn* ou *blastx*) considerando o parâmetro *max score*; inspeção da estrutura de *ORFs* com verificação da continuidade e extensão geral em todo o *contig*; avaliação do perfil de pequenos RNAs analisando picos simétricos de 20-22 nt sem preferência de base como evidências de siRNAs e picos assimétricos de 24-29 nt com preferência 5'U como evidência de piRNAs. Também foi avaliada a cobertura por pequenos RNAs de cada *contig*. *Contigs* com *ORFs* não interrompidas, perfil claro de siRNA e cobertura coerente de pequenos RNAs foram curados como virais com alto grau de confiança. *Contigs* contendo *ORFs* truncados e ausência de picos simétricos de 21 nt foram considerados EVEs putativas. Esses *contigs* confirmados como virais foram agrupados por similaridade de sequência utilizando o programa *CDHIT* (FU et al., 2012) com os cortes de 90% de identidade do alinhamento global (*-c*) e 90% de cobertura do alinhamento para a menor sequência *(-aS*) para remoção da redundância.

4.1.9 - Coocorrência dos contigs virais para determinação de vírus únicos utilizando a carga de pequenos RNAs

Para análise de coocorrência e determinar o número de vírus únicos, os contigs virais não redundantes foram usados para quantificação da abundância de pequenos RNAs em cada uma das bibliotecas. Os agrupamentos hierárquicos das bibliotecas e dos contigs foi gerado com base nos valores de RPKM *(Reads Per Kilobase of transcript per Million*) de pequenos RNAs de 20 a 22 nt alinhados aos contigs em cada uma das bibliotecas separadamente. Os dendrogramas e heatmap foram gerados com o pacote *ComplexHeatmap* (GU; EILS; SCHLESNER, 2016). Todas as combinações possíveis de métodos de agrupamento (*hclust*) e distâncias (*dist*) disponíveis foram testadas sistematicamente e visualmente inspecionadas para escolha da combinação que melhor permitiu inferir os grupos de contigs pertencentes ao mesmo vírus combinando as evidências do resultado de similaridade de sequência.

A robustez estatística dos agrupamentos de contigs gerados foi acessada pelo cálculo da probabilidade dos agrupamentos usando o método de "*Multiscale bootstrap"* com o pacote *pvclust* (SUZUKI; SHIMODAIRA, 2006) na linguagem R avaliando 1000 pseudorepetições. Para avaliar a influência da escolha do intervalo de pequenos RNAs quantificados na estruturação dos agrupamentos de contigs virais foram realizados cálculos de dissimilaridade com o pacote *seriation* (HAHSLER; HORNIK; BUCHTA, 2008) na linguagem R. A versão online interativa do heatmap de coocorrência com os agrupamentos hierárquicos de bibliotecas e contigs foi criada com o pacote *InteractiveComplexHeatmap* (GU; HÜBSCHMANN, 2022) e a disponibilizada online com funções do pacote s*hiny* (CHANG et al., 2023) para construção de aplicativos em *R*. O heatmap interativo foi hospedado utilizando o limite gratuito do servidor *shinyapps.io*.

4.1.10 - Extensão da montagem de contigs virais

Para tentar estender a montagem das sequências virais, *contigs* pertencentes ao mesmo agrupamento hierárquico com similaridade de sequência a vírus próximos foram submetidos a uma nova rodada de montagem com o programa SPAdes usando os como entrada para o parâmetro *–trusted-contigs* na remontagem dos reads das bibliotecas nas quais esses contigs virais foram inicialmente montados.

4.1.11 - Referência de arbovírus comumente transmitidos por mosquitos

Para verificar a presença dos principais arbovírus transmitidos por mosquitos que ocasionalmente poderiam estar presentes com baixa carga viral impossibilitando a montagem *de novo* de contigs foi criada uma referência com as sequências representativas obtidas do Refseq NCBI para os vírus: Dengue vírus (NC_001477.1, NC_001474.2, NC_001475.2, NC_002640.1, NC_075403.1, NC_075435.1); Zika vírus (NC_035889.1, NC_075414.1, NC_075423.1, NC_075429.1); Chikungunya vírus (NC_004162.2, NC_075020.1); Yellow fever vírus (NC_002031.1); Usutu vírus (NC_006551.1); West Nile vírus (NC_001563.2, NC_009942.1).

4.1.12 - Comparação dos perfis de pequenos RNAs virais

Para a análise comparativa dos perfis de pequenos RNAs dos vírus encontrados, foi selecionado o maior contig montado com similaridade de sequência com o gene ou sequência de aa da RdRP do vírus mais próximo ou apenas o maior contig com similaridade de sequência com a sequência viral mais próxima encontrada

no Genbank quando fragmentos da replicasse do potencial vírus não foram montados. Para os vírus segmentados, o mesmo critério foi aplicado para os contigs dos segmentos virais que não codificam replicases. A biblioteca na qual o contig representativo foi montado foi alinhada contra o contig para estabelecimento o perfil de pequenos RNAs representativo. A quantidade de pequenos RNAs de diferentes tamanhos alinhada a cada contig foi normalizada por Z-score (AGUIAR et al., 2015), estabelecendo um vetor numérico com atributos quantitativos para cada contig viral. A similaridade entre os perfis de pequenos RNAs dos *contigs* virais representativos foi analisanda via análise de aprendizado de máquina não supervisionado com agrupamentos hierárquico utilizando o pacote *ComplexHeatmap* (GU; EILS; SCHLESNER, 2016) na linguagem R. Todas as combinações possíveis de métodos de agrupamento (*hclust*) e distâncias (*dist*) disponíveis foram testadas sistematicamente e visualmente inspecionadas para escolha da combinação que apresentou maior coerência com o conhecimento prévio sobre a segmentação genômica dos vírus encontrados.

4.1.13 - Análises dos piRNAs virais

Sobreposições e padrões dos pequenos RNAs virais foram analisadas para determinar a presença de piRNAs secundários. As distâncias entre os pequenos RNAs foram calculadas conforme descrito anteriormente (GAINETDINOV et al., 2018; HAN et al., 2015b). Resumidamente, a frequência das distâncias 5'-5' entre das reads de 24 a 29 nt de alinhadas em fitas opostas foi calculada e normalizada por Z-score. O enriquecimento das bases A-U na primeiras e décima posição das reads de 24 a 30 nt foi alinhadas aos contigs virais foi analisado via inspeção dos gráficos de logos de sequências gerados com o pacote *ggseqlogo* na linguagem R (WAGIH, 2017).

4.1.14 - Análises filogenéticas

Para os potenciais novos vírus descobertos nas espécies *A. aegypti*, *A. albopictus* e *Ae. japonicus*, selecionamos os contigs contendo a maior sequência de RdRP para cada vírus identificado. Para os bunyavírus de *Ae. japonicus*, também selecionamos os contigs que continham as maiores sequências de nucleocapsídeo e glicoproteína. A presença de domínios proteicos conservados de RdRp, nucleocapsídeo e glicoproteína foi confirmada com a ferramenta web Conserved Domain Search – NCBI. Buscas por sequências virais homólogas para compor o conjunto de sequências para inferências filogenéticas foram realizadas usando os

programas Blastx e Blastp versões web. Os alinhamentos múltiplo globais foram realizados com o programa MAFFT (KATOH et al., 2002). Os modelos evolutivos mais adequado foram selecionados usando o programa MEGA-X (KUMAR et al., 2018) sob o critério de Informação de Akaike. A inferência filogenética foi executada com o programa MEGA-X usando o método de Máxima Verossimilhança. Para todas as árvores filogenéticas, a robustez dos clados foi avaliada usando o método de *bootstrap* (1000 pseudorréplicas). As árvores foram visualizadas e editadas usando o programa iTOL versão 6.5 (LETUNIC; BORK, 2021).

4.1.15 - Alinhamento de Sequências e Modelagem de Estrutura de RNA

As sequências de RNA dos Narnavírus foram alinhadas usando o programa *MUSCLE* versão 3.8.1551 (EDGAR, 2004). As estruturas secundárias de RNA foram preditas usando o servidor web *RNAstructure* versão 6.3 com temperatura ajustada para 28°C (REUTER; MATHEWS, 2010). Pseudo-nós no RNA do Totivírus montado em bibliotecas de *Ae, japonicus* foram preditos com o programa *DotKnot* versão 1.3.2 (SPERSCHNEIDER; DATTA, 2010). As estruturas de RNA foram visualizadas usando o programa *VARNA* (DARTY; DENISE; PONTY, 2009).

4.2 - Análises das estruturas proteicas de *Ae. japonicus Narnavirus* 1 segmento 2

As propriedades físico-químicas da estrutura primária foram calculadas com o programa *ProtParam* (WILKINS et al., 1999). Alinhamentos globais foram realizados com o programa *Clustal Omega* (SIEVERS et al., 2011) e editados com o servidor web *ESPcript* (ROBERT; GOUET, 2014). O programa *PSIPRED Workbench* foi utilizado para avaliação das estruturas secundárias preditas com o programa *PSIPRED* 4.0 (BUCHAN et al., 2013; BUCHAN; JONES, 2019). O programa *DISOPRED*3 (JONES; COZZETTO, 2015) foi utilizado para avaliação de regiões desordenadas e o MEMSAT-SVM (NUGENT; JONES, 2009) para avaliação de regiões de interações com membranas. A modelagem de estruturas terciárias foi feita com o programa *AlphaFold* (JUMPER et al., 2021) rodando localmente em servidores instanciados na plataforma AWS. Utilizando a base de dados de MSA completa e com parâmetros padrão. A base de dados utilizada foi obtida em "06/01/2022".

4.3 - Análise do EVEroma

Para a análise do compêndio de EVEs das espécies de mosquitos estudadas nesse trabalho, aqui denominado EVEroma, todas as 122 bibliotecas geradas por nosso grupo passaram por uma segunda execução do *pipeline* do viroma porém, sem excluir as reads alinhadas nos genomas de mosquitos para permitir a montagem de contigs das EVEs presentes nos genomas disponíveis. Foi estabelecido como corte que as porções dos contigs com alinhamentos contra sequências virais tivessem no mínimo 100nt ou 33aa. Para concentrar a análise apenas nas sequências com potencial similaridade com vírus cognatos circulantes, as porções com alinhamentos não virais ou sem alinhamentos significativos dos contigs foram removidas. O perfil de pequenos RNAs dos contigs maiores que 200nt e com similaridade de sequência com sequências virais a nível de nt e aa foram inspecionadas visualmente. Sequências virais que apresentaram perfil de piRNAs (24-30nt) sem evidências de produção de siRNAs (dsRNAs de 21nt) foram considerados potenciais EVEs. A redundância das sequências de EVEs foi removida com o programa CDHIT com os cortes de 80% de identidade do alinhamento global (*-c*) e 90% de cobertura do alinhamento para a menor sequência *(-aS*) para remoção da redundância. As sequências de EVEs das espécies *A. aegypti* e *A. albopictus* foram alinhadas contra os respectivos genomas de referência GCA_002204515.1 e GCA_006496715.1 utilizando o programa blastn+.

Para análise de coocorrência das EVEs nas diferentes espécies de mosquitos, os contigs não redundantes foram usados para quantificação da abundância de pequenos RNAs em cada uma das bibliotecas. Os agrupamentos hierárquicos das bibliotecas e dos contigs foi gerado com base nos valores de RPKM *(Reads Per Kilobase of transcript per Million*) de pequenos RNAs de 24 a 30nt (piRNAs) alinhados aos contigs em cada uma das bibliotecas separadamente. Os dendrogramas e heatmap foram gerados com o pacote *ComplexHeatmap* (GU; EILS; SCHLESNER, 2016).

4.4 - Diferenciação de contigs virais exógenos e EVEs utilizando aprendizado de máquina não-supervisionada e supervisionado

Sequências montadas a partir das 122 bibliotecas de pequenos RNAs preparadas por nosso grupo foram inicialmente classificadas por similaridade de sequência em “virais”, “não-virais” e “desconhecidas”. Após curadoria dos perfis de pequenos RNAs (**Figura 9A**) as sequências virais foram divididas em 2 classes: “Virais” ou “EVEs”. Sequências dessas duas classes passaram por uma estruturação

de dados que resultou em vetores representativos compostos por 48 atributos gerados a partir da quantificação de pequenos RNAs para cada sequência (**Figura 9B**). Os 48 atributos de pequenos RNAs usados para representar as sequências estão descritos na **Tabela 1**.

Com o objetivo de reduzir a dimensionalidade para a visualização eficaz dos dados e avaliar a capacidade distintiva entre as duas classes por meio da combinação de atributos representativos selecionados, empregamos técnicas de aprendizado de máquina não supervisionado. Utilizamos os métodos de PCA (*Principal Component Analysis*) e t-SNE (*t-distributed stochastic neighbor embedding*) para realizar essa análise.

Para construção de um modelo classificador capaz de distinguir sequências virais de EVEs automaticamente utilizando bibliotecas de pequenos RNAs, treinamos classificadores utilizando métodos de aprendizado supervisionado com o conjunto de dados de sequências virais curadas dívidas em “Virais” e “EVEs”. As etapas a seguir foram realizadas com scripts escritos na linguagem Python implementando funções e pacotes com a biblioteca *scikit-learn.*

Tabela 1 – Descrição dos 48 atributos de pequenos RNAs utilizados para representar as sequências virais.

Descrição dos atributos	Nro. de atributos	Referência
Número de pequenos RNAs separados por tamanho de 15-35nt , por polaridade de fita e normalizados por Z-score	42	AGUIAR et al., 2015
Número de pequenos RNAs de 15-18nt normalizado por tamanho do contig	1	AGUIAR et al., 2016
Número de pequenos RNAs de 20-22nt (siRNAs) normalizado por tamanho do contig	1	AGUIAR et al., 2016
Número de pequenos RNAs de 24-30nt (piRNAs) normalizado por tamanho do contig	1	AGUIAR et al., 2016
Razão entre siRNA e piRNAs	1	AGUIAR et al., 2016
Razão do número de pequenos RNAs 20-22nt entre fita senso e antisenso	1	AGUIAR et al., 2016
Numero de pequenos RNAs de 18-35nt normalizados pelo tamanho do contig.	1	AGUIAR et al., 2016
Total	**48**	

Foram treinados e avaliados modelos utilizando 10 algoritmos de aprendizado supervisionado: *Naive Bayes*, *Decision Tree*, *Support Vector Machines* (SVM), *k-Nearest Neighbors* (KNN); e os modelos do tipo *ensemble*: *Random Forest*, *Adaptative Boosting*, *Gradient Tree Boosting, CatBoost*, *XGBoost* e *LightGBM*. Os dados foram estratificados em 70% das instâncias para treino e 30% para teste. Os experimentos *in sílico* para cada algoritmo foram realizados com validação cruzada com *k-fold* = 5. O refinamento dos hiper parâmetros dos diferentes modelos construídos com os diferentes algoritmos foi realizado automaticamente com o método *GridSearchCV*(). Os hiper parâmetros foram otimizados em função da acurácia. Para comparação do desempenho dos melhores modelos de cada algoritmo foram avaliadas as métricas: acurácia, precisão, revocação e f1-score. As médias das métricas precisão, revocação e f1-score entre classes foram calculadas de forma macro e balanceada. Também foram avaliadas as curvas ROC (*Receive Operating Characteristic*) acompanhadas pelos valores de AUC (*Area Under the Curve*) e as curvas DET (*Detection Error Trade-Off*) dos melhores modelos de cada algoritmo após refinamento dos hiper parâmetros.

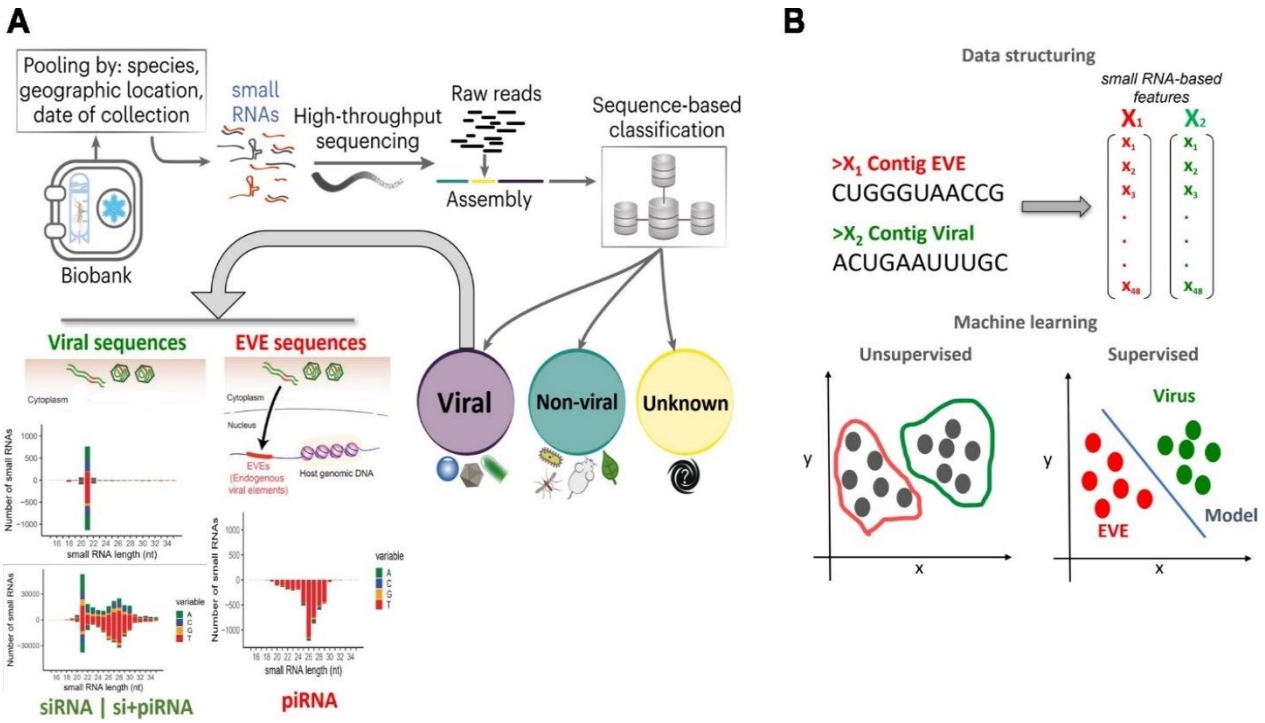


Figura 9 – **Estratégia de diferenciação de sequências virais de EVEs**. **A.** Após processamento dos reads, montagem de contigs e classificação por similaridade de sequência as reads foram realinhadas aos contigs incialmente classificados como virais para geração dos perfis de pequenos RNAs para cada contig. Os perfis e as sequências foram inspecionados e classificados como "Virais" quando apresentaram perfil de siRNA (dsRNAs simétricos de 21nt) e janelas de leitura não-interrompidas por códons de paradas ou "EVEs", quando apresentavam perfil de piRNAs (~24-30 nt sem simetria de fita, comumente enriquecidos com U na 5'). **B**. Os contigs virais passaram por uma etapa de estruturação de dados em que cada sequência passou a ser representada por um vetor contendo 48 atributos numéricos obtidos a partir de diferentes quantificações de pequenos RNAs alinhados aos mesmos. Com os dados estruturados, foram aplicados algoritmos de aprendizado de máquina não-supervisionados para visualização dos dados e foram treinados modelos classificadores para distinguir as duas classes

4.5 - Automatização e criação de um container do pipeline de análise de viroma com pequenos RNAs

Com o objetivo de oferecer uma solução viável, portátil e compacta para análises de viromas utilizando bibliotecas de pequenos RNAs, todos os passos mencionados nas seções anteriores (**Figura 10**) que envolvem: controle de qualidade das bibliotecas; montagem de contigs; classificação dos contigs por similaridade de sequência; quantificações de pequenos RNAs; geração de gráficos de perfis de pequenos RNAs e cobertura dos contigs virais; e diferenciação de sequências de vírus de EVEs utilizando atributos de pequenos RNAs foram automatizados e encapsulados em um único contêiner que pode ser utilizado por linha de comando em sistemas Unix.

Uma imagem Docker foi criada para a construção de um contêiner preparado para executar o pipeline de forma transparente ao usuário. O desenvolvimento e testes do encapsulamento da aplicação nesta imagem linux foram realizados em servidores linux (Ubuntu 22.04 e CentOS). Os testes de execução de contêineres foram feitos com uso das runtimes Docker e Podman em quatro máquinas, com diferentes configurações de hardware, havendo sucesso em todas as combinações.

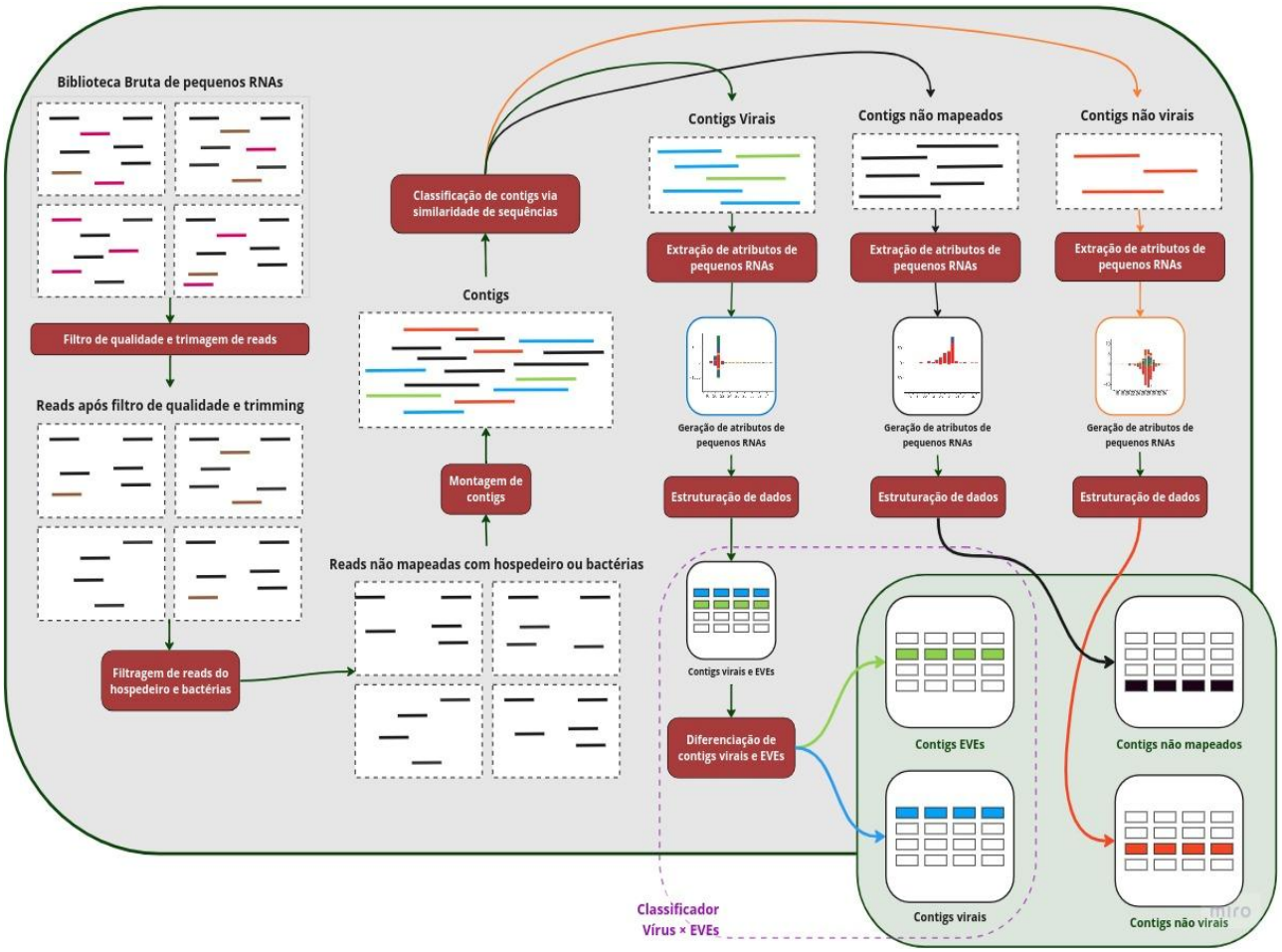


Figura 10 – Etapas do pipeline de metagenômica baseada em pequenos RNAs. As etapas mostradas foram automatizadas e disponibilizadas em um container de fácil instalação para uso da ferramenta computacional

A imagem base utilizada, a partir da qual se construiu a nova imagem customizada, é a imagem Docker oficial da linguagem Perl, versão habilitada para funcionalidades *multithreaded* (LEINO; MÜLLER, 2009). Essa imagem oferece um ambiente, baseado na distribuição Linux Debian, pronto e extensível para execução de scripts em linguagem Perl e inclui também o gerenciador de pacotes Perl cpanm e a linguagem Python (https://metacpan.org/dist/App-cpanminus/view/bin/cpanm). Usando o paradigma Docker Multi-stage build, a inclusão das dependências necessárias do pipeline foram segmentadas num único arquivo Dockerfile, agrupando blocos de declarações correlacionadas em 6 estágios que permitem a divisão lógica de cada bloco de dependências:

1. **Stage perl**: Estágio inicial no qual a instrução FROM aponta para a imagem base. Utilizando o gerenciador cpanm, todas as dependências Perl são instaladas nesse estágio (“cpanm - get, unpack build and install modules from CPAN - metacpan.org”, [s.d.]);

2. **Stage R:** Estágio que acrescenta a instalação da linguagem R e todas as bibliotecas R usadas no pipeline;
3. **Stage Dependencies:** Adiciona a instalação de todos os softwares livres de bioinformática utilizados no processamento de reads e contigs ao longo das etapas do pipeline;
4. **Stage Main:** Inclui os diretórios padrão utilizados na execução do pipeline;
5. **Stage Final:** Estágio final onde são feitos procedimentos de limpeza com o objetivo de reduzir ao máximo o tamanho da imagem final resultante. A limpeza envolve a remoção de arquivos internos de interesse específico dos procedimentos de instalação de dependências nos estágios anteriores.;
6. **Stage Dev:** Estágio extra opcional que substitui o Stage Final realizando uma configuração diferente, especialmente para facilitar o uso da ferramenta em ambiente de desenvolvimento. Nesse caso não são realizados os procedimentos de limpeza e recursos adicionais são inseridos, como por exemplo editores de texto;

Dentre os recursos encapsulados neste container está incluso um modelo do tipo ensemble *Random Forest* separador de contigs virais exógenos e EVEs resultante de treinamento realizado num segundo container Docker gerado a partir da imagem pública docker.io/jupyter/datascience-notebook. O modelo treinado foi exportado num arquivo com extensão "*.joblib"* utilizado internamente pelo pipeline na etapa de refinamento da classificação de contigs virais desdobrando-as entre contigs virais exógenos e EVEs.

4.6 - Quantificação de RNA de *Wolbachia* em bibliotecas de pequenos RNAs de mosquitos

As reads trimadas foram previamente alinhadas aos genomas de *A. aegypti* e *A. albopictus* com o programa bowtie permitindo 1 missmatch (*-v* 1). As reads não alinhadas nos genomas de mosquitos foram utilizadas para quantificar os pequenos RNAs de *Wolbachia*s. Foram baixados todos os 67 genomas completos de *Wolbachia*s disponíveis no Refseq do Genbank (abril de 2023). Os genomas de *Wolbachia* foram indexados em uma referência única e os reads das bibliotecas foram alinhados a essa referência com o programa bowtie sem permitir missmatchs (parâmetros: *-a* e *-v* 0). As reads alinhadas foram quantificadas e normalizadas por RPM (*reads per million*). Os valores normalizados foram utilizados para execução de agrupamentos hierárquicos com o pacote *ComplexHeatmap* na linguagem R que permitiu comparar a quantificação e especificidade da detecção de linhagens de *Wolbachia* nas diferentes bibliotecas.

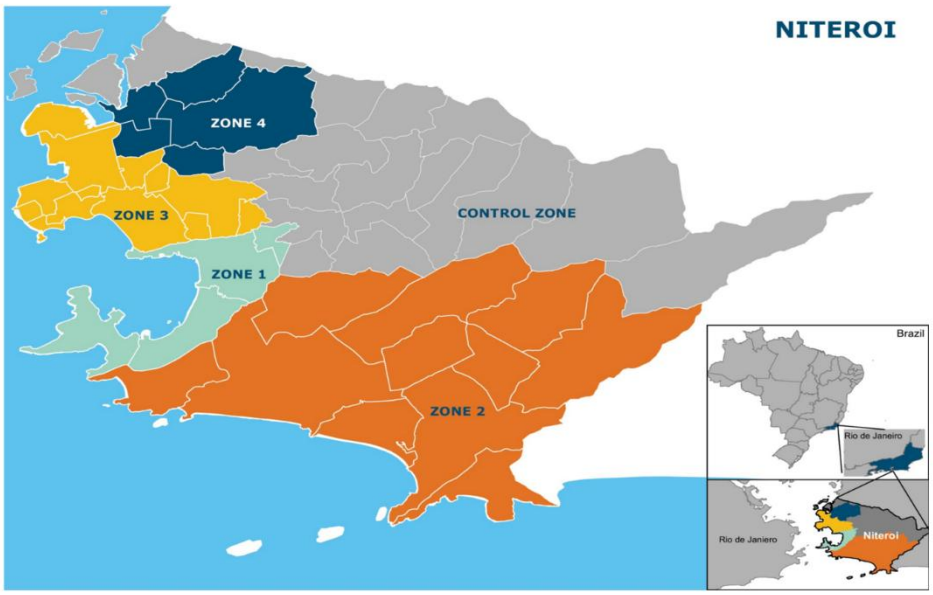
Para quantificações especificas de pequenos RNAs da linhagem *wAlb* (Genbank: NZ_CP031221.1) (SINHA et al., 2019) e da *wMel* infectando *A. aegypti* artificialmente nos estudos de campo (GV_2018_1, Genbank: CP072672.1) (DAINTY et al., 2021), foram indexados os dois genomas de referência representativos e as reads foram alinhados sem permitir missmatches e contabilizando apenas os melhores alinhamentos para cada read com o software bowtie (parâmetros: *-v* 0 *-k* 1). Para quantificação de reads alinhados aos genes de *wAlb* foi utilizado o arquivo GFF de anotações do genoma NZ_CP031221.1 e as contagens foram computadas com o programa HTSeq (modo *intersection_strict*) (PUTRI et al., 2022).

4.6.1 - Análise de bibliotecas de pequenos RNAs de experimento controle para a detecção de pequenos RNAs de *Wolbachia*

Para verificar a sensibilidade e precisão da detecção de reads de *Wolbachia*s em bibliotecas de pequenos RNAs, obtivemos 18 bibliotecas públicas de um experimento em que nove pools de *A. aegypti* artificialmente infectados pela *Wolbachia wAlb* foram sequenciados concomitantemente com nove pools de *A. aegypti* tratados com o antibiótico tetraciclina para a remoção da infecção pelo endossimbionte (BISHOP et al., 2022). As bibliotecas foram processadas seguindo os passos já descritos anteriormente e o pipeline para detecção de vírus também foi utilizado para verificar a presença de infecções virais não descrita pelos autores do trabalho que gerou as bibliotecas.

4.6.2 - Análise do impacto da infecção artificial por *wMel* nos ISVs infectando *A. aegypti* em campo

A



B

Id_biblioteca	Zona	Periodo_coleta	Nro. de Femeas	spp.
Aaeg_1	Zona3	Jan_2018	1	*A. aegypti*
Aaeg_2	Zona3	Jan_2018	2	*A. aegypti*
Aaeg_3	Zona3	Jan_2019	5	*A. aegypti*
Aaeg_4	Zona3	Jan_2019	6	*A. aegypti*
Aaeg_5	Zona3	Jan_2020	5	*A. aegypti*
Aaeg_6	Zona3	Jan_2020	1	*A. aegypti*
Aaeg_7	Zona4	Dez_2020-Jan_2021	17	*A. aegypti*
Aaeg_8	Zona1	Dez_2020-Jan_2021	16	*A. aegypti*
Aaeg_9	Zona3	Dez_2020-Jan_2021	14	*A. aegypti*
Aaeg_10	Zona2	Dez_2020-Jan_2021	6	*A. aegypti*
Aaeg_11	Zona5	Dez_2020-Jan_2021	14	*A. aegypti*
Aaeg_12	Zona4	Abril_2021-Maio_2021	21	*A. aegypti*
Aaeg_13	Zona1	Abril_2021-Maio_2021	17	*A. aegypti*
Aaeg_14	Zona3	Abril_2021-Maio_2021	19	*A. aegypti*
Aaeg_15	Zona2	Abril_2021-Maio_2021	12	*A. aegypti*
Aaeg_16	Zona5	Abril_2021-Maio_2021	20	*A. aegypti*
Aalb_18	Zona1	Abril_2021-Maio_2021	28	*A. albopictus*
Aalb_20	Zona2	Abril_2021-Maio_2021	21	*A. albopictus*
Aalb_21	Zona5	Abril_2021-Maio_2021	30	*A. albopictus*

Figura 11 – Amostragem de mosquitos coletados na cidade de Niterói-RJ em áreas de avaliação do projeto *Wolbachia* para controle da transmissão de arbovírus. A. Mapa da cidade de Nitéroi-RJ dividida em quatro Zonas onde ocorrem solturas de mosquitos *A. aegypti* artificialmente infectados por *Wolbachia* e uma zona controle (Zona 5, cinza) onde não houve solturas de mosquitos com *Wolbachia* (Mapa obtido da publicação PINTO et al., 2021). **B.** Informações das amostras de mosquitos que deram origem as 19 bibliotecas de pequenos RNAs analisadas.

Em colaboração com o grupo do Dr. Luciano Moreira (Fiocruz-MG) obtivemos amostras e sequenciamos bibliotecas de pequenos RNAs de *A. aegypti* da região de Niterói – RJ em que o projeto de controle de transmissão de arbovírus por mosquitos infectados artificialmente com *Wolbachia*s se encontra em andamento. A cidade foi dividida em cinco zonas (**Figura 11A**) (PINTO et al., 2021), nas quais quatro foram soltos mosquitos infectados com *wMel* (DAINTY et al., 2021) e uma região controle onde não foram soltos mosquitos. A soltura e coleta de mosquitos foi realizada pelo grupo do Dr. Luciano Moreira. As coletas ocorreram entre janeiro de 2018 e maio de 2021. O processamento dos mosquitos, extração de RNA e controlo de qualidade das amostras foram realizados pelo Dr. Yaovi Mathias Todjro e doutoranda Ellen Caroline no laboratório RNAi UFMG. Foram preparadas e sequenciados 16 amostras de *A. aegypti* e *três* de *A. albopictus* (**Figura 11B**). As bibliotecas foram preparadas pelo Dr. Roenick Olmo na Universidade de Estrasburgo – FR seguindo os protocolos descritos nas seções anteriores, porém, sem procedimentos de oxidação de RNA. As bibliotecas foram enviadas para sequenciamento na plataforma GenomEast, França com o equipamento Illumina Next-seq 2000. A bibliotecas sequenciadas passaram por todos os passos de controle de qualidade, análise de viroma e quantificação de pequenos RNAs de *Wolbachia*s descritos nas seções anteriores.

5 - Resultados

5.1 - Análise do viroma global de mosquitos vetores

Como parte do consórcio ZikaAlliance (https://zikalliance.tghn.org/), nosso grupo de pesquisa concluiu o sequenciamento de 107 bibliotecas de pequenos RNAs preparadas a partir de mosquitos selvagens pertencentes a 10 espécies (*A. aegypti*, *A. albopictus*, *A. japonicus*, *Haemagogus sp*., *Sabethes sp*., *A. furcifer*, *A. vexans*, *A. vittatus*, *A. luteocephalus*, *A. taylori*) coletadas em 6 países (Brasil, França, Suriname, Singapura, Gabão e Senegal) (**Figura 12A; Tabela Suplementar 1**). Foram também sequenciadas 11 bibliotecas de mosquitos *A. aegypti* de linhagens de laboratório e duas bibliotecas de cultura de células Aag2. Adicionalmente, duas bibliotecas públicas de *A. japonicus* da Holanda (ABBO et al., 2020) que não tiveram o viroma previamente analisado também foram incluídas ao conjunto de bibliotecas analisadas, totalizando 122 bibliotecas de pequenos RNAs nesse estudo (**Figura 12B; Tabela Suplementar 1**).

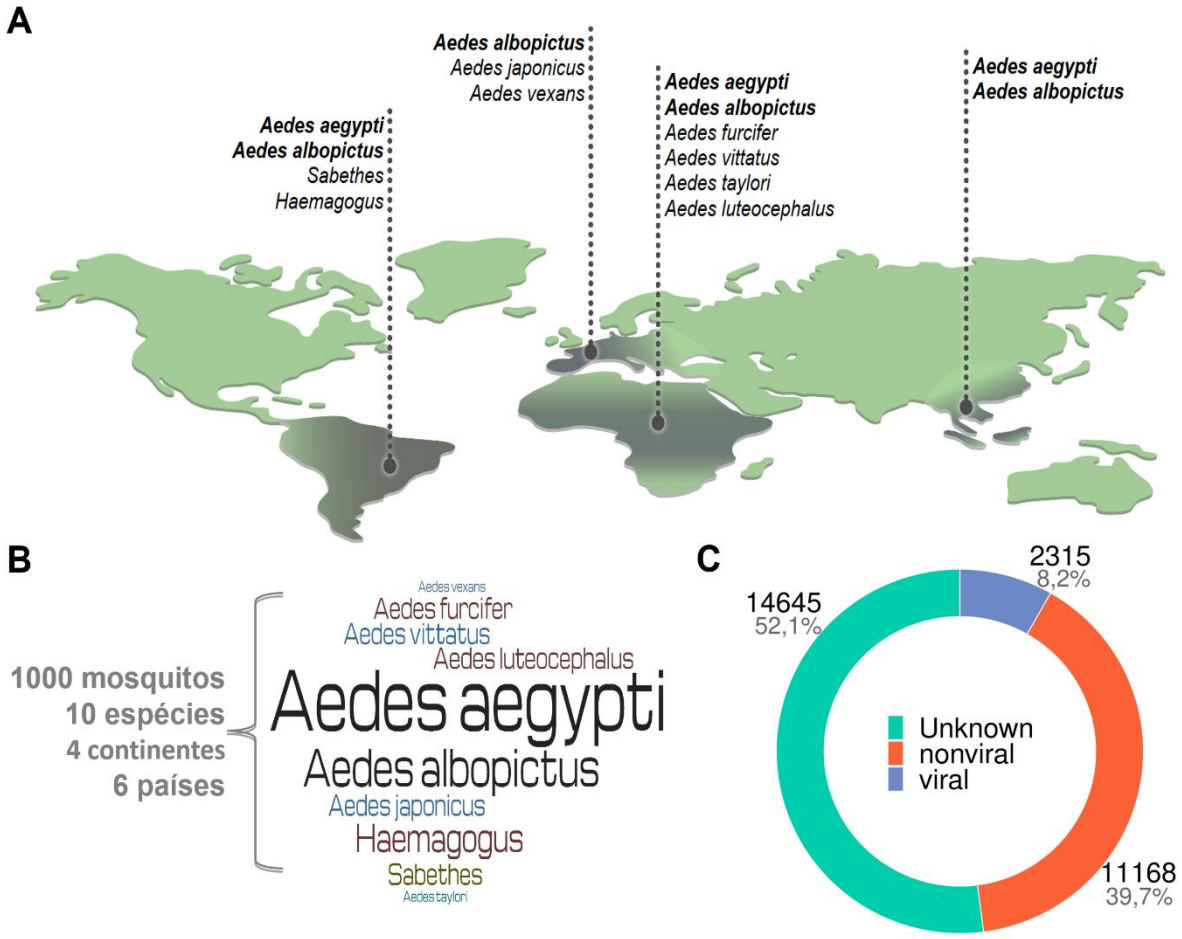


Figura 12 – Visão geral das amostras de mosquitos que deram origem as bibliotecas de pequenos RNAs e resultados globais das análises de similaridade de sequência. A. Espécies de mosquitos coletados por região. **B.** Composição das amostras sequenciadas. O tamanho da fonte no gráfico de logo é proporcional a quantidade de bibliotecas sequenciadas da espécie de mosquito. **C.** Proporções de contigs maiores que 200 nt montados nas 122 bibliotecas classificados por similaridade de sequência como: “Virais”, “Não-virais”, “Desconhecidos”**.**

Foram montados 28128 contigs maiores que 200 nt que foram classificados por similaridade de sequência em 2315 virais, 11168 não virais e 14645 desconhecidos (**Figura 12C; Tabela Suplementar 2**). A maioria dos contigs montados pertence a classe dos “desconhecidos”, sequências sem similaridade estatisticamente significativa com sequências contidas nos bancos de dados nt e nr do Genbank e sem evidências da presença de domínios proteicos conservados.

A maioria dos genomas animais contém sequências virais integradas conhecidas como elementos virais endógenos (EVEs) que são transcritos e geram pequenos RNAs. Assim, é importante que trabalhos de metagenômica viral utilizem estratégias para distinguir entre sequências derivadas de vírus exógenos e EVEs para uma melhor estimativa dos vírus realmente em circulação (AGUIAR; OLMO; MARQUES, 2016; DE ALMEIDA et al., 2021). Para abordar esse problema, utilizamos o perfil de pequenos RNAs associado à análise de ORFs dos contigs para separar sequências derivadas de vírus exógenos de EVEs (**Figura 9**). Com esse filtro, identificamos 994 sequências de EVEs putativas que foram removidas dos 2315 contigs inicialmente classificados como virais. As estatísticas de montagem dos contigs das quatro classes são mostradas na **Tabela 2**.

Tabela 2 - Estatísticas de montagem dos contigs classificados como “Viral”, “EVE”, “Não-Viral” e “Desconhecido” montados nas 122 bibliotecas.

	contigs		médio (nt)		Padrão	contig	> 1K	contigs > 1K
Viral	1321	1346	813.22	484	983.33	8702	292	635595
EVE	994	387	383.07	288	274.84	2796	45	61952
Não-Viral	11168	301	311.59	265	150.63	2449	83	111689
esconheci	14645	271	283.52	249	108.30	1875	32	40333

Utilizando o programa CD-HIT, os 1321 contigs virais restantes foram agrupados em 267 clusters de similaridade representados por somente uma sequência. Esses 267 contigs foram utilizados para avaliar a coocorrência dos vírus e determinar a quantidade de vírus únicos nas 122 bibliotecas. A ocorrência das

sequências virais em cada biblioteca foi inferida pela quantificação de reads de 20 a 22 nt mapeados nos contigs (**Figura 13**). A quantificação dos reads em cada biblioteca associada a inspeção dos resultados de similaridade de sequência e agrupamentos resultantes da clusterização hierárquica nos permitiram inferir os contigs pertencentes aos mesmos segmentos genômicos virais e a quantidade de vírus únicos nas 122 bibliotecas.

Diferentes combinações de métodos de agrupamentos e distâncias foram testadas para verificar a coerência dos agrupamentos formados avaliando contigs pertencentes a vírus previamente conhecidos como referência. A melhor combinação foi a de distância “euclidiana” com método “complete” (**Figura 13**). A robustez estatística dos agrupamentos formados foi inferida por cálculo de probabilidade com método de *bootstrap*. A maioria dos nós utilizados para suportar a inferência de contigs pertencentes ao mesmo vírus possuem valores de bootstrap maiores ou próximos de 90% (**Figura 14**). Uma versão interativa do heatmap que permite amplificar os agrupamentos e inspecionar os valores de RPKM dos contigs em cada biblioteca está disponível como aplicativo *Shiny* R no link: https://jpalmeida.shinyapps.io/Viral_smallRNA_cooccorrunce_20to22ntRPKM/.

A influência do intervalo de tamanhos de pequenos RNAs quantificados na estruturação dos dados que permitiu inferir os contigs pertencentes aos mesmos vírus foi avaliada pelo cálculo de dissimilaridade entre os contigs comparando a quantificação com reads de 20-22nt, 24-30nt, todos os reads, todos os reads removendo os de 20-22nt (**Figura 15**). O intervalo de 20-22nt apresenta uma melhor definição das estruturas internas do conjunto de dados.

A análise da coocorência dos contigs virais considerando os agrupamentos formados e alinhamentos dos contigs com segmentos virais similares, nos permitiu definir a presença de 34 vírus únicos nas 122 bibliotecas analisadas (**Figura 13,14**). Desses, dois arbovírus identificados, *Zika vírus* e *Usutu vírus* (*Flaviviridae*), serviram de controle positivo para a estratégia computacional aplicada, pois as bibliotecas em que foram identificados foram construídas a partir de mosquitos *Ae. japonicus* coletados em campo e artificialmente infectados com esse arbovírus em laboratório (ABBO et al., 2020).

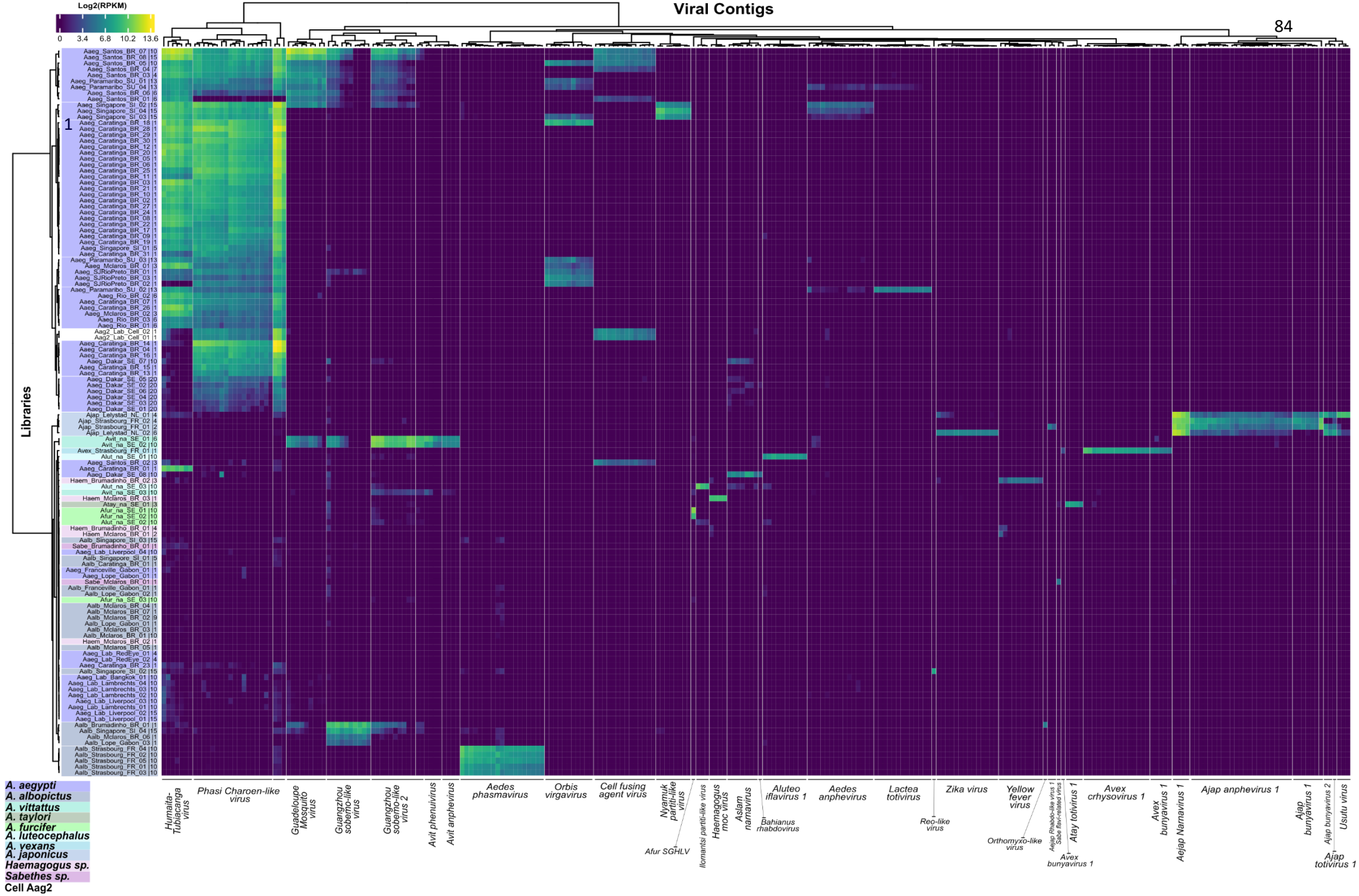

Log2(RPKM)
0
3.4
6.8
10.2
13.6
Viral Contigs
1
84
Libraries
Aaeg_Santos_BR_07 10
Aaeg_Santos_BR_08 15
Aaeg_Santos_BR_05 10
Aaeg_Santos_BR_04 7
Aaeg_Santos_BR_03 4
Aaeg_Paramaribo_SU_01 13
Aaeg_Paramaribo_SU_04 13
Aaeg_Santos_BR_06 6
Aaeg_Santos_BR_01 6
Aaeg_Singapore_SI_02 15
Aaeg_Singapore_SI_04 15
Aaeg_Singapore_SI_03 15
Aaeg_Caratinga_BR_18 1
Aaeg_Caratinga_BR_28 1
Aaeg_Caratinga_BR_29 1
Aaeg_Caratinga_BR_30 1
Aaeg_Caratinga_BR_12 1
Aaeg_Caratinga_BR_20 1
Aaeg_Caratinga_BR_05 1
Aaeg_Caratinga_BR_06 1
Aaeg_Caratinga_BR_25 1
Aaeg_Caratinga_BR_11 1
Aaeg_Caratinga_BR_03 1
Aaeg_Caratinga_BR_21 1
Aaeg_Caratinga_BR_10 1
Aaeg_Caratinga_BR_02 1
Aaeg_Caratinga_BR_27 1
Aaeg_Caratinga_BR_24 1
Aaeg_Caratinga_BR_08 1
Aaeg_Caratinga_BR_22 1
Aaeg_Caratinga_BR_17 1
Aaeg_Caratinga_BR_09 1
Aaeg_Caratinga_BR_19 1
Aaeg_Singapore_SI_01 5
Aaeg_Caratinga_BR_31 1
Aaeg_Paramaribo_SU_03 13
Aaeg_Mclaros_BR_01 3
Aaeg_SJRioPreto_BR_01 1
Aaeg_SJRioPreto_BR_03 1
Aaeg_SJRioPreto_BR_02 1
Aaeg_Paramaribo_SU_02 13
Aaeg_Rio_BR_02 6
Aaeg_Caratinga_BR_07 1
Aaeg_Caratinga_BR_26 1
Aaeg_Mclaros_BR_02 3
Aaeg_Rio_BR_03 6
Aaeg_Rio_BR_01 6
Aag2_Lab_Cell_02 1
Aag2_Lab_Cell_01 1
Aaeg_Caratinga_BR_14 1
Aaeg_Caratinga_BR_04 1
Aaeg_Caratinga_BR_16 1
Aaeg_Dakar_SE_07 10
Aaeg_Caratinga_BR_15 1
Aaeg_Caratinga_BR_13 1
Aaeg_Dakar_SE_05 20
Aaeg_Dakar_SE_02 20
Aaeg_Dakar_SE_06 20
Aaeg_Dakar_SE_04 20
Aaeg_Dakar_SE_03 20
Aaeg_Dakar_SE_01 20
Ajap_Lelystad_NL_01 4
Ajap_Strasbourg_FR_02 4
Ajap_Strasbourg_FR_01 2
Ajap_Lelystad_NL_02 6
Avit_na_SE_01 6
Avit_na_SE_02 10
Avex_Strasbourg_FR_01 1
Alut_na_SE_01 10
Aaeg_Santos_BR_02 3
Aaeg_Caratinga_BR_01 1
Aaeg_Dakar_SE_08 10
Haem_Brumadinho_BR_02 3
Alut_na_SE_03 10
Avit_na_SE_03 10
Haem_Mclaros_BR_03 1
Atay_na_SE_01 3
Afur_na_SE_01 10
Afur_na_SE_02 10
Alut_na_SE_02 10
Haem_Brumadinho_BR_01 4
Haem_Mclaros_BR_01 2
Aalb_Singapore_SI_03 15
Sabe_Brumadinho_BR_01 1
Aaeg_Lab_Liverpool_04 10
Aalb_Singapore_SI_01 5
Aalb_Caratinga_BR_01 1
Aaeg_Franceville_Gabon_01 1
Aaeg_Lope_Gabon_01 1
Sabe_Mclaros_BR_01 1
Aalb_Franceville_Gabon_01 1
Aalb_Lope_Gabon_02 1
Afur_na_SE_03 10
Aalb_Mclaros_BR_04 1
Aalb_Mclaros_BR_07 1
Aalb_Mclaros_BR_02 9
Aalb_Lope_Gabon_01 1
Aalb_Mclaros_BR_03 1
Aalb_Mclaros_BR_01 10
Haem_Mclaros_BR_02 1
Aalb_Mclaros_BR_05 1
Aaeg_Lab_RedEye_01 4
Aaeg_Lab_RedEye_02 4
Aaeg_Caratinga_BR_23 1
Aalb_Singapore_SI_02 15
Aaeg_Lab_Bangkok_01 10
Aaeg_Lab_Lambrechts_04 10
Aaeg_Lab_Lambrechts_03 10
Aaeg_Lab_Lambrechts_02 10
Aaeg_Lab_Liverpool_03 10
Aaeg_Lab_Lambrechts_01 10
Aaeg_Lab_Liverpool_02 15
Aaeg_Lab_Liverpool_01 15
Aalb_Brumadinho_BR_01 1
Aalb_Singapore_SI_04 15
Aalb_Mclaros_BR_06 1
Aalb_Lope_Gabon_03 1
Aalb_Strasbourg_FR_04 10
Aalb_Strasbourg_FR_02 10
Aalb_Strasbourg_FR_05 10
Aalb_Strasbourg_FR_01 10
Aalb_Strasbourg_FR_03 10
A. aegypti
A. albopictus
A. vittattus
A. taylori
A. furcifer
A. luteocephalus
A. vexans
A. japonicus
Haemagogus sp.
Sabethes sp.
Cell Aag2
Humaita-Tubiacanga virus
Phasi Charoen-like virus
Guadeloupe Mosquito virus
Guangzhou sobemo-like virus
Guangzhou sobemo-like virus 2
Avit phenuivirus
Avit anphevirus
Aedes phasmavirus
Orbis virgavirus
Cell fusing agent virus
Nyamuk partiti-like virus
Afur SGHLV
Ilomantsi partiti-like virus
Haemagogus moc virus
Aslam narnavirus
Aluteo iflavirus 1
Bahianus rhabdovirus
Aedes anphevirus
Lactea totivirus
Reo-like virus
Zika virus
Yellow fever virus
Orthomyxo-like virus
Aejap Rhabdo-like virus 1
Sabe flavi-related virus
Atay totivirus 1
Avex bunyavirus 1
Avex chrysovirus 1
Avex bunyavirus 1
Aejap Narnavirus 1
Ajap anphevirus 1
Ajap bunyavirus 1
Ajap bunyavirus 2
Ajap totivirus 1
Usutu virus

Figura 13 – Coocorrência dos 267 contigs virais não-redundantes nas bibliotecas das 10 espécies de mosquitos amostradas. O heatmap representa a contagem normalizada (RPKM) de pequenos RNAs de 20-22 nt alinhados a cada um dos 267 contigs virais não-redundantes (colunas) nas 122 bibliotecas analisadas (linhas). Os nomes das bibliotecas indicam a espécies, local de coleta e o número de mosquitos na amostra sequenciada (“spp_local_ID” | “nro. de mosquitos”). A coocorrência dos contigs no mesmo agrupamento hierárquico em conjunto a curadoria dos resultados de alinhamentos locais de cada contig permitiram a definição de 34 vírus únicos. As bibliotecas da mesma espécie foram destacadas com a mesma cor. Os dendrogramas apresentados tem como combinação de método e distancia para agrupamentos das linhas "ward.D" com “euclidean” e para as colunas “complete” com “euclidean”.

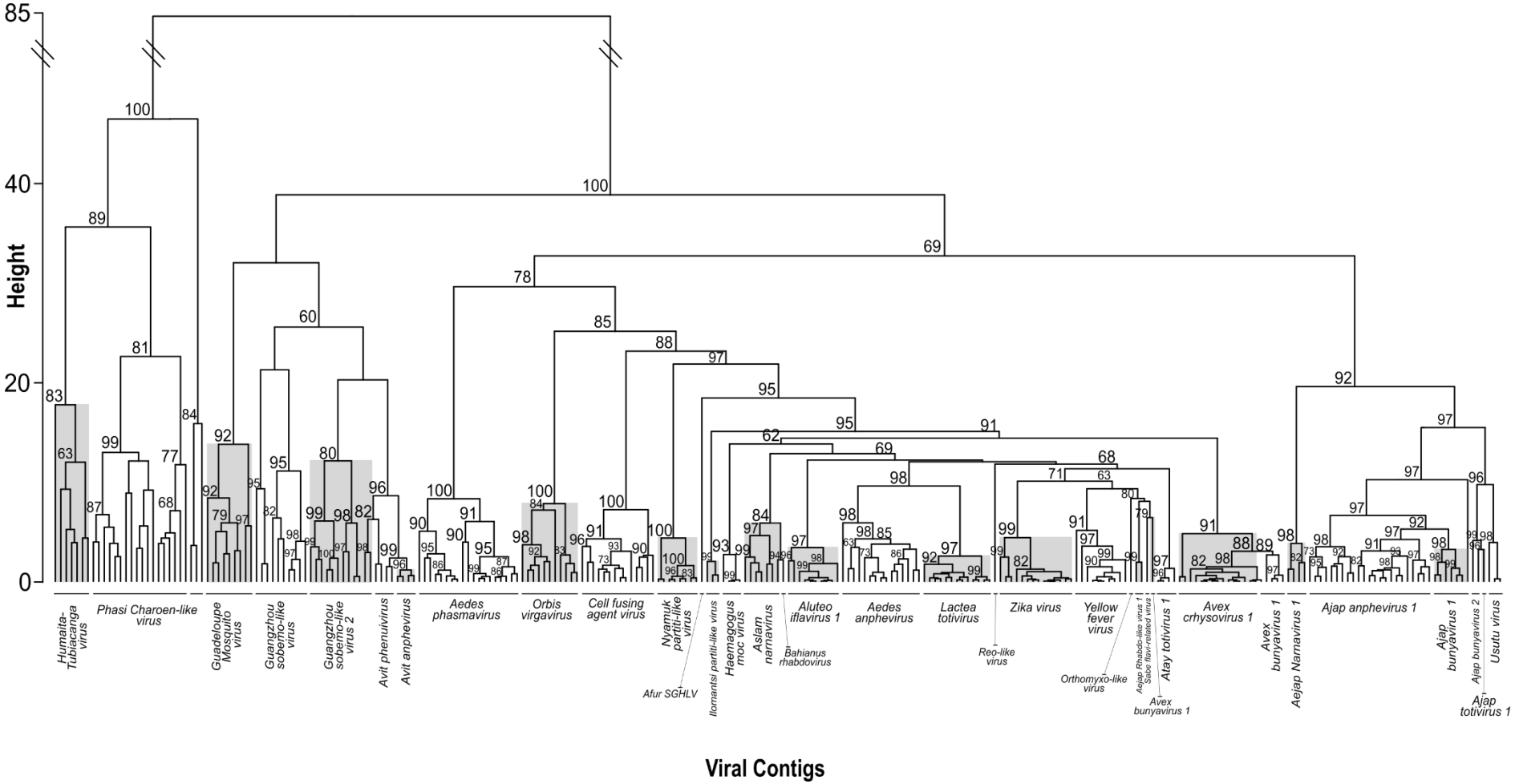


Figura 14 – Suporte estatístico dos agrupamentos de contigs utilizados para inferir a quantidade de vírus únicos nas 122 bibliotecas. Os valores de *bootstrap* dos agrupamentos inspecionados para inferir os contigs pertencentes ao mesmo vírus nas bibliotecas (**Figura 12**) estão anotados próximos aos nós de origem dos agrupamentos. O eixo y (height) representa a distância euclidiana dos ramos. Destaques intercalados na cor cinza foram usados apenas para facilitar a visualização de agrupamentos próximos.

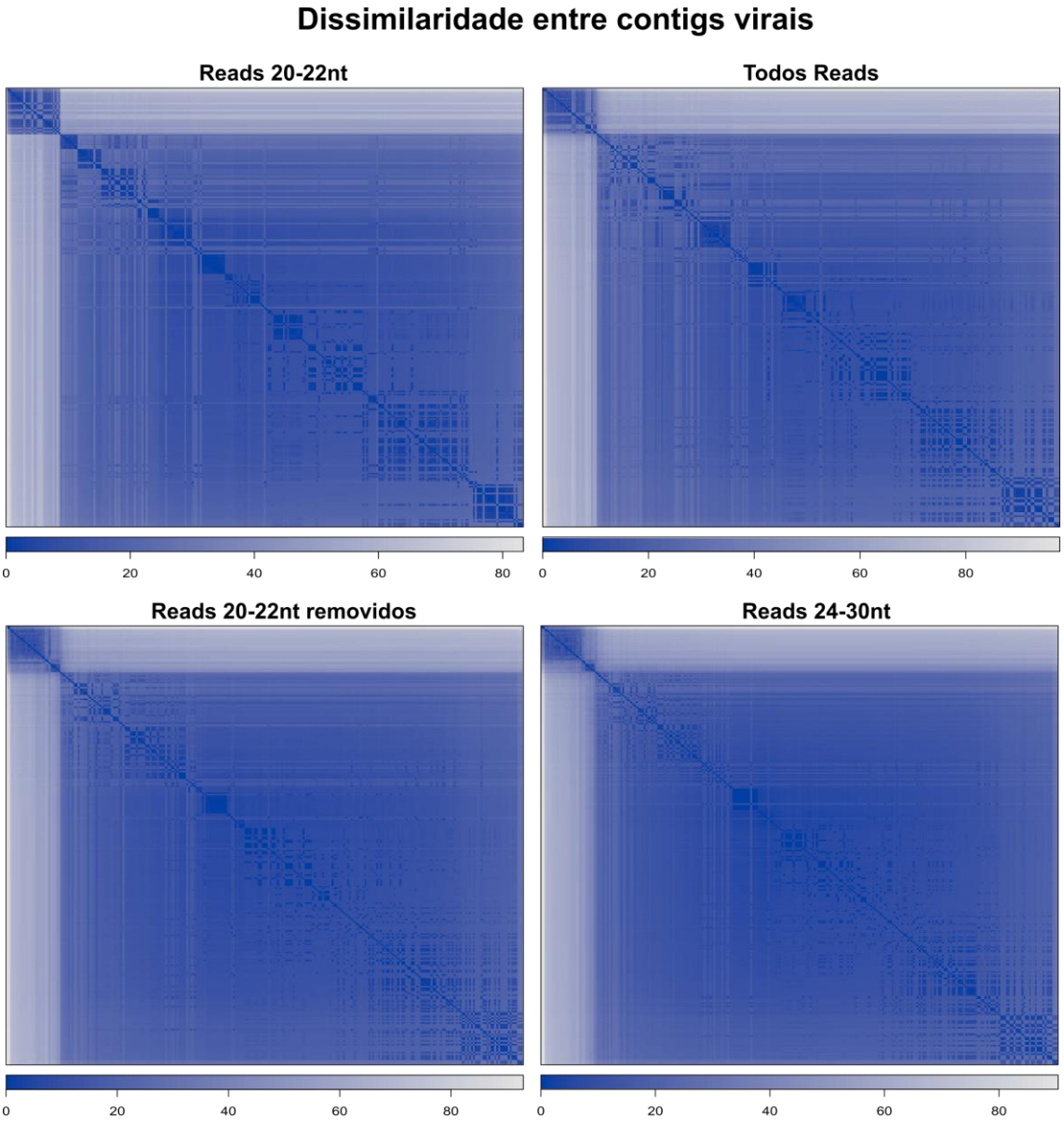


Figura 15 – Matrizes de dissimilaridade dos contigs representados pela quantificação de diferentes intervalos de pequenos RNAs. Os gráficos representam matrizes simétricas com os valores de dissimilaridade entre os contigs (azul baixa distância euclidiana, branco alta distância euclidiana) representados pela quantificação de quatro intervalos de pequenos RNAs. O intervalo de 20-22nt sozinho apresenta uma estruturação mais definida dos dados quando comparada aos demais intervalos;

A **Tabela 3** contém os resultados dos alinhamentos locais dos maiores contigs montados para cada arbovírus e a porcentagem total de cobertura por todos contigs

montados para cada referência. A **Figura 16A** contém o perfil de pequenos RNAs e cobertura de reads de 21nt dos maiores contigs montados para esses dois arbovírus nas bibliotecas de *A. japonicus*.

Tabela 3 – Cobertura por contigs montados dos arbovírus ZIKV e USUV.

Arbovirus	Maior Contig (nt)	Cobertura Contig	Ident	E value	Referência	Cobertura Referência	Cobertura Referência por todos contigs
Usutu virus	8702	99%	99.99%	0.0	MT188658.1	74,00%	99,00%
Zika virus	1267	98%	100.00%	0.0	MK566202.1	11,00%	81,00%

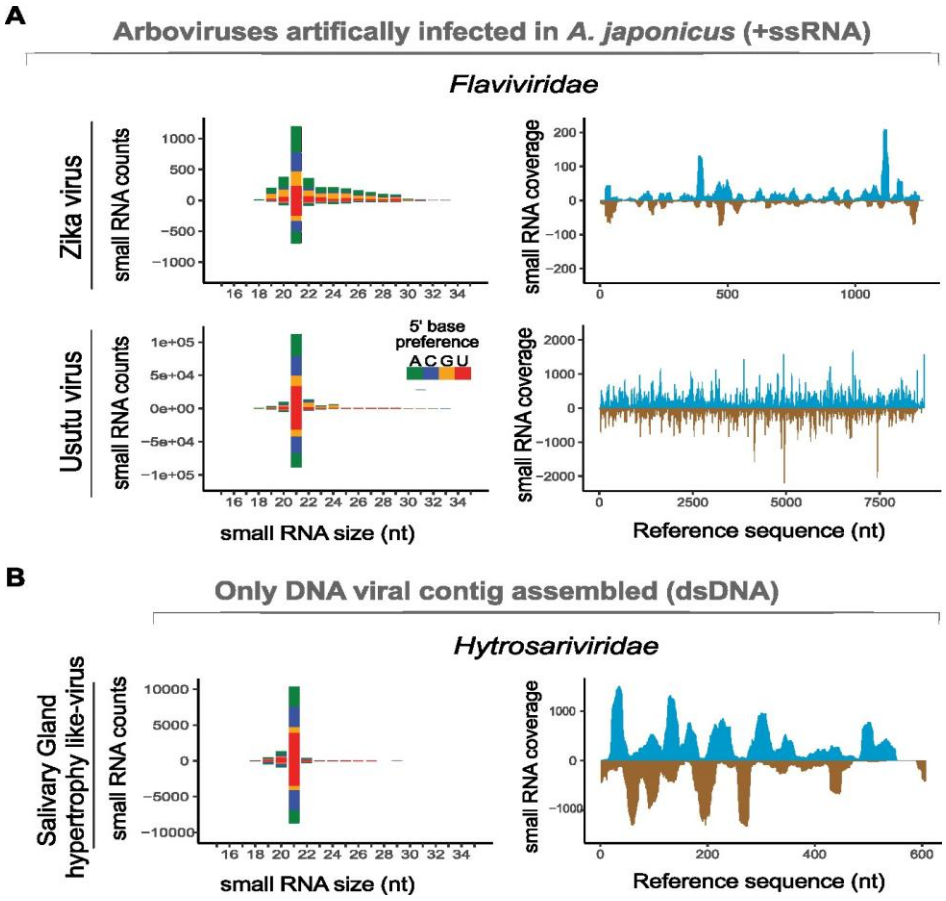


Figura 16 – Perfil e cobertura de pequenos RNAs das sequências de arbovírus infectando *Ae. japonicus* e do único segmento de um vírus de DNA encontrado em *A. furcifer*. Figuras da esquerda mostram a distribuição de tamanho dos pequenos RNAs separados por fita (senso acima e antisenso abaixo do 0 do eixo y) e a frequência de nucleotídeo da primeira base na extremidade 5' representada pelas cores nas barras. Figuras da direita mostram a cobertura de pequenos RNAs de 21nt (azul fita senso, marrom fita antisenso). Os reads das bibliotecas de *Ae. japonicus* foram alinhados aos maiores contigs montados para os arbovírus ZIKV e USUV.

Tabela 4 – Resumo dos 32 potenciais vírus montados nas 122 bibliotecas com resultados de alinhamentos locais dos contigs representativos.

Spp.	Baltimore	Família	Vírus	Segmento	Tipo	Contig (nt)	Cobertura Contig	Ident.	E-value	Referência
A. aegypti	+ssRNA	*Flaviviridae*	*Cell Fusing Agent virus*	ns	nt	1073	99,00%	98.78%	0.0	LR694073.1
		Narnaviridae	*Aslam narnavirus*	ns	aa	380	93,00%	46.61%	3,00E-23	QED42904.1
		Virgaviridae	*Orbis virgavirus*	ns	aa	787	78%	41.0%	4.0e-8	BBQ04461.1
		Unclassified	*Humaita-Tubiacanga virus*	1	nt	2786	100,00%	99.21%	0.0	KR003801.1
				2		1360	100,00%	99.63%	0.0	KR003802.1
		Unclassified	*Guadeloupe mosquito virus*	1	nt	1131	100,00%	99.56%	0.0	MW434816.1
				2		1664	98,00%	98.05%	0.0	MW434804.1
	-ssRNA	*Phenuiviridae*	*Phasi Charoen-like vírus*	L	nt	6149	99,00%	99.61%	0.0	NC_038262.1
				M		3672	100,00%	99.59%	0.0	NC_038261.1
				S		840	99,00%	97.62%	0.0	MN866293.1
		Xinmoviridae	*Aedes anphevirus*	ns	nt	1057	99,00%	99.81%	0.0	MH430652.1
		Rhabdoviridae	*Bahianus rhabdovirus*	ns	aa	911	94,00%	32.18%	3,00E-28	AJG39213.1
	dsRNA	*Totiviridae*	*Lactea totivirus*	ns	nt	1623	87,00%	94.83%	0.0	MW434943.1
		Partitiviridae	*Nyamuk partiti-like virus*	1	aa	909	99,00%	77.74%	8,00E-178	QHA33901.1
				2		1286	82,00%	39.11%	1,00E-55	QHA33900.1
A. albopictus	+ssRNA	*Unclassified*	*Guangzhou sobemo-like virus*	1	nt	2441	100,00%	98.16%	0.0	MT096519.1
				2		411	98,00%	98.27%	0.0	MW434910.1
	-ssRNA	*Phasmaviridae*	*Aedes phasmavirus*	L	nt	4977	99,00%	98.63%	0.0	MW434659.1
				M		1192	97,00%	99.31%	0.0	MT361042.1
				S		2045	99,00%	99.07%	0.0	MW434639.1
		Orthomyxoviridae	*Orthomyxo-like virus*	5	nt	268	100,00%	100.00%	1,00E-135	BK059435.1
	dsRNA	*Unclas. Reovirales*	*Reo-like virus*	unknown	aa	295	80,00%	35.37%	1,00E-04	QPN36924.1
A. japonicus	+ssRNA	*Narnaviridae*	*Aedes japonicus Narnavirus 1*	1	nt	3152	100%	100%	0.0	MK984721.2
	-ssRNA	*Phenuiviridae*	*Aedes japonicus Bunyavirus 1*	L	aa	1814	96%	35%	5.0e-9	QHA33858.1
				M		2029	85%	42%	1.0e-142	QHA33860.1
				S		1432	55%	44%	8.0e-62	QHA33859.1
		Phasmaviridae	*Aedes japonicus Bunyavirus 2*	M	aa	2133	88%	44%	0	YP_009305132.1
				S		2042	52%	54%	5.0e-126	YP_009305134.1
		Xinmoviridae	*Aedes japonicus Anphevirus 1*	ns	aa	2932	99%	58%	0.0	AWW13479.1
		Rhabdoviridae	*Aedes japonicus Rhabdovirus 1*	ns	aa	482	57%	38%	2.0e-8	QIS62330.1
	dsRNA	*Totiviridae*	*Aedes japonicus Totivirus 1*	ns	aa	6498	47%	60%	0.0	AJT39583.1
A. vexans	-ssRNA	*Unclas Bunyavirales*	*Aedes vexans Bunyavirus 1*	M	nt	401	100,00%	89.53%	3,00E-139	MH703046.1
				S	nt	892	88,00%	92.55%	0.0	NC_040758.1
	dsRNA	*Chrysoviridae*	*Aedes vexans chrysovirus 1*	1	aa	3180	95,00%	64.86%	0.0	UUV42200.1
				2		1500	88,00%	48.87%	3,00E-147	UUV42192.1
				3		791	88,00%	56.84%	4,00E-89	UUV42194.1
A. vittattus	+ssRNA	*Unclassified*	*Guadeloupe mosquito virus 2*	1	aa	534	98,00%	90.91%	1,00E-115	QEM39257.1
				2	nt	785	80,00%	80.98%	4,00E-136	MN053804.1
	-ssRNA	*Phenuiviridae*	*Aedes vittattus phenuivirus 1*	L	aa	818	98,00%	53.16%	5,00E-100	UYE93924.1
				M		1953	79,00%	32.12%	5,00E-81	QRW41943.1
				S		1698	73,00%	41.01%	1,00E-92	API61886.1
		Xinmoviridae	*Aedes vittattus anphevirus 1*	ns	aa	3133	99,00%	79.58%	0.0	AWW13453.1
A. luteocephalus	+ssRNA	*Iflaviridae*	*Aedes luteocephalus iflavirus 1*	ns	aa	398	99,00%	88.64%	2,00E-67	UYE93723.1
	dsRNA	*Partitiviridae*	*Illomantsi partiti-like virus*	1	nt	565	96,00%	86.81%	1,00E-169	OP019950.1
A. taylori	dsRNA	*Totiviridae*	*Aedes taylori totivirus 1*	ns	aa	346	96,00%	48.65%	2,00E-23	YP_007761589.1
A. furcifer	dsDNA	*Hytrosariviridae*	*Salivary Gland hypertrophy like-virus*	ns	aa	607	40,00%	35.63%	3,00E-05	YP_001686988.1
Haemagogus sp.	+ssRNA	*Flaviviridae*	*Yellow fever virus*	ns	nt	307	100%	100%	4E-157	OR052147.1
	dsRNA	*Sedoreoviridae*	*Haemagogus sedoreovirus*	1	aa	4349	87,00%	70.52%	0.0	DAZ85690.1
				2		2784	95,00%	63.99%	0.0	DAZ85691.1
				3		1167	87,00%	33.92%	3,00E-58	UPT53659.1
				4		1224	88,00%	70.91%	0.0	DAZ85694.1
S. purpureos	+ssRNA	*Flaviviridae*	*Sabethes flavi-related virus*	ns	aa	318	100,00%	54.63%	8,00E-29	QTJ63572.1

Os 32 potenciais vírus naturalmente infectando os mosquitos estão distribuídos em, ao menos, 17 famílias virais levando em conta a classificação taxonômica do melhor resultado de alinhamento local dos contigs representativos dos segmentos virais de cada vírus **(Tabela 4)**. Dos 32 vírus únicos, 13 apresentaram similaridade de sequência significativa a nível de nucleotídeos, evidenciando a alta proximidade com as sequências das referências contidas nos bancos de dados e provavelmente se trata de vírus previamente descobertos. Os 19 prováveis vírus restantes apresentaram similaridade de sequência significativa apenas a nível de aminoácido, representando sequências pertencentes a potenciais vírus novos.

Analisando os resultados dos alinhamentos locais e das buscas por domínios proteicos conservados buscamos por contigs com similaridade com replicases virais. Para seis potenciais vírus (**Tabela 4**), não foram montadas sequências com similaridade ou presença de domínio de replicases virais. Dois desses vírus são da classe Bunyavirales, vírus tri segmentados de fita negativa, encontrados em mosquitos diferentes, *A. japonicus* e *A. vexans*, ambos com sequências correspondentes aos segmentos gnômicos que codificam a glicoproteína (M) e o capsídeo (S). Os demais vírus para os quais não foram encontradas sequências de replicases pertencem a vírus monopartites, sendo um deles o único vírus de DNA que encontramos similaridade de sequência em nossos contigs virais montados em bibliotecas de *A. furcifer*. Porém, apesar de um perfil de pequenos RNAs evidente da via de siRNA (**Figura 16B**), foi montado apenas um contig de 607 nucleotídeos com similaridade ao vírus de DNA *Glossina pallidipes salivary gland hypertrophy virus* com genoma de cerca de 190 Kb. Devido a repetição do fenômeno em bibliotecas independentes de mosquitos distintos, consideramos que os bunyavirus *Aedes japonicus bunyavirus* 2 e *Aedes vexans bunyavirus* podem se tratar de novos vírus capazes de usar a RdRP de outros vírus. Os demais vírus para os quais não encontramos replicases, *Orthomyxo-like vírus*, *Reo-like vírus*, *Aedes japonicus Rhabdovirus* e *Salivary Gland hypertrophy like-virus* foram chamados de vírus putativos e considerados com maior incerteza de que se trata de sequências representativas de vírus únicos. Na **Figura 17,** é mostrado um resumo da classificação de Baltimore e das famílias virais e os valores de RPKM dos reads de 20-22nt contabilizados para os contigs montados para os potenciais 28 vírus únicos

encontrados no viroma das nove espécies. Além dos fragmentos do potencial vírus de DNA, não foram encontradas sequências virais nas três bibliotecas de *A. furcifer.*

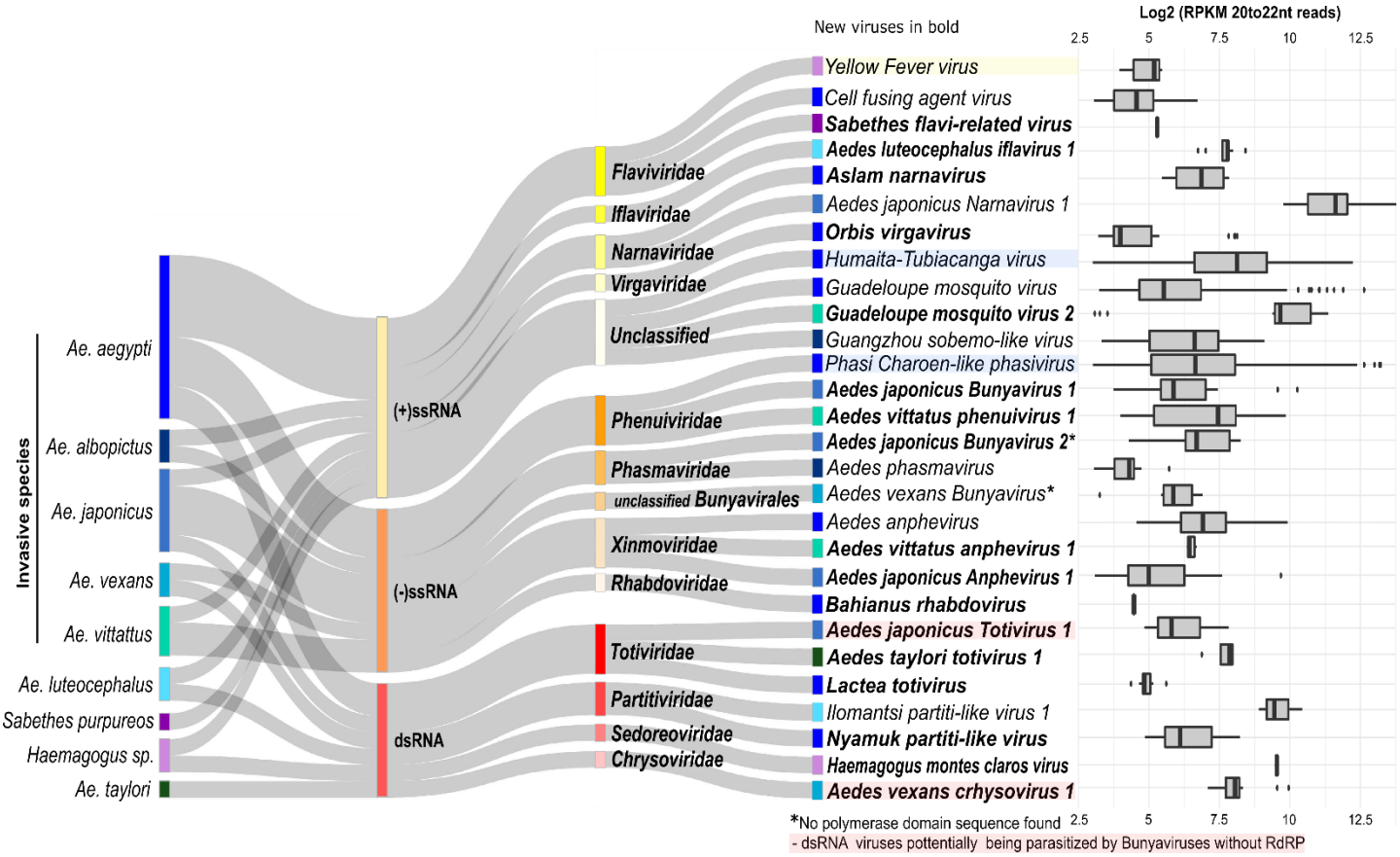


Figura 17 - Diagrama de Sankey com Classificação de Baltimore, Famílias Virais e resumo da carga de pequenos RNAs virais dos vírus encontrados nas 10 espécies de mosquitos analisadas. *São mostrados os 28 vírus pertencentes a, ao menos, 14 famílias virais diferente, todos vírus com genoma de RNA. São mostradas as médias de RPKM da carga de pequenos RNAs virais dos contigs virais de cada vírus nas bibliotecas em que foram encontrados. As cores dos quadrados de cada vírus remetem a espécie hospedeira. Os 17 potenciais vírus novos estão destacados com fonte em negrito. PCLV e HTV, os vírus mais prevalentes e que atingem as maiores cargas virais em A. aegypti estão destacados em azul. O único arbovírus naturalmente infectando mosquitos encontrados nesse trabalho, YFV, está destacado em amarelo, com um dos menores intervalos de RPKM de pequenos RNAs virais coerente com o alto CT de qPCR da amos7ra em que foi encontrado. Foram marcados com * os dois Bunyavirus para os quais não conseguimos encontrar sequências de RdRP, mas que os demais segmentos possuem informação o suficiente para inferir a presença desses vírus. Foram destacados em vermelho os dois vírus de dsRNA que coocorrem com os Bunyavirus sem RdRP em mosquitos diferentes (Ae. japonicus e A.vexans).*

O único arbovírus detectado infectando os mosquitos naturalmente foi o *Yellow Fever virus* em uma biblioteca de *Haemagogus sp.* de mosquitos coletados em Brumadinho – MG (**Figura 13,17; Tabela 4**). Foram montados contigs pequenos,

todos menores que 500 nt e a carga de pequenos RNAs virais é uma das mais baixas entre os vírus identificados (**Figura 17**). Tais resultados de detecção estão relacionados a baixa carga viral do arbovírus que foi confirmado na amostra por RT-qPCR com valor de Ct = 32 (resultado do RT-qPCR gerado pela doutoranda Ana Luiza Cruz, laboratório da Profa. Betânia Drumond ICB-UFMG). Para verificar a presença de outros arbovírus comumente transmitidos por mosquitos que podiam possuir carga viral insuficiente para a montagem de contigs, as reads de todas as bibliotecas foram alinhadas contra uma referência de arbovírus. Apenas as bibliotecas de mosquitos sabidamente infectados com ZIKV e USUV (*Ae. japonicus* da Holanda) e a biblioteca de *Haemagogus de* Brumadinho - MG tiveram alinhamentos com referências genômicas correspondentes aos arbovírus com contigs montados (**Tabela 5**).

Tabela 5 – Bibliotecas com alinhamentos contra a referência de sequências de arbovírus.

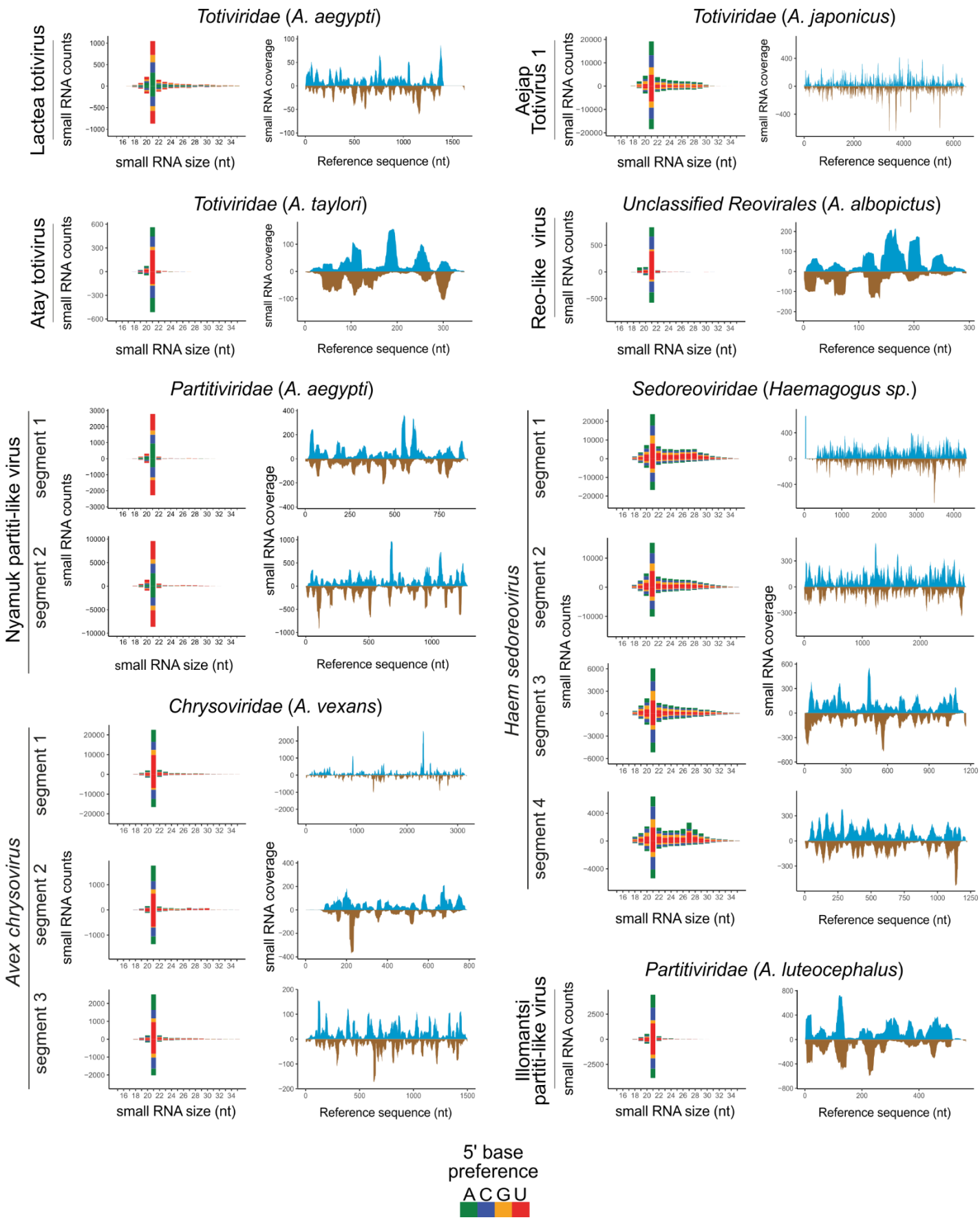
Biblioteca	Referência	reads alinhadas		Cobertura (%)		Profundidade média	
		v0	v1	v0	v1	v0	v1
Ajap_Lelystad_NL_01	USUV (NC_006551.1)	168479	281285	87,81	99,23	324,22	542,16
Ajap_Lelystad_NL_02	ZIKV (NC_075423.1)	11779	12516	98,71	99,25	24,66	26,21
Haem_Brumadinho_BR_02	YFV (NC_002031.1)	426	1316	9,72	30,45	0,9	2,76

***v = nro de não pareamentos (*missmatchs*) permitidos**

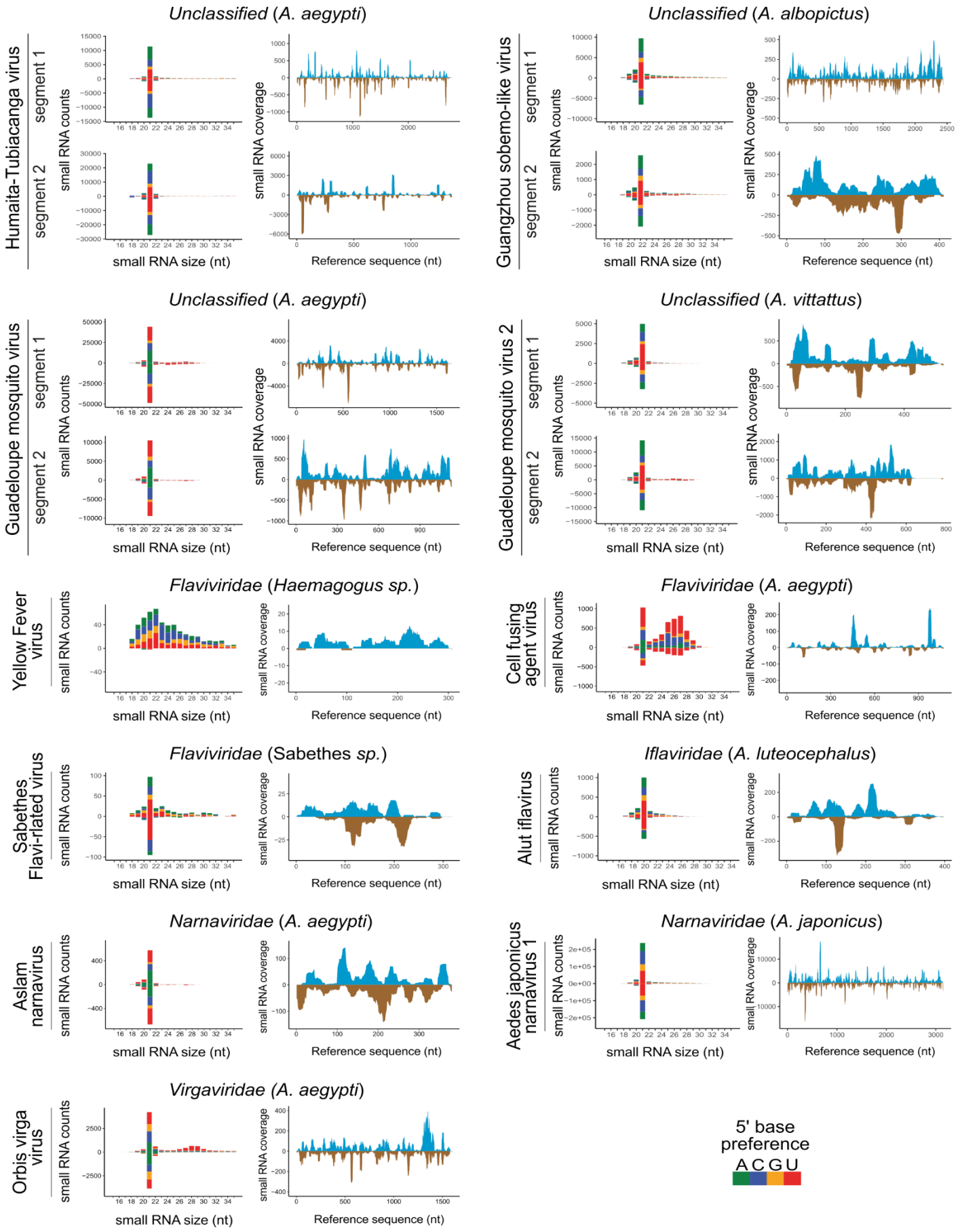
5.1.1 - Análise do pequenos RNAs virais do viroma de mosquitos vetores

O perfil de pequenos RNAs e cobertura de reads de 21 nt dos contigs representativos dos segmentos virais obtidos foram agrupados pela classificação de Baltimore do vírus de origem e mostrados na **Figura 18**. Todos os potenciais vírus encontrados apresentam um perfil de ativação da via antiviral de siRNA com picos simétricos da quantificação de reads de 21nt oriundos das fitas senso e antisenso. A simetria da cobertura dos reads de 21 nt das fitas sense e antisenso evidencia o processamento de substrato dsRNA pela DIcer-2. O único vírus que não apresenta perfil de ativação da via de siRNA arbovírus, YFV. O vírus possui pequenos RNAs oriundos apenas da fita senso com assimetria de fita do perfil de pequenos RNAs gerado similar ao de vírus que conseguem inibir as vias de RNAi levando a degradação por outras RNases.

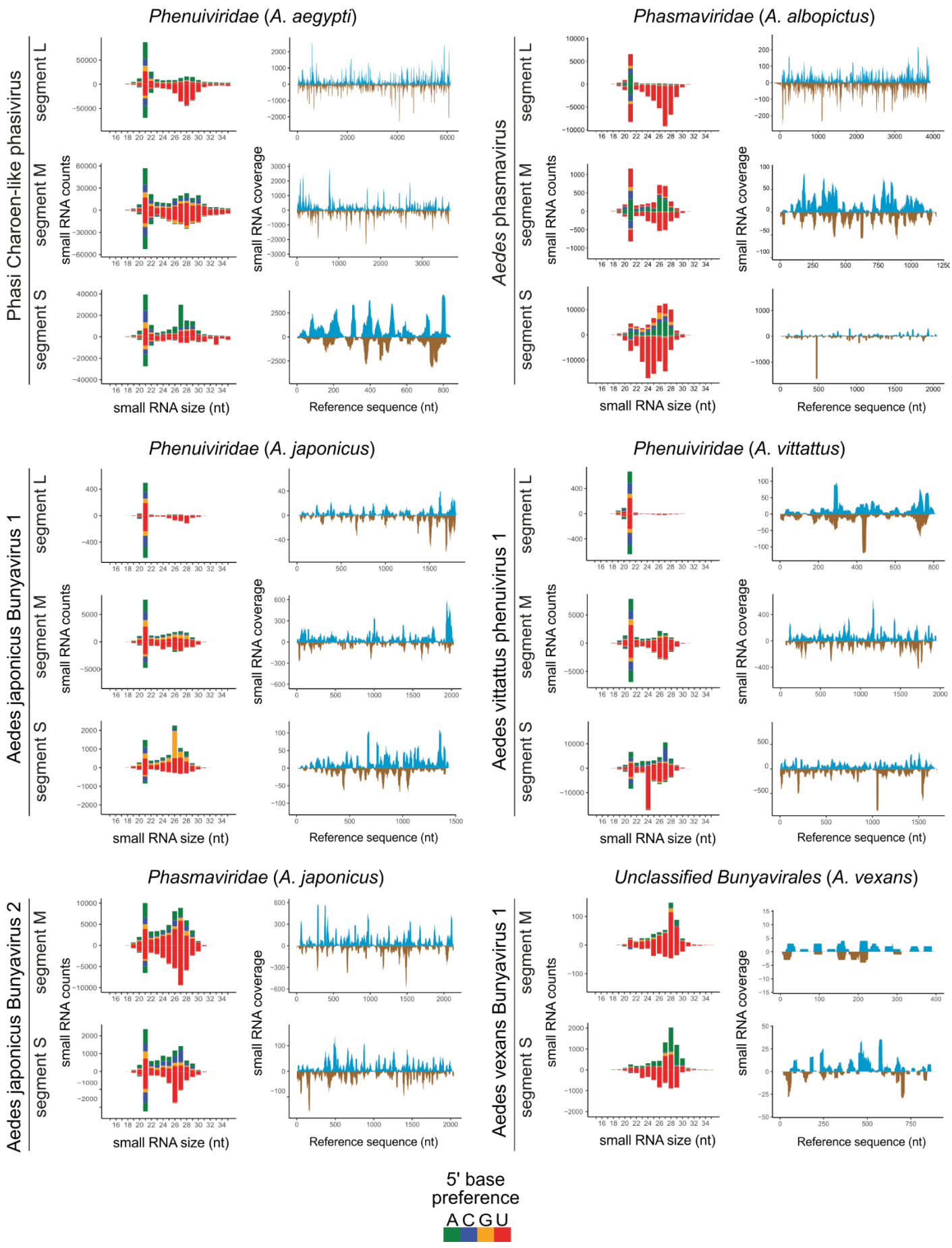
dsRNA viruses (Class III)



+ssRNA viruses (Class IV)



-ssRNA segmented viruses (Classe V)



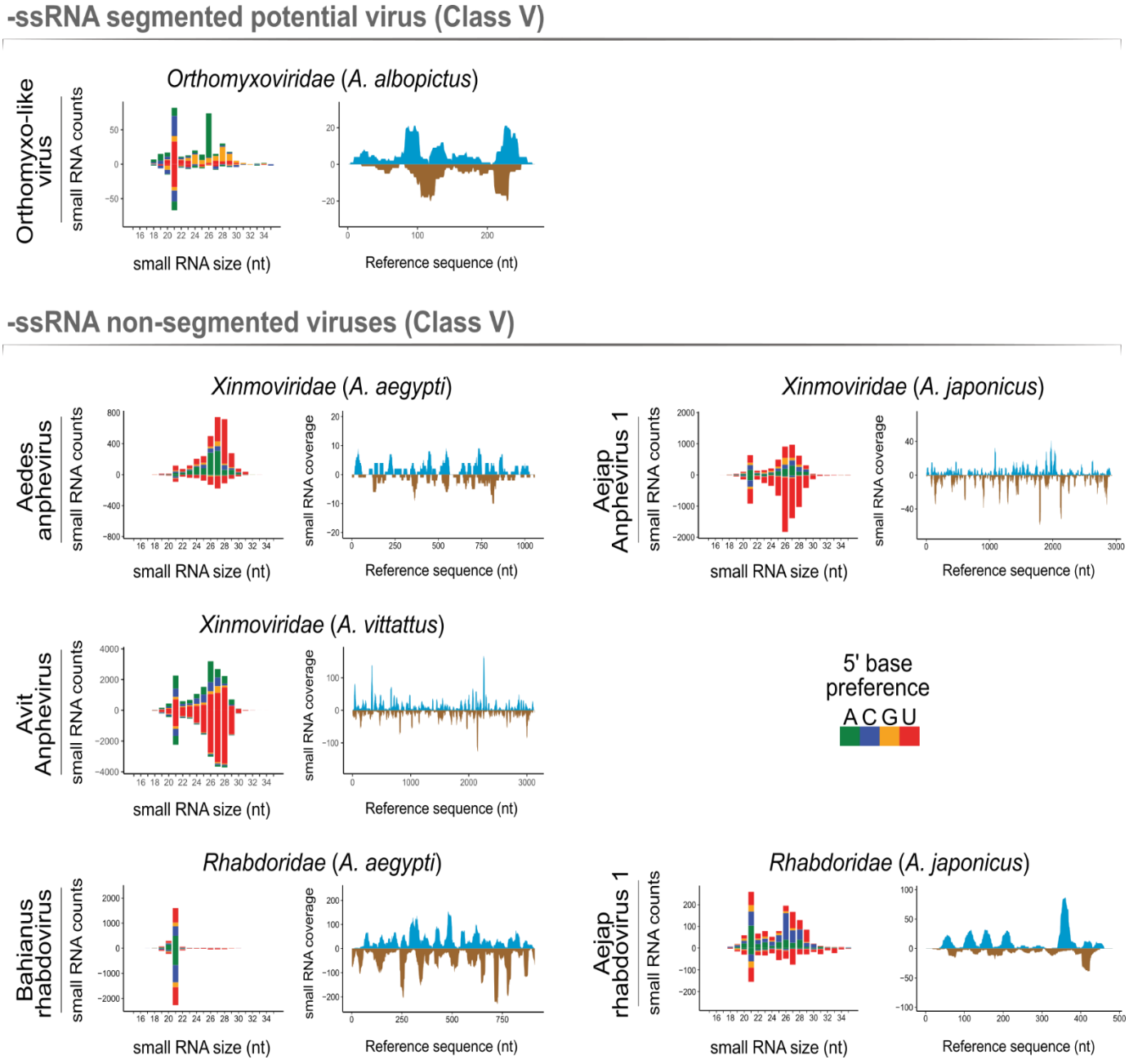
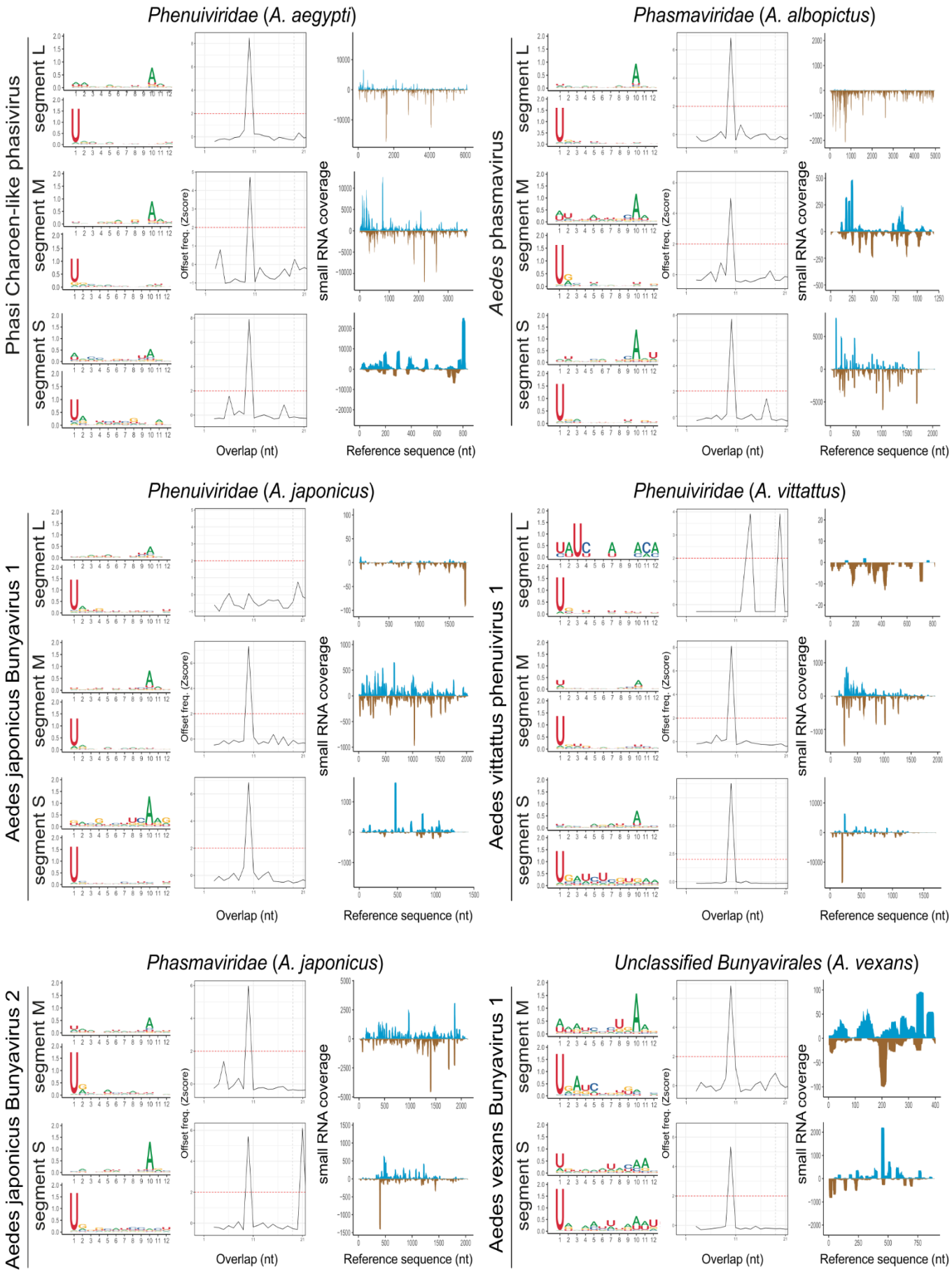


Figura 18 – Catálogo do perfil de pequenos RNAs virais de nove espécies de mosquitos coletados ao redor do globo. Os perfis de pequenos RNAs das sequências representativas de 31 vírus foram agrupadas por Classificação de Baltimore. Para cada vírus está descrito a Família viral e espécie hospedeira. Para os vírus segmentados foram mostrados os perfis para os maiores contigs montados para cada segmento. Os reads alinhados aos contigs representativos são oriundos da biblioteca em que o maior contig do vírus foi montado. Para cada vírus, figuras da esquerda mostram a distribuição de tamanho dos pequenos RNAs separados por fita (senso acima e antisenso abaixo do 0 do eixo y) e a frequência de nucleotídeo da primeira base na extremidade 5' representada pelas cores nas barras; figuras da direita mostram a cobertura de pequenos RNAs de 21nt (azul fita senso, marrom fita antisenso).

Para alguns vírus também foi observado perfil de pequenos RNAs coerente com ativação da via de piRNAs. Todos os vírus com genoma -ssRNA segmentados, todos da ordem *Bunyavirales*, apresentam segmentos genômicos com perfil de piRNAs em conjunto ao de siRNA. Há uma prevalência na geração de piRNAs primários oriundos da fita correspondente ao genoma (antisense a orientação das ORFs usadas para a quantificação) caracterizados pelo enriquecimento de uridina na extremidade 5' das reads de ~24-30 nt. A exceção em relação a produção de piRNAs é o segmento L do *Aedes vittaus phenuivirus* 1 que não parece gerar piRNAs. O mesmo segmento do também phenuivirus,, *Aedes japonicus Bunyavirus* 1, apresenta uma produção de piRNAs primários residuais. Os segmentos L dos demais bunyavirus, PCLV e Aedes anphevirus apresentam apenas piRNAs primários quase que completamente oriundos da fita do genoma. Os vírus de genoma -ssRNA não segmentados também apresentaram perfil de geração de piRNAs, com exceção do *Bahianos rhabdovirus*. Para muitos desses vírus e segmentos genômicos virais a produção de piRNAs supera a de siRNAs. O único vírus não -ssRNA que apresenta pequenos RNAs com perfil de piRNAs é o CFAV com genoma +ssRNA.

Para os 12 potenciais vírus com evidência de produção de piRNAs virais foi avaliada a produção de piRNAs secundários pela ativação do mecanismo de amplificação "*ping pong*" (**Figura 19**). Com exceção dos segmentos L (RdRP) dos vírus AejapBV1, AejapPV1 e o fragmento do Orthomyxo-like vírus, todos os 12 vírus apresentaram um padrão de offset com alta frequência de sobreposição de 10 nt. Avaliando os gráficos de logo, todos os vírus -ssRNA segmentados possuem um padrão de enriquecimento de 5'U nos reads que alinham na fita antisenso e A na posição das dos reads que alinham na fita senso. O enriquecimento de A10 é variável entre os segmentos desses vírus. Dos vírus não-segmentados -ssRNA. As reads alinhadas ao fragmento do vírus Orthomyxo-like vírus não apresenta enriquecimento 5'U antisenso ou 10A senso. AAV não parece ter um enriquecimento evidente de 10A. CFAV, o único +ssRNA com evidência de produção de piRNAs também não possui enriquecimento 10A. A cobertura de reads de 24-30nt para os contigs representativos apresenta uma assimetria entre as fitas senso e antisenso que distingue essa fração dos pequenos RNAs correspondentes aos siRNAs com 21 nt.

-ssRNA segmented viruses (Class V)



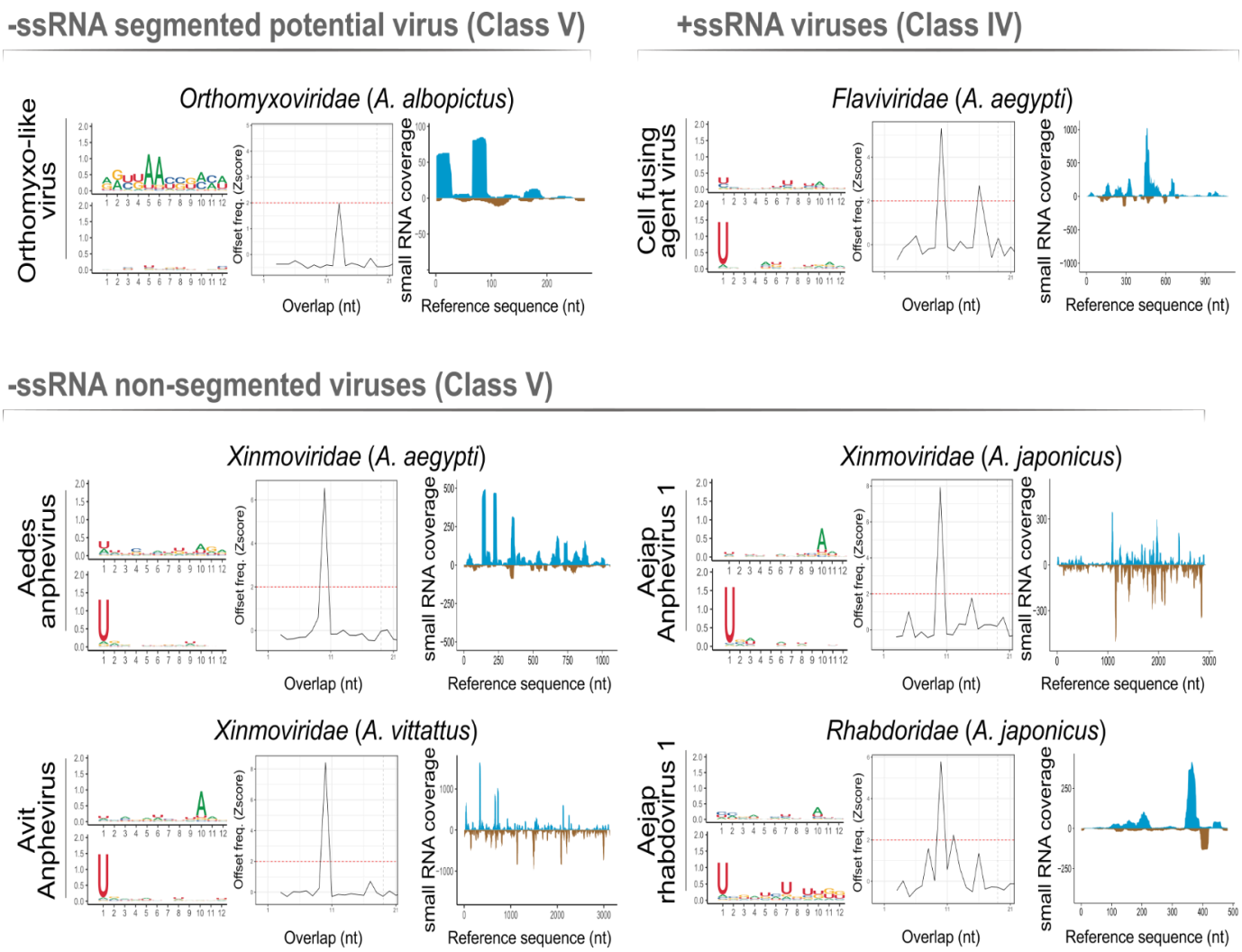


Figura 19 – Análise dos piRNAs virais. 11 vírus de genoma -ssRNA (Baltimore V) e o CFAV de genoma +ssRNA apresentaram perfis de pequenos RNAa que evidenciam a produção de piRNAs virais e foram analisados para investigarmos a presença de características típicas dessa classe de pequenos RNAs. Para cada vírus (ou segmento viral), a primeira figura (esquerda) mostra um gráfico de logo com valores de bitscore para a preferência de nucleotídeos (direção 5'-3') encontrados nas reads de 24-30nt alinhadas aos contigs na fita senso (logo superior) e antisenso (logo inferior). As figuras do meio são histogramas com a frequência normalizada dos tamanhos das sobreposições dos reads de 24-30nt alinhados às fitas senso e antisenso. A última figura (direita), mostra a cobertura de reads de 24-30nt da fita senso (azul) e antisenso (marrom) dos contigs representativos.

Na **Figura 20**, foi utilizada uma abordagem de agrupamento hierárquico para mostrar as relações entre os diferentes vírus representados pelas quantificações de cada tamanho de pequenos RNAs de 15 a 35nt alinhados aos contigs representativos, separados por polaridade de fira e normalizados por Z-score. O primeiro padrão evidente é o grande agrupamento formado pela maioria

Figura 20 – Análise integrada do perfil de pequenos RNAs do viroma dos mosquitos vetores. No heatmap são mostrados os valores de Z-score da quantidade de pequenos RNAs de 15-35nt separados por fita (senso e antisenso). Os valores de Z-score representando as sequências virais foram usados para clusterizações hierárquicas de cada vírus e seus segmentos genômicos usando método “ward.D2” e distância “camberra”. Os vírus foram destacados de acordo com suas classificações de Baltimore: dsRNA em vermelho, +ssRNA em amarelo e -ssRNA em laranja.

dos segmentos genômicos dos vírus de genoma -ssRNA, incluindo o CFAV (+ssRNA). Também pode se observar um agrupamento em que todos os arbovírus são localizados proximamente. Com exceção da maioria dos vírus segmentados -ssRNA nos quais os segmentos genômicos apresentam perfil de pequenos RNAs com maior semelhança com seus correspondentes em outro vírus similar do que entre os segmentos do mesmo vírus, a representação dos atributos de pequenos RNAs apresenta resolução o suficiente para inferir que contigs de diferentes segmentos podem pertencer ao mesmo vírus pela proximidade em que essas sequências se localizam nos agrupamentos e pela observação da quantificação de atributos pelo heatmap.

5.2 - Estudo de caso: O viroma do mosquito invasor *Ae. japonicus* na Europa

Nos dedicamos a caracterização detalhada do viroma de um subgrupo de quatro bibliotecas de pequenos RNAs construídas a partir de duas amostras da França e duas da Holanda do mosquito *Ae. japonicus* (**Figura 21A**), um vetor competente para a transmissão de arbovírus em ascensão na Europa e América do Norte e com potencial de invadir a América do Sul (OUTAMMASSINE; ZOUHAIR; LOQMAN, 2022). Em parceria com o Laboratório do Professor Pijlman Gorben (Universidade de Wageningen, Holanda), tivemos acesso a amostras para estudos de prevalência e validações dos vírus encontrados nessa espécie.

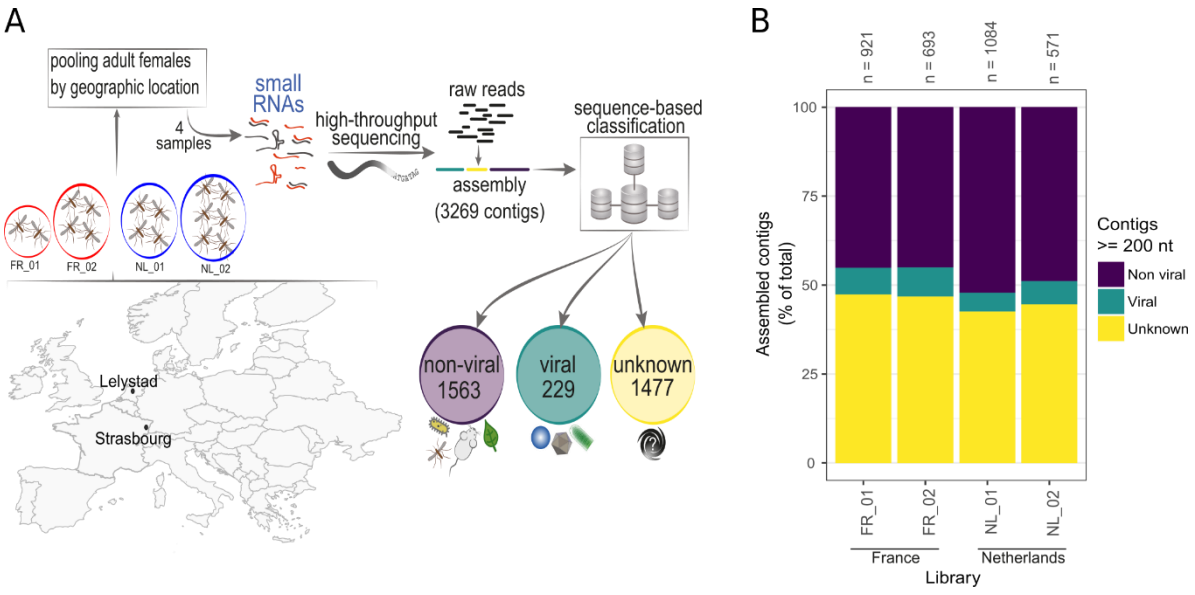


Figura 21 - Análise do viroma *de Ae. japonicus* usando uma abordagem metagenômica baseada em pequenos RNAs. **A.** Mapa da Europa indicando os locais de coleta de mosquitos: Estrasburgo, França (vermelho) e Lelystad, Holanda (azul). As amostras com o número de mosquitos de Estrasburgo são indicadas dentro dos círculos vermelhos (FR_01 dois mosquitos e FR_02 quatro mosquitos) e de Lelystad dentro dos círculos azuis (NL_01 quatro mosquitos e NL_02 seis mosquitos). Os mosquitos capturados foram previamente identificados morfologicamente por espécie. As amostras foram usadas para preparar bibliotecas de pequenos RNAs para sequenciamento. Os resultados do sequenciamento foram analisados usando nossa pipeline metagenômica descrita previamente. Os contigs montados foram classificados em sequências "não-virais", "virais" e "desconhecidas" com base na similaridade de sequência em relação a bancos de dados de referência. **B.** Resultados individuais de nossa análise de similaridade de sequência para cada uma das quatro bibliotecas de pequenos RNAs. O número total de contigs maiores ou iguais a 200 nt (n) e a proporção de contigs "não-virais", "virais" e "desconhecidos" são mostrados

Foram montados 3269 contigs maiores que 200 nt a partir das quatro bibliotecas individuais. Um resumo dos resultados das montagens é mostrado na **Tabela 6**. Com base nos resultados de similaridade de sequência, os contigs foram classificados em 229 sequências virais, 1563 não virais e 1477 sequências desconhecidas (**Figura 21A**; **Tabela 7**). As proporções de cada classe por biblioteca são mostradas na **Figura 21B**. Os 229 contigs inicialmente classificados como virais passaram por curadoria de sequência e perfis de pequenos RNAs e posteriormente classificados em 93 virais e 136 EVEs. Após remoção de redundância com o programa CDHit e retirada dos contigs oriundos dos arbovírus infectando os mosquitos artificialmente nas bibliotecas da Holanda, obtivemos um total de 39 sequências virais representativas.

Tabela 6 – Estatísticas de montagem dos contigs montados nas bibliotecas de *Ae. japonicus.*

SRA	ID	Contigs > 200nt	N50	Tamnho médio(nt)	Mediana	Desvio Padrão	Maior Contig (nt)	Contigs > 1K	Total de bases em contigs > 1k
SRR9131261	NL_01	1.084	316	333,00	265	332,46	8.702	19	36.916
SRR9131262	NL_02	571	291	319,00	255	325,89	6.498	9	18.239
SRR17146598	FR_01	921	320	330,00	264	208,38	2.253	17	25.240
SRR17146597	FR_02	693	279	297,00	255	143,59	1.609	8	9.991

Para associar diferentes segmentos que podem pertencer ao mesmo vírus, avaliamos a coocorrência dos contigs virais com base nas contagens de pequenos RNAs nas quatro bibliotecas, como anteriormente realizado para todas as 122 bibliotecas (**Figura 13)**. Nessa análise das bibliotecas de *Ae. japonicus*, incluímos 22 contigs inicialmente classificados como desconhecidos com tamanhos maiores que

500 nt e que apresentaram perfis de pequenos RNAs semelhante a sequências virais. A aplicação de um algoritmo de agrupamentos hierárquicos permitiu a distinção de diversos agrupamentos de contigs coocorrentes **(Figura 22)**. Contigs coocorrentes foram considerados fragmentos prováveis do mesmo vírus, especialmente quando compartilhavam um perfil de tamanho de pequenos RNA semelhante (**Figura 22**). Além disso, os resultados de similaridade de sequência com as mesmas referências virais guiaram a determinação de contigs pertencentes ao mesmo vírus. Com base nesses dois critérios, coocorrência e referência mais próxima no banco de dados, definimos 5 grupos de contigs virais.

Tabela 7 – Classificação dos contigs montados nas bibliotecas de *Ae. japonicus.*

SRA	ID	contigs > 200nt	Não-viral	Viral	Desconhecidos	EVEs	Desconhecidos com siRNA
SRR9131261	NL_01	1.084	564	56	464	22	111
SRR9131262	NL_02	571	270	52	249	18	78
SRR17146598	FR_01	921	417	66	438	33	44
SRR17146597	FR_02	693	312	55	326	20	15

O primeiro grupo, em vermelho, formou um único agrupamento de contigs contendo quatro contigs com alta similaridade de sequência com o AejapNV1 e um contig desconhecido (NL_02_Contig12989_12988) **(Figura 22)**. O segundo grupo, em azul, inclui contigs com similaridade significativa com diferentes segmentos do *Narangue virus*, um bunyavírus, todos contigs pertencem ao mesmo agrupamento com exceção de um contig com alta similaridade com uma glicoproteína de bunyavírus (FR_01_Contig9275_9274). Outro grande grupo de sequências, em verde, foi dividido em dois aglomerados principais e os contigs possuem alta similaridade com o *Aedes anphevirus*. Dois outros grupos, um contendo dois contigs com similaridade significativa de sequência com um Bunyavírus (em roxo) e um contig com similaridade a nível de aa com um Totivírus (em amarelo), se agruparam proximamente. Por último, dois contigs com alta similaridade de sequência com um Rhabdovírus também formaram um grupo definido (em marrom). Os outros potenciais agrupamentos não foram destacados porque são compostos por contigs desconhecidos que não puderam ser claramente associados a um vírus específico.

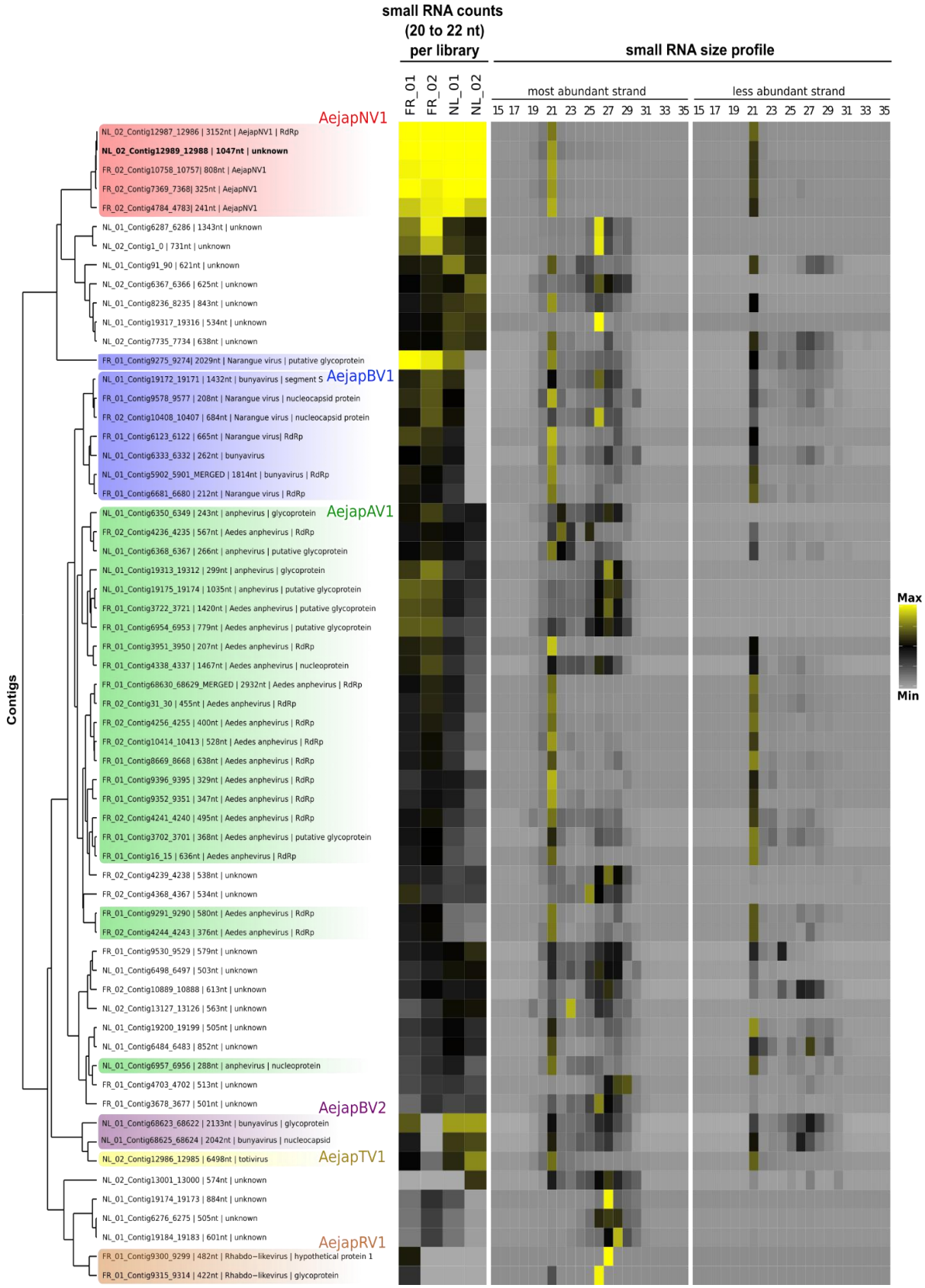


Figura 22 - Coocorrência de contigs virais e desconhecidos. Agrupamento hierárquico de contigs "virais" e "desconhecidos" montados a partir de pequenos RNAs de *Ae. japonicus*, baseado na distância euclidiana dos valores RPKM (método UPGMA). Contigs do mesmo vírus foram coloridos de forma uniforme. O mapa de calor à esquerda mostra a abundância de pequenos RNAs (20-22 nt) por contig curado (Log_2 RPKM, máx: 10, mín: 0). À direita, o mapa exibe os valores de *Z-score* para pequenos RNAs (15-35 nt) por polaridade de fita (máx: 7, mín: -1).

Para caracterizar precisamente os vírus representados pelos contigs dos agrupamentos destacados na **Figura 22**, concentramo-nos em sequências que codificam polimerases virais. Foi possível identificar contigs que codificam polimerases virais em quatro agrupamentos (**Tabela 4**). O agrupamento vermelho (**Figura 22**) possui sequências com alta similaridade a nível de nucleotídeo com o vírus AejapNV1 (gênero *Narnavirus*, família *Narnaviridae*), que havia sido identificado anteriormente como um possível ISV (ABBO et al., 2020). De acordo com a análise filogenética (**Figura 23**), este vírus agrupou-se com outros Narnavírus do clado Alphanarnavirus que são associados a mosquitos e possuem estruturas de ORFs ambigramáticas com longas ORFs reversas (antisense) sobrepondo as fORFs (fita senso) (DINAN et al., 2020).

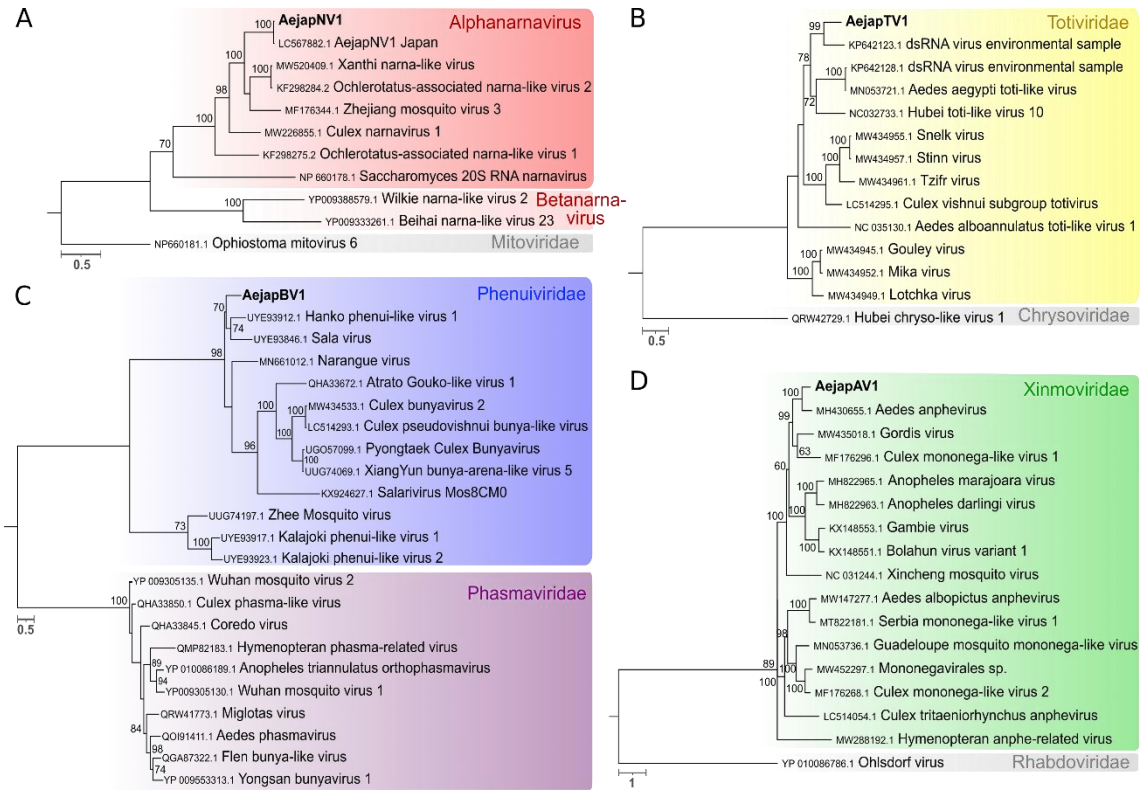


Figura 23- Filogenia dos vírus identificados em mosquitos *Ae. japonicus*. As árvores filogenéticas foram geradas usando alinhamentos múltiplos de sequências de sequências de aminoácidos da proteína RdRp. As árvores foram inferidas usando o método da Máxima Verossimilhança. A árvore com a verossimilhança mais alta (*log likelihood*) é mostrada para cada vírus. Número de sítios conservados e os modelos de substituição usados para cada árvore: **A.** AejapNV1, 1446 sítios, LG+G+F; **B.** AejapTV1, 1269 sítios, LG+G; **C.** AejapBV1, 616 sítios, LG+G+F; **D.** AejapAV1, 1188 sítios, LG+G+I+F. Os valores de bootstrap dos nós foram calculados com 1000 replicatas e são mostrados próximos a cada clado. Valores de bootstrap menores que 60 foram omitidos. As árvores foram enraizadas no ponto médio (*midpoint-rooted*), e sequências de RdRp de diferentes famílias virais foram incluídas nos alinhamentos como grupos externos. As árvores foram plotadas em escala e os comprimentos

dos ramos representam o número esperado de substituições por sítio de aminoácido. Os números de acesso para as sequências de nucleotídeos a partir das quais as sequências de proteínas correspondentes foram derivadas ou as sequências de proteínas diretas são mostrados com os nomes dos vírus. Os vírus identificados neste estudo estão destacados em negrito.

Os demais contigs contendo sequências de polimerase viral apresentaram similaridade de sequência significativa, porém baixa, com sequências referências de Totivírus, Anphenvírus e Bunyavírus apenas a nível de aminoácidos (*e-value* < 1e-3) (**Tabela 4**). Análises filogenéticas evidenciam que esses são provavelmente novos vírus pertencentes às famílias *Totiviridae*, *Xinmoviridae* e *Phenuiviridae*, denominados *Ae. japonicus Totivirus* 1 (AejapTV1), *Ae. japonicus Anphenvírus* 1 (AejapAV1) e *Ae. japonicus bunyavirus* 1 (AejapBV1), respectivamente (**Figura 23**). Na **Figura 24**, são mostradas as prováveis organizações genômicas do vírus para os quais não obtivemos montagens completas. Os três vírus novos possuem proximidade filogenética com ISVs conhecidos (**Figura 23**; **Tabela 4**), porém uma classificação precisa requer experimentos para evidenciar que esses vírus não são capazes de infectar vertebrados. Dos quatro vírus identificados com sequências de polimerases virais, um conhecido e três novos, todos possuem genomas de RNA, seja de fita simples (de polaridade positiva ou negativa) ou fita dupla (**Tabela 4**; **Figura17**). Não pudemos identificar sequências codificadoras de polimerases entre os contigs pertencentes a outros dois vírus putativos, que nomeamos *Ae. japonicus Bunyavirus* 2 (AejapBV2) e *Ae. japonicus Rhabdovirus* 1 (AejapRV1) com base nos alinhamentos com as referências mais próxima (**Tabela 4; Figura 24 B,D**).

Os contigs de AejapBV2 agruparam-se com a sequência do genoma completo do AejapTV1 com base na coocorrência nas quatro bibliotecas (**Figura 22**), mas não há evidências adicionais que associem essas sequências ao mesmo vírus. Pelo contrário, nossa análise geral, incluindo perfis de tamanho de pequenos RNAs (**Tabela 4**, **Figura 18,19,22**) e comparação de similaridade de sequência, indica que AejapBV2 e AejapTV1 são vírus completamente distintos. Embora não tenham sido encontrados contigs correspondentes a um suposto segmento L (RdRP) pertencente ao AejapBV2 em nossas bibliotecas, montamos com sucesso os segmentos genômicos M e S inteiros (**Figura 24B**).

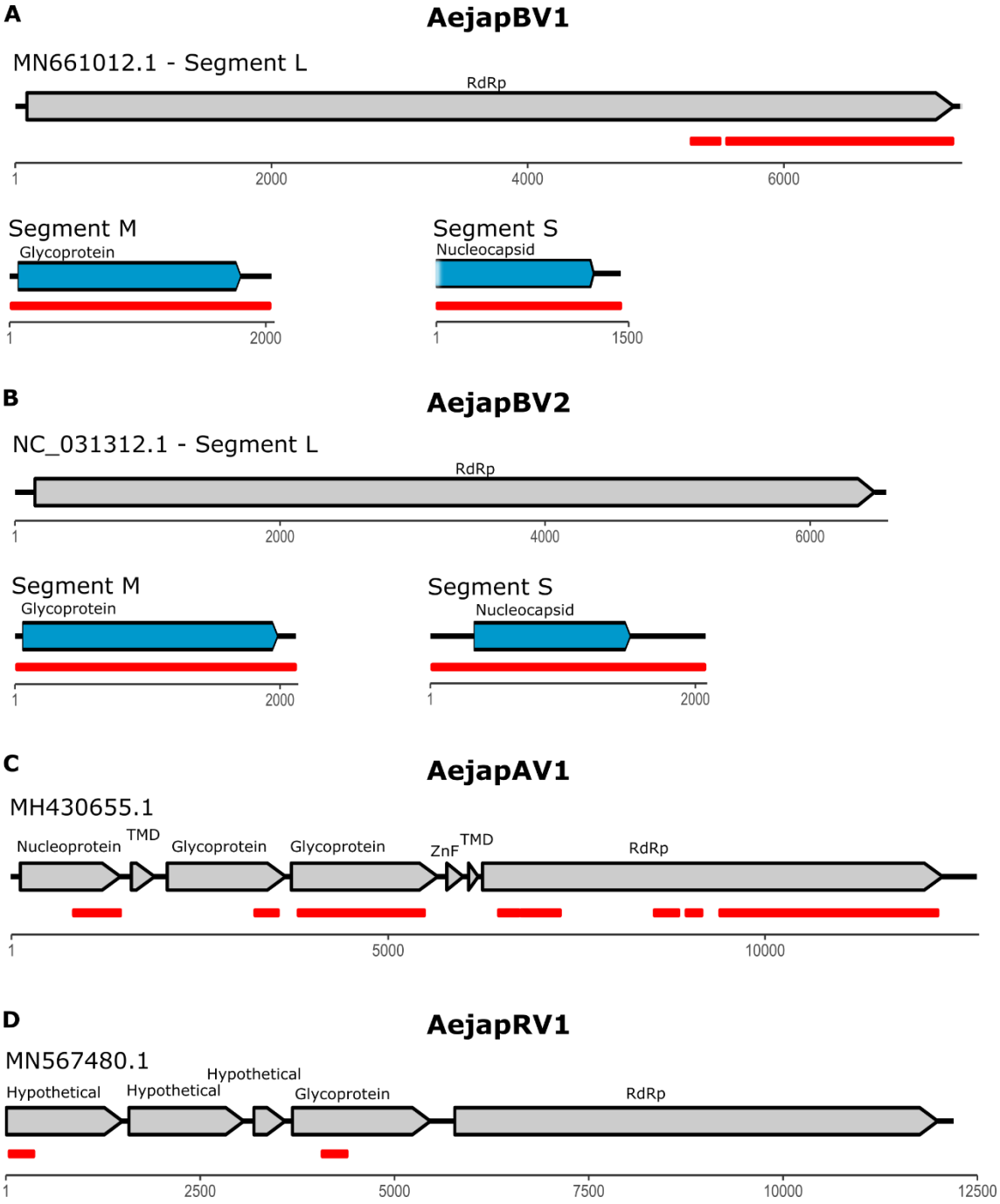


Figura 24 - Organização genômica de vírus parcialmente montados. Setas cinza representam ORFs da sequência viral de referência mais próxima no GenBank. Linhas vermelhas indicam regiões genômicas de referência viral cobertas por nossos contigs montados. Setas azuis representam ORFs de segmentos virais completamente montados neste trabalho. Linhas pretas indicam regiões não traduzidas. **A.** AejapBV1. A ausência da 5' UTR para o segmento S é representada como uma região de cor desvanecida. Apesar da falta de uma 5' UTR e de um códon de início, o tamanho total do ORF do segmento S é semelhante ao de sua sequência mais próxima no GenBank (QHA33859.1). **B.** AejapBV2, **C.** AejapAV1, **D.** AejapRV1.

A análise filogenética utilizando sequências de aminoácidos mostra que os segmentos do AejapBV2 codificando a glicoproteína e o nucleocapsídeo estão

distantes das sequências equivalentes de AejapBV1 (**Figura 25**). As análises filogenéticas combinadas aos resultados de similaridade de sequência do nucleocapsídeo e da glicoproteína de AejapBV2 indicam que este vírus pertence à família *Phasmaviridae*, enquanto AejapBV1, pertence à família *Phenuiviridae* (**Tabela 4; Figura 24,25**). Na tentativa de encontrar reads do segmento genômico codificador de uma potencial RdRp pertencente ao AejapBV2 em nossas bibliotecas, mapeamos os reads de cada biblioteca contra a sequência do segmento L do vírus no qual os segmentos M e S eram mais semelhantes ao AejapBV2 (**Tabela 4**), *Wuhan mosquito virus* 2 (GenBank: NC_031312.1) (**Figura 25**). Mesmo permitindo múltiplos missmatches, não observamos uma quantidade expressiva de alinhamentos nem uma cobertura contínua do segmento L desse vírus (**Figura 26**).

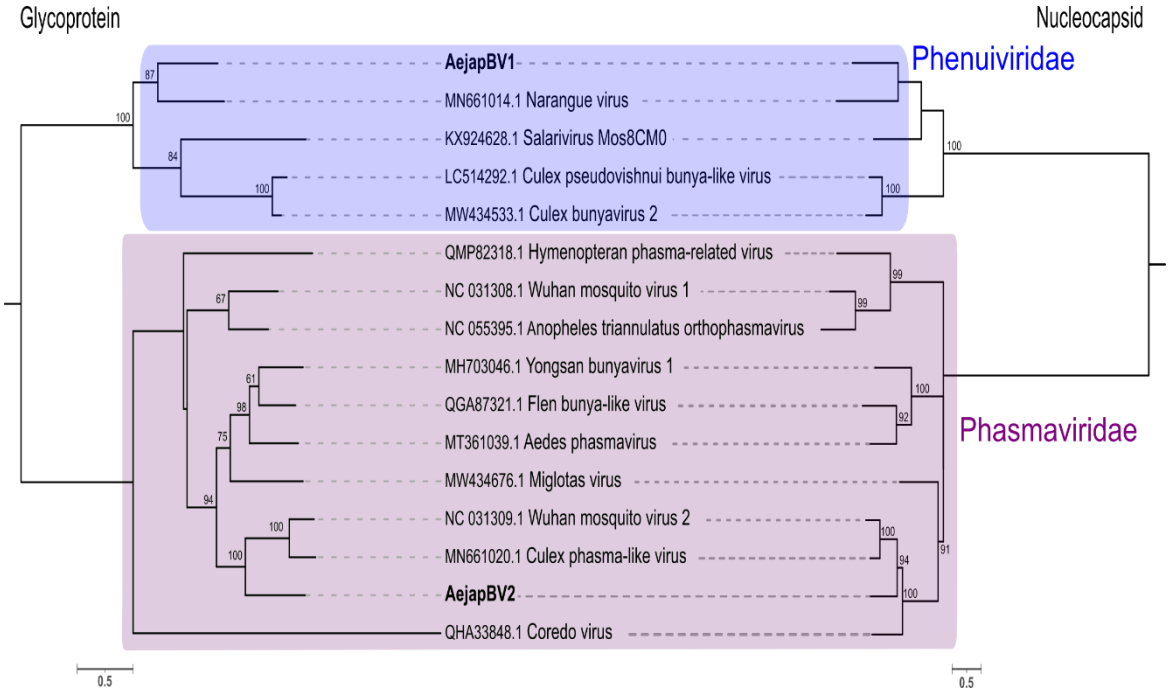


Figura 25 – Análise filogenética da glicoproteína e nucleocapsídeo de AejapBV1 e AejapBV2. Árvores filogenéticas foram geradas usando as sequências de aminoácidos da glicoproteína e do nucleocapsídeo. As árvores foram inferidas utilizando o método da Máxima Verossimilhança. As árvores com as maiores verossimilhança (*log likelihood*) estão mostradas. Número de sítios conservados e os modelos de substituição usados para cada árvore: glicoproteína, 796 sítios, WAG+G+F, e nucleocapsídeo, 573 sítios, LG+G. Os valores de bootstrap dos nós foram calculados com 1000 replicatas e são mostrados próximos a cada clado. Valores de bootstrap menores que 60 foram omitidos. As árvores foram enraizadas no ponto médio (*midpoint-rooted*), e sequências de diferentes famílias virais foram incluídas nos alinhamentos. As árvores foram plotadas em escala e o comprimento dos ramos representa o número esperado de substituições por sítio de aminoácido. Números de acesso para as sequências de nucleotídeos das quais as sequências de proteínas correspondentes foram derivadas ou as sequências de proteínas diretas são mostradas com os nomes dos vírus. Vírus identificados neste estudo estão destacados em negrito.

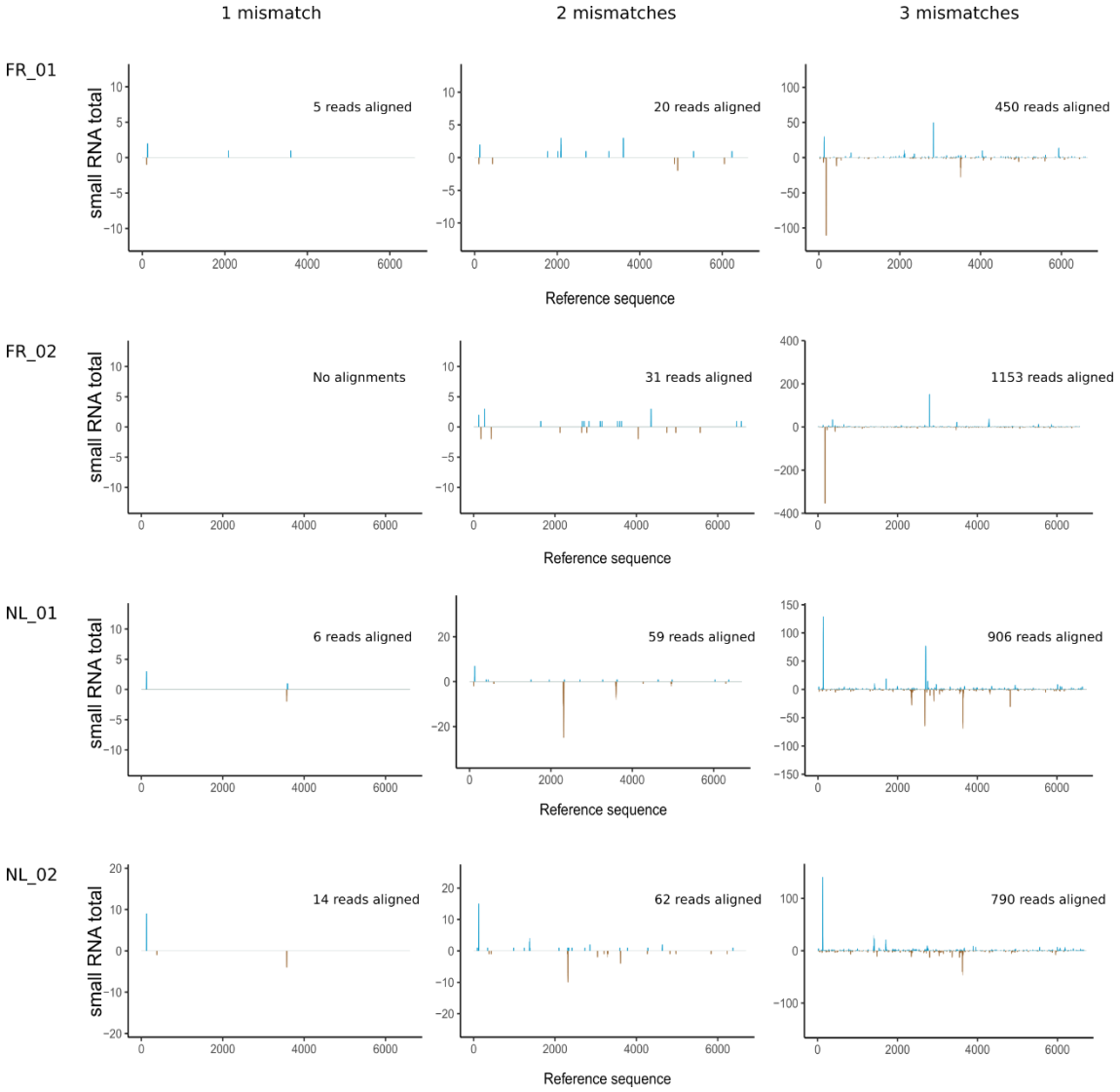


Figura 26 - Cobertura de pequenos RNAs do segmento L do vírus *Wuhan mosquito* 2. Reads totais de cada biblioteca (linhas) foram alinhadas ao segmento L do vírus *Wuhan mosquito* 2 (NC_031312.1), uma sequência potencialmente homóloga ao segmento L ausente identificado no nosso AejapBV2. Para confirmar a ausência de um segmento L altamente divergente inexplicavelmente não montado, alinhamos cada biblioteca de pequenos RNAs permitindo de um a três mismatches por read (colunas), variando o parâmetro -v do programa Bowtie. Cada painel representa a cobertura e as reads totais alinhadas para a combinação de uma biblioteca e o número máximo de mismatches permitidos por read. As linhas azuis indicam a cobertura de reads na fita senso e as linhas marrons na fita antisenso. Somente quando foram permitidos três mismatches, uma quantidade expressiva de reads foram alinhadas à referência, porém sem sinal de cobertura contínua. Esses resultados indicam alinhamentos provavelmente espúrios.

5.2.1 - Perfil de pequenos RNAs e organização genômica de AejapNV1

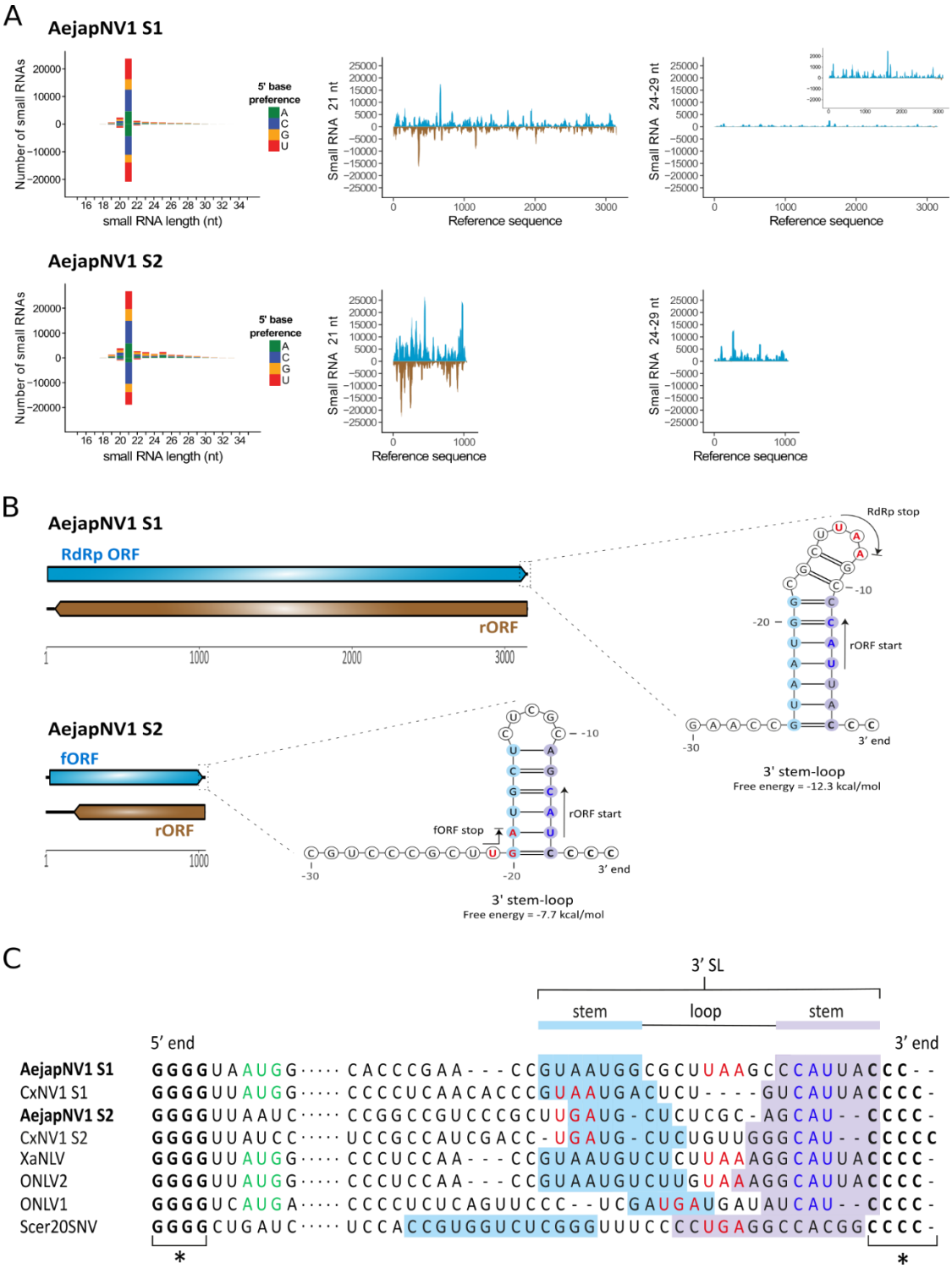


Figura 27 - Perfis de pequenos RNAs e organização do genoma de AejapNV1. A. Esquerda: distribuição de tamanho e preferência da base 5' de pequenos RNAs derivados de AejapNV1 S1 e S2. Meio e direita: cobertura de pequenos RNAs de tamanho 21 e 24-29 nt alinhados em S1 e S2. Reads alinhadas na fita senso são mostradas em azul e na fita antisenso em marrom. **B.** Organização do genoma de AejapNV1. Estratégia de codificação ambigramática de S1 e S2. ORF da RdRp de S1 e ORF senso de S2 (fORF) são mostradas em azul, enquanto ORFs antisenso (rORFs) mostradas em marrom. Regiões não traduzidas (UTRs) são indicadas por linhas pretas. Estruturas de stem-loop previstas na3' UTR do das fitas senso também são mostradas para ambos os segmentos. As localizações dos códons de início e término são indicadas por setas coloridas em azul e vermelho na estrutura de RNA.

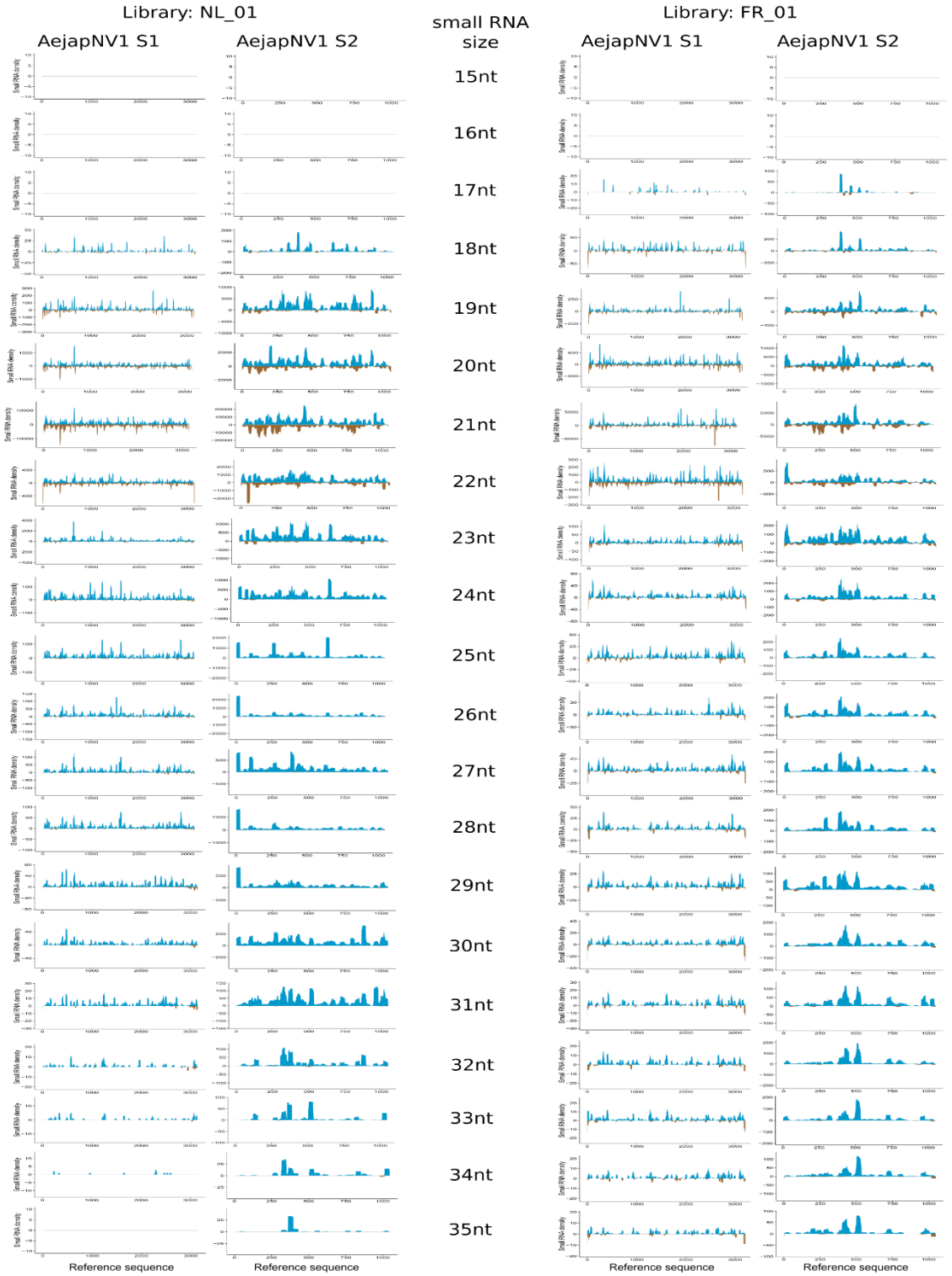
C. Alinhamento global múltiplo das sequências UTR 5’ e 3’ da fita positiva dos Narnavírus indicados. Asteriscos (*) indicam regiões repetitivas conservadas e complementares ricas em nucleotídeos G ou C no final de 5' ou 3', respectivamente. Pontos representam o restante do genoma viral omitido no alinhamento. Os códons de início das ORFs das RdRp estão em letras verdes, os códons de término do ORF de RdRp / fORF estão em letras vermelhas, e os códons de início do rORF estão em letras azuis. Os códons de início dos fORFs de AejapNV1 S2 e CxNV1 S2, assim como o códon de início do ORF de RdRp de Scer20SNV, estão localizados a mais de 10 nt a jusante do 5' e, portanto, não são mostrados no alinhamento. Os nucleotídeos envolvidos na formação do stem-loop 3' (3' SL) foram destacados com sombras azul e roxa.

O perfil e cobertura de pequenos RNAs, e a estrutura de ORFs ambigramáticas do contig desconhecido de aproximadamente 1 kb (NL_02_Contig12989_12988) (**Figura 22**) são semelhantes à sequência do genoma do AejapNV1 de cerca de 3 kb que codifica a RdRp (**Figura 22**, **Figura 27A,B**). Associadas a proximidade dessas sequências no agrupamento hierárquico (**Figura 22**) essas características indicam que este contig pertence ao vírus AejapNV1 e foi, portanto, nomeado AejapNV1 segmento 2 (S2), enquanto a sequência que codifica a RdRp agora é referida como segmento 1 (S1).

5.2.2 - Determinação das ORFs senso e antisenso de AejapNV1 S2

Como o novo segmento genômico descoberto possui ORFs ambigramáticas e não apresenta similaridade de sequência com outros vírus ou domínios proteicos conservados, utilizamos o padrão de degradação de pequenos RNAs para determinar a polaridade de fita do segmento descoberto. S1 apresenta um viés de pequenos RNAs mapeados de 24 a 29 nt para a fita positiva na qual a ORF da proteína RdRp está codificada (**Figura 27B**). Da mesma forma, S2 também mostrou preferência de pequenos RNAs mapeados de 24 a 29 nt para uma fita específica (**Figura 27A**). A análise da cobertura de pequenos RNAs do AejapNV1 S1 e S2 para cada comprimento de pequeno RNA separadamente indicou um viés de pequenos RNAs (18-35 nt de comprimento) para a mesma fita específica para cada comprimento de pequeno RNA, exceto para pequenos RNAs de 21 nt, que foram encontrados em quantidades simétricas em ambas as fitas (**Figura 28**). Este padrão assimétrico para pequenos RNAs de 18-20 nt e 22-35 nt é provavelmente causado pela degradação não específica de RNAs virais, indicando a abundância relativa e a exposição de cada fita (HAN et al., 2011). Para o AejapNV1 S1, o viés de pequenos RNAs foi de fato para a fita que codifica a RdRp, que é considerada a fita positiva (**Figura 28**). Portanto,

propomos que a fita positiva do AejapNV1 S2 é aquela para a qual a maioria dos pequenos RNAs mapeou.



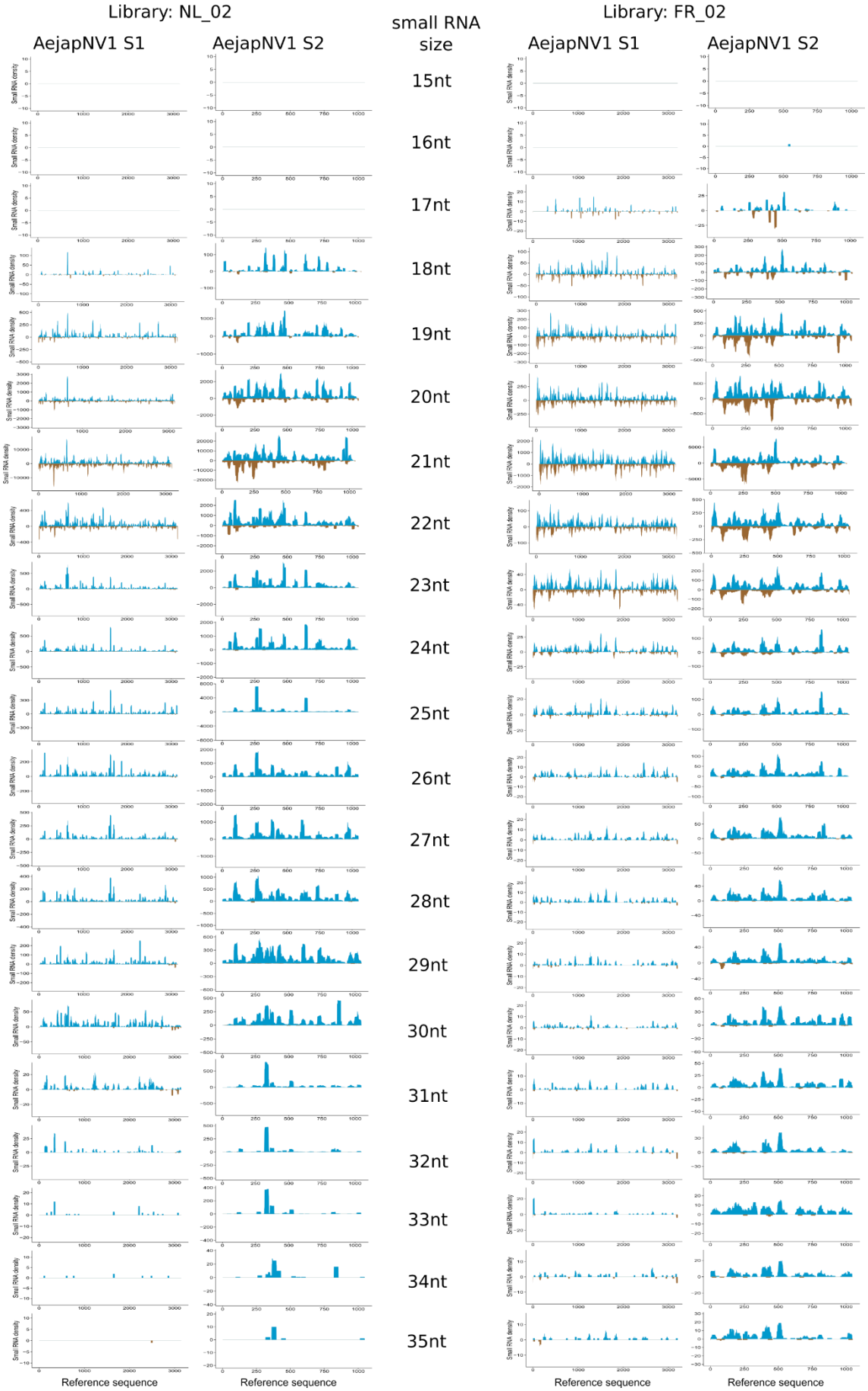

Library: NL_02
AejapNV1 S1
AejapNV1 S2
small RNA size
Library: FR_02
AejapNV1 S1
AejapNV1 S2
15nt
16nt
17nt
18nt
19nt
20nt
21nt
22nt
23nt
24nt
25nt
26nt
27nt
28nt
29nt
30nt
31nt
32nt
33nt
34nt
35nt
Small RNA density
Reference sequence

Figura 28 - Viés de cobertura de pequenos RNAs fita-específicos dos segmentos AejapNV1 S1 e S2. Reads de tamanhos de 15 a 35 nt de cada biblioteca foram alinhadas separadamente aos segmentos S1 e S2. A área azul indica a cobertura de reads da fita senso e a área marrom da fita antisenso. A orientação da sequência de S1 foi determinada com base na direção do ORF codificadora da RdRp. Observamos um viés de cobertura de pequenos RNAs na fita senso em relação a RdRp de S1, e um viés de cobertura semelhante foi observado para uma das fitas de S2, evidenciando que essa fita com maior cobertura se trata da fita senso desse segmento descoberto.

5.2.3 - Análises das regiões UTR dos segmentos genômicos de AejapNV1

Para obtermos mais evidências de que o segmento putativo S2 pertence ao vírus AejapNV1 as regiões terminais 5' e 3' de ambos os segmentos foram analisados quanto à presença de motivos de RNA conservados. As sequências terminais de AejapNV1 S1 e S2 foram comparadas com as de quatro Narnavírus ambigramáticos encontrados em mosquitos e também com o Narnavírus que infecta leveduras Scer20SNV. Para todos os vírus e segmentos analisados, regiões repetitivas de G e C estão presentes nos terminais 5' e 3' (**Figura 27C**). Com base na modelagem da estrutura de RNA, foi possível predizer estruturas de *Stem-Loops* (SL) conservada ocorresse no terminal 3' de AejapNV1 S1 e S2 (**Figura 27B,C**). Estruturas de SL conservadas semelhantes no terminal 3', diferindo em tamanho e com pares de bases covariantes na região do stem (**Figura 27C**), já foram observadas em outros Narnavirus. A presença dessas estruturas conservadas nos terminais genômicos tanto de S1 quanto de S2 de AejapNV1 evidencia que nosso método de sequenciamento de pequenos RNA foi capaz de recuperar sequências genômicas completas e permitiu a associação entre AejapNV1 S1 e o recém-descoberto S2 ao mesmo vírus.

5.2.4 - Inferências sobre as proteínas hipotéticas codificadas pelas ORFs de AejapNV1 S2

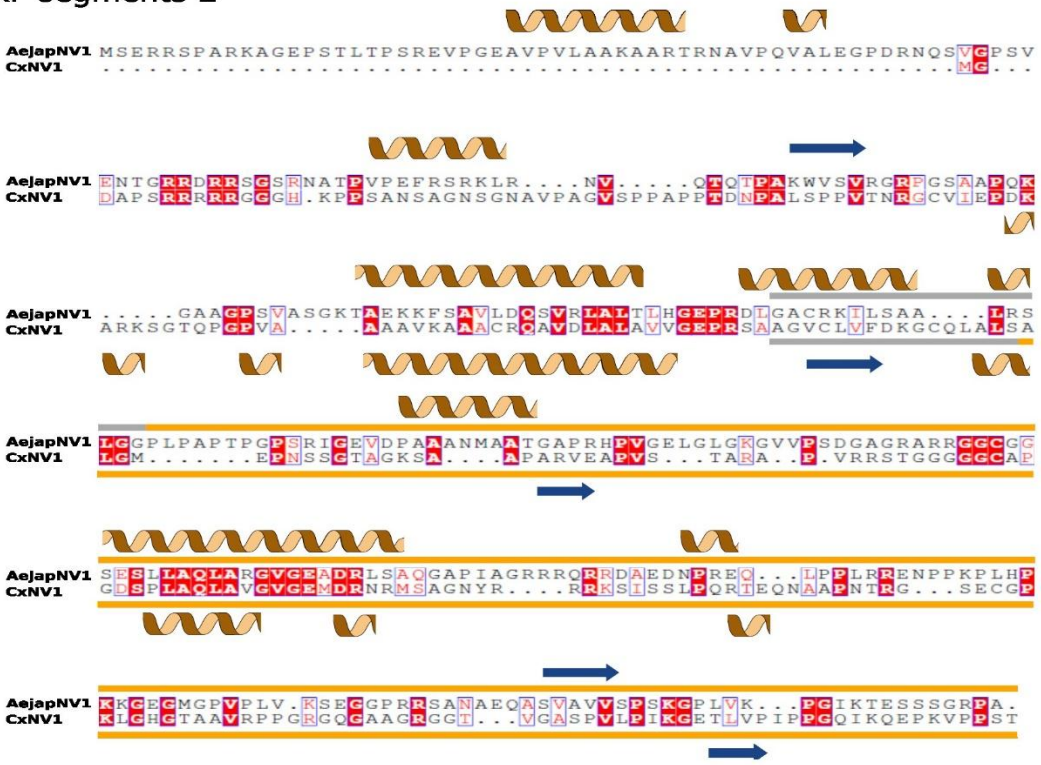
A sequência de CxNV1 S2 é a única sequência de um segundo segmento disponível publicamente de um Narnavírus que infecta mosquitos. Apesar dos terminais genômicos conservados e da mesma estratégia de codificação ambigramática (**Figura 27B,C**), os segmentos S2 de AejapNV1 e CxNV1 são extremamente divergentes, apresentando ~46% de identidade global no nível de nucleotídeos e não são detectados como significativamente similares por alinhamentos locais com a ferramenta blast. Comparamos as sequências das proteínas putativas codificados pelos segmentos S2 desses dois vírus. Ambas as

ORFs, senso e antisenso, apresentaram propriedades bioquímicas semelhantes com pontos isoelétricos altos, sugerindo uma natureza básica dessas proteínas (**Tabela 8**). No nível da sequência de aminoácidos, observamos 29,34% de identidade entre as fORFs e 26,89% entre as rORFs (**Figura 29**). As predições de estruturas secundárias mostraram regiões estruturadas e desordenadas para as proteínas codificadas pelas fORFs de ambos os vírus, com a presença de muitas α-hélices preditas em regiões potencialmente homólogas (**Figura 29A**). O mesmo padrão não foi observado para os rORFs (**Figura 29B**). Potenciais regiões transmembranas foram preditas para ambas fORFs entre os resíduos 149 a 164 de AejapNV1 S2 e resíduos 99 a 114 de CxNV1 S2 (**Figura 29A**). Usando o AlphaFold, não obtivemos predições de estrutura terciária altamente confiáveis (valores de pLDDT > 90) para as estruturas geradas a partir das fORFs e rORFs de ambos os vírus (**Figura 30**). Com valores de pLDDT > 70, o AlphaFold modelou uma região central estruturada composta por α-hélices compartilhadas pelas fORFs de ambos os vírus (**Figura 30A,C**) coerente com a organização de potenciais estruturas secundárias conservadas mostradas na **Figura 29A**. Os gráficos de PAE (*Predicted Aligned Error*) das estruturas preditas também mostram apenas um pequeno núcleo de região estruturada similar entre as fORFs (**Figura 31**).

Tabela 8 - Propriedades bioquímicas das proteínas codificadas pelas ORFs dos segmentos 2 de AejapNV1 e CxNV1.

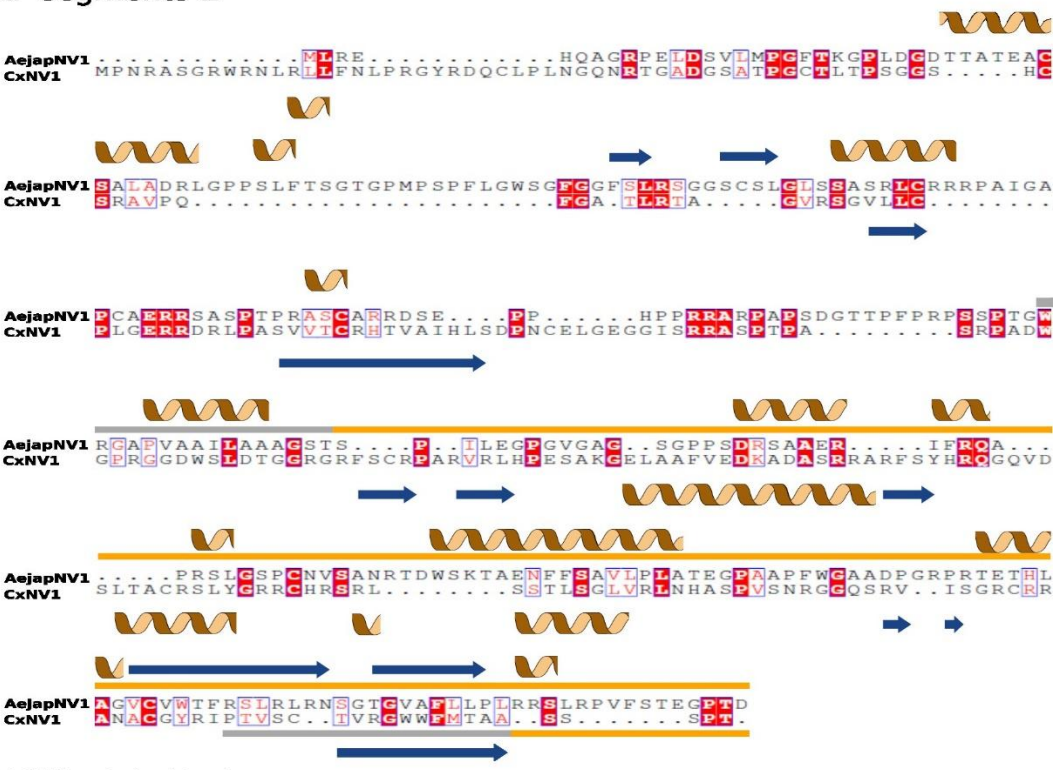
Segmento genômico	ORF	aa	Peso molecular (kDa)	pI teórico	Índice de instabilidade	Índice Alifático	GRAVY
AejapNV1 S2	fORF	333	34.44	11.39	64.35	67.48	-0,681
	rORF	285	29.65	10.91	73.88	59.40	-0,428
CxNV1 S2	fORF	268	26.58	10.70	66.95	63.84	-0,355
	rORF	268	29.01	11.35	59.74	63.40	-0,46

A fORF segments 2



*29.34% pairwise identity

B rORF segments 2



*26.89% pairwise identity

α-helix
β-sheet
Membrane Interaction
Extracellular

Figura 29 - Comparações das estruturas primárias e secundárias das proteínas dos ORFs do segmento 2 de AejapNV1 e CxNV1. Alinhamento de sequências dos ORFs fORF e rORF do segmento 2 de AejapNV1 e CxNV1 (GenBank MW226856.1) mostrados junto com as predições de estruturas secundárias e regiões de interação com membrana. **A.** São mostradas as comparações dos fORFs (senso) e em **B**. os rORFs (antisenso). Resíduos com idênticos são destacados com caixas vermelhas, e resíduos com propriedades físico-químicas de cadeia lateral semelhantes são destacados com uma caixa azul e escritos em vermelho. Acima das sequências estão representados os resultados de predição de estrutura secundária pelo PSIPRED e MEMSAT-SVM de AejapNV1 S2, e sob as sequências, a predições para as ORFs de CxNV1 S2. As regiões de α-hélice são representadas por hélices marrons e as folhas-β com setas azuis escuras. As regiões com predição de interação com membrana estão em caixas cinzas e as regiões preditas como extracelulares em amarelo

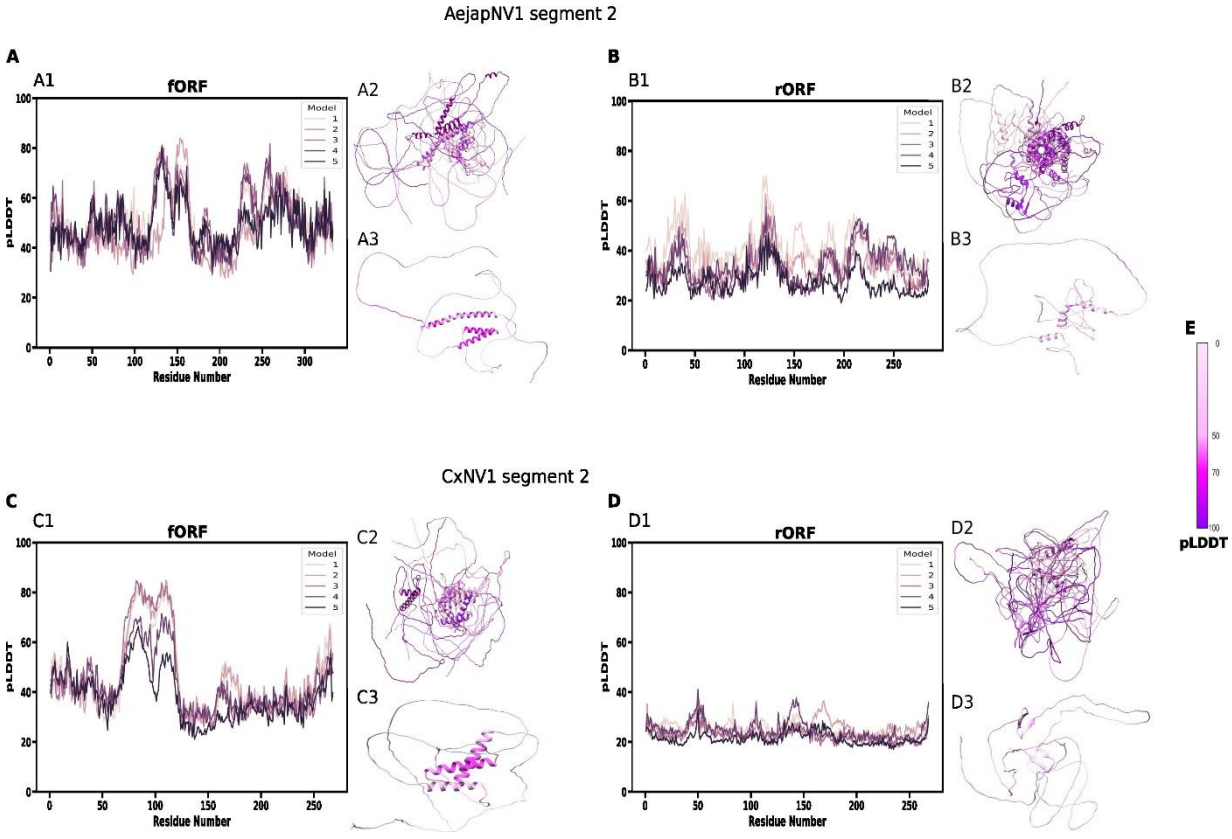


Figura 30 - Predições de estruturas proteicas terciárias para as ORFs do segmento 2 de AejapNV1 e CxNV1. Estão apresentadas a estimativa de confiança por resíduo (pLDDT) para as predições obtidas com o programa AlphaFold e os modelos calculados para os ORFs senso e antisenso dos segmentos 2 de AejapNV1 e CxNV1 (GenBank MW226856.1). Em **A.** é mostrado são mostradas as predições para a fORF do segmento 2 de AejapNV1, em **B.** da rORF do mesmo vírus; em **C.** as predições da fORF do segmento 2 de CxNV1 e em **D.** de sua rORF. **A1, B1, C1** e **D1** mostram os valores de pLDDT dos cinco modelos calculados para cada ORF. **A2, B2, C2** e **D2** mostram os cinco modelos de saída preditos superpostos. **A3, B3, C3** e **D3** mostram o modelo com o maior valor de confiança para cada ORF, com valores de pLDDT renderizados em sua estrutura terciária. A escala de cores de pLDDT está representada em (E). Todos os modelos foram representados no gráfico e na superposição estrutural usando a paleta de cores púrpura, variando de roxo claro (modelo 1) a roxo escuro (modelo 5). A escala de cores de pLDDT podem ser interpretadas como: confiança muito baixa (pLDDT < 50), confiança baixa (70 > pLDDT > 50), confiante (90 > pLDDT > 70) e confiança muito alta (pLDDT > 90).

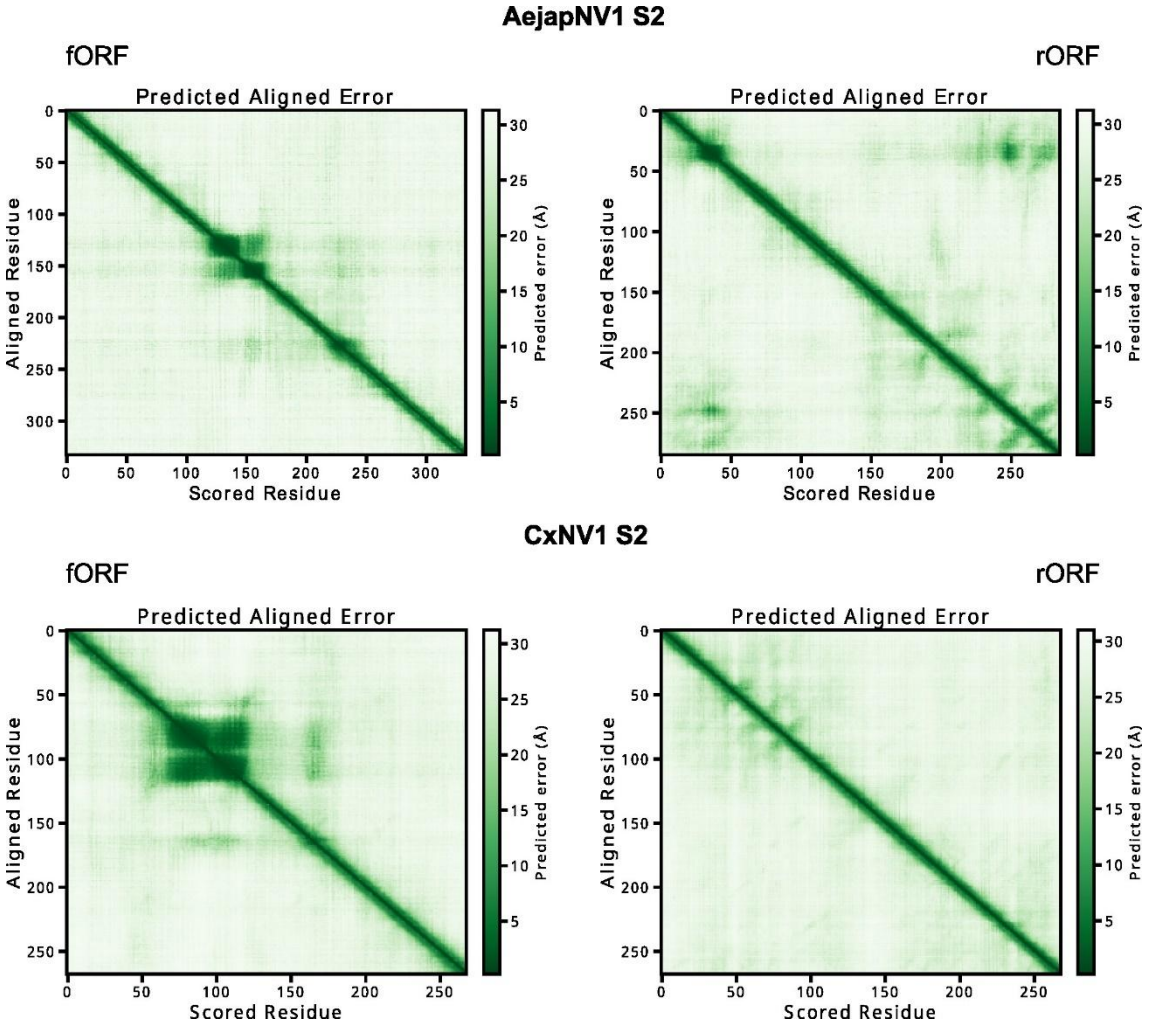


Figura 31 – Valores de PAE (*Predicted Aligned Error*) para os resíduos das estruturas proteicas preditas com Alphafold para as ORFs dos segmentos 2 de AejapNV1 e CxNV1. Os valores PAE medem a confiança nas posições relativas e orientações de partes das estruturas preditas. Foram calculados os PAEs das predições estruturais das ORFs senso e antisenso dos segmentos 2 de AejaNV1 (gráficos superiores) e CxNV1 (gráficos inferiores). A escala em tom de verde a direita de cada gráfico representa o erro predito do alinhamento dos resíduos em Ångström (Å).

5.2.5 - Organização genômica do *Aedes japonicus Totivirus* 1

Para o vírus AejapTV1 (*Totiviridae*), montamos um único contig de ~6.5 Kb (**Tabela 4**; **Figura 22**) correspondente ao genoma completo desse vírus, contendo duas ORFs dos genes que codificam as proteínas do Capsídeo e RdRP (**Figura 32**), coerente com a estrutura genômica de vírus da mesma *frame shifiting* para a tradução de ORFs em janelas de leituras diferentes. Investigamos se essa estratégia poderia ser potencialmente empregada por AejapTV1. Com base na modelagem de estrutura de RNA, uma área putativa de mudança de janela de leitura -1 foi descoberta no final do ORF do Capsideo (**Figura 32**). Um sítio deslizante (GGAAAAC) presente logo

antes do códon de parada do ORF do Capsideo corresponde ao motivo de consenso heptamérico típico para mudanças de quadro ribossômico -1 e representa a área onde o ribossomo volta para a janela leitura (BEKAERT et al., 2003). Logo após o sítio deslizante, foi encontrada uma região de espaçamento com 5 nucleotídeos de comprimento. Esta região foi seguida por uma área altamente estruturada consistindo em um pseudonó de três hastes, que se espera que seja responsável por pausar e realocar o ribossomo.

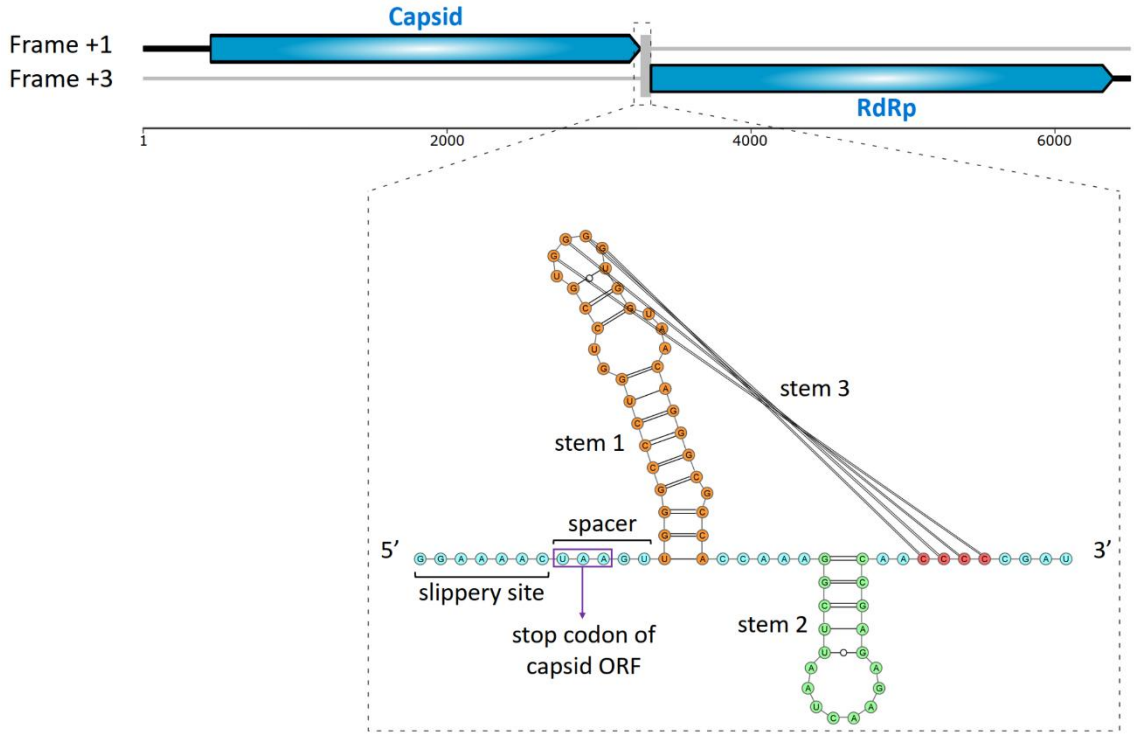


Figura 32 – Organização genômica do AejapTV1. Regiões UTR são indicadas por linhas pretas. ORFs das proteínas do Capsídeo e da RdRp são mostrados em azul. As ORFs são codificadas em janelas de leitura (frames) distintos, e uma área putativa de mudança de quadro ribossômico -1 foi observada entre as duas ORFs. Essa área consiste em um heptâmero deslizante, uma região espaçadora e uma pseudonó de três hastes com valor energia livre das estruturas preditas de -33,97 kcal/mol.

5.3 - Análise do EVEroma

Para análise do EVEroma, as 122 biblioteca passaram por uma segunda vez pelo pipeline do viroma, porém sem a remoção das reads que alinharam no genoma dos hospedeiros. Foram montados 174576 contigs maiores que 200nt que foram classificados por similaridade de sequência em 9117 virais, 106625 não virais e 58834 desconhecidos (**Tabela Suplementar 2**).

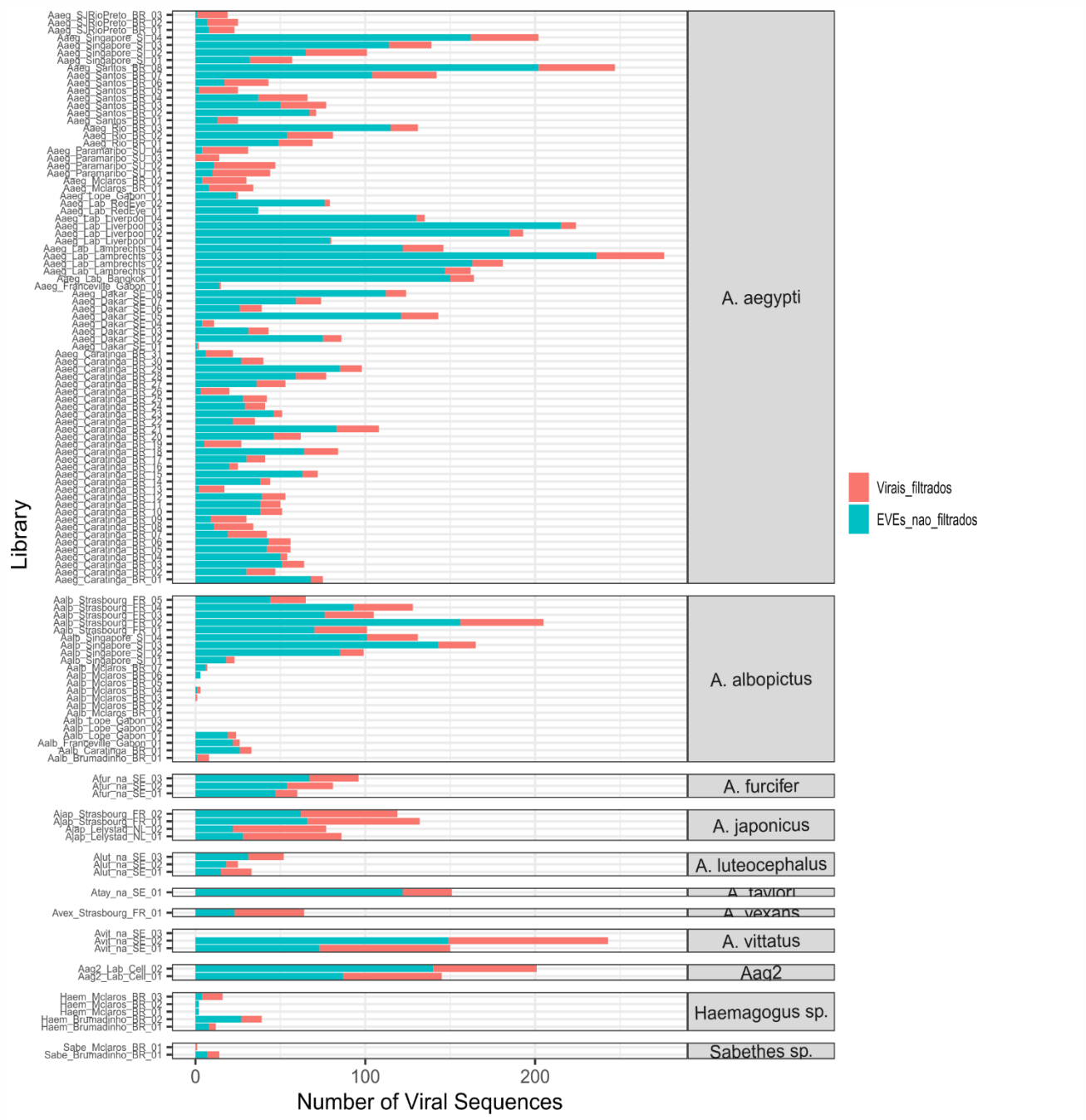


Figura 33 – Proporções de contigs virais e de EVEs montados por bibliotecas. A porporção de contigs de cada biblioteca é mostrada separadamente por espécies.

Após inspeção dos perfis de pequenos RNAs virais das sequências classificadas como virais (Figura 9), 6191 contigs foram classificados como potenciais EVEs. Dentre esses, apenas 744 tiveram similaridade significativa com sequências virais conhecidas a nível de nt, os outros 5447 contigs tiveram alinhamentos estatisticamente significativos com sequências virais apenam em nível de aa como resultado de alinhamentos no modo blastx da ferramenta *Diamond.* Dos 6191 contigs de EVEs, Na **Figura 33** é mostrada a proporção de contigs de EVEs obtidos a partir

da montagem de reads não filtrados nos genomas de referência de mosquitos junto com a proporção de contigs virais montados a partir de reads filtrados (alta confiança de que foram montados com reads de sequências virais exógenas) por biblioteca. Dentre as 6191 sequências de EVEs, 1783 possuem similaridade de sequência com genes de capsídeos virais, 1612 com glicoproteínas virais, 1534 com genes de polimerases virais, 640 com proteínas hipotéticas virais e 622 com outros genes virais. Apesar das proporções de sequências de capsídeos, glicoproteínas e polimerases parecem similares, 582 sequências de polimerases são pequenos fragmentos dos retrovírus *Aedes aegypti To vírus 1* e 2 que enviesaram a proporção de sequências dessa classe.

Após remoção das porções de sequências não virais com os resultados de alinhamentos locais e remoção da com o programa CDHIT, obtivemos 1736 sequências representativos de EVEs. Estatísticas dos contigs de EVEs pré e pós processamento são mostradas na **Tabela 9**.

Tabela 9 - Estatísticas dos contigs de EVEs pré e pós processamento e remoção de redundâncias.

Contigs	Nro. de contigs	N50	Tamanho médio (nt)	Mediana	Desvio Padrão	Maior contig	Contigs > 1K	Bases em contigs > 1k
EVE	6191	362	355.91	299	179.48	2775	76	96161
EVE porção viral	6191	255	243.81	213	130.35	1524	20	23297
EVE porção viral não redundante	1736	297	275.22	228	167.57	1524	15	1515

As sequências não redundantes foram utilizadas para análise de coocorrência na qual reads de 24-30 nt (piRNAs) de todas as bibliotecas foram alinhados contra cada contig não redundante de EVE (**Figura 34**). Pelo fato de as EVEs representarem elementos genômicos, podemos notar que, em comparação a coocorrência do viroma (**Figura 13**) o EVEroma é mais homogêneo entre as bibliotecas (linhas) de mosquitos da mesma espécie, como pode ser observado para *A. aegypti*. Apesar desse padrão geral, podemos notar também a formação de agrupamentos internos de mosquitos de diferentes regiões para as espécies *A. aegypti* e *A. albopictus*. A coocorrência dos contigs de EVEs em mesmo agrupamentos (colunas) não foi informativa para inferir sequências oriundas do mesmo vírus, Uma versão interativa do heatmap de coocorrência em formato aplicativo shiny R pode ser acessada no link https://jpalmeida.shinyapps.io/Everome_heatmap/ .

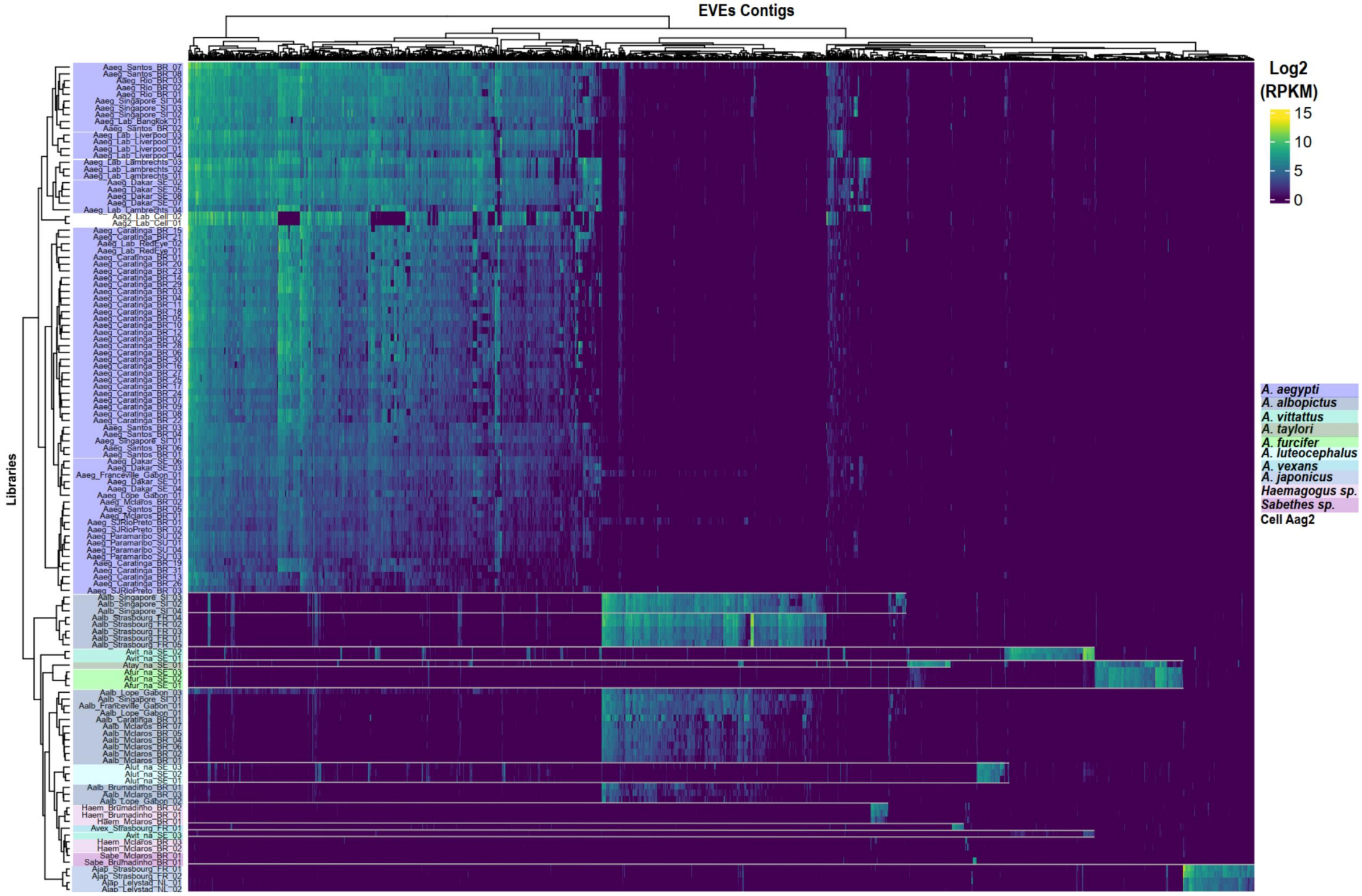

EVEs Contigs
Log2 (RPKM)
15
10
5
0
Libraries
A. aegypti
A. albopictus
A. vittattus
A. taylori
A. furcifer
A. luteocephalus
A. vexans
A. japonicus
Haemagogus sp.
Sabethes sp.
Cell Aag2

Figura 34 – Coocorrência das 1736 sequências não redundantes $_{d+e}$ EVEs nas bibliotecas das 10 espécies de mosquitos amostradas. O heatmap representa a contagem normalizada (RPKM) de pequenos RNAs de 24-30 nt alinhados a cada uma das 1736 sequências não-redundantes de EVEs (colunas) nas 122 bibliotecas analisadas (linhas). Os nomes das bibliotecas indicam a espécies, local de coleta e o número de mosquitos na amostra sequenciada ("spp_local_ID" | "nro. de mosquitos"). As bibliotecas da mesma espécie foram destacadas com a mesma cor. Os dendrogramas apresentados tem como combinação de método e distância "complete" com "euclidean" para as linhas e colunas

Com os resultados das análises de similaridade de sequência podemos inferir que as sequências de EVEs não redundantes pertencem a 115 potenciais vírus distintos que representam 24 famílias virais. Na **Figura 35** é mostrada uma comparação geral da classificação de Baltimore e de famílias virais entre o viroma e o EVEroma das 122 bibliotecas de 10 espécies de mosquitos analisadas nessa tese. A ocorrência (presença/ausência simples) de EVEs representadas por suas espécies virais em cada espécies de mosquito é mostrada na **Figura 36,** são mostradas as classificações de Baltimore, Família viral e número de contigs não redundantes de EVEs que possuem similaridade de sequência com cada espécie viral**.**

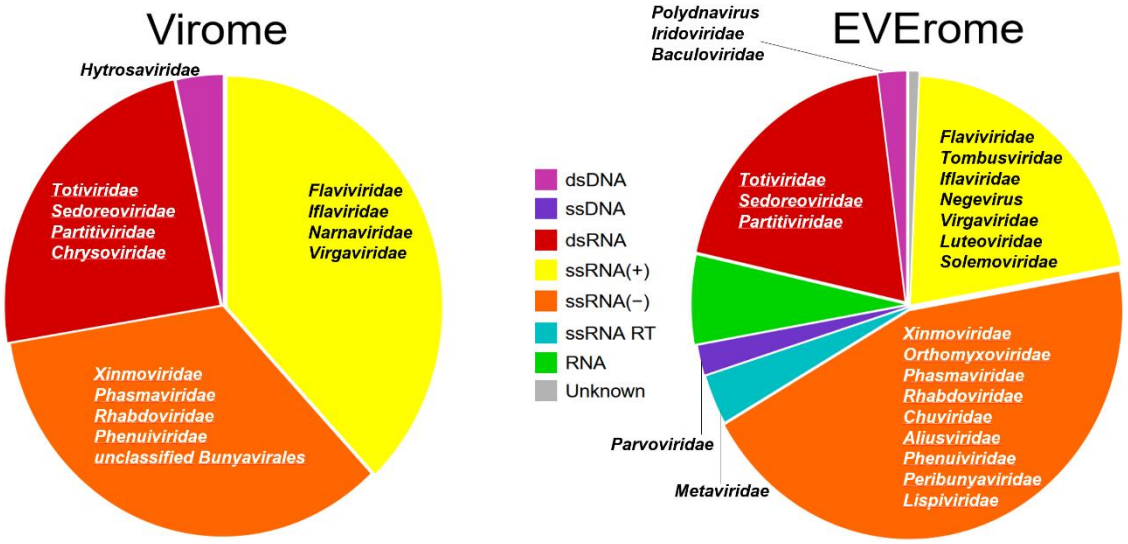


Figura 35 – Proporção de Classes de Baltimore e Familiais virais encontradas no EVEroma e no Viroma. As proporções de Classes de Baltimore são mostradas em relação aos vírus únicos encontrados no Viroma e potenciais vírus únicos que deram origem as sequências do EVEroma. Além das classes de Baltimore, a classe "RNA" inclui vírus de RNA sem genoma classificado e a classe "Unknown" com sequências virais sem qualquer informação sobre o genoma viral de origem.

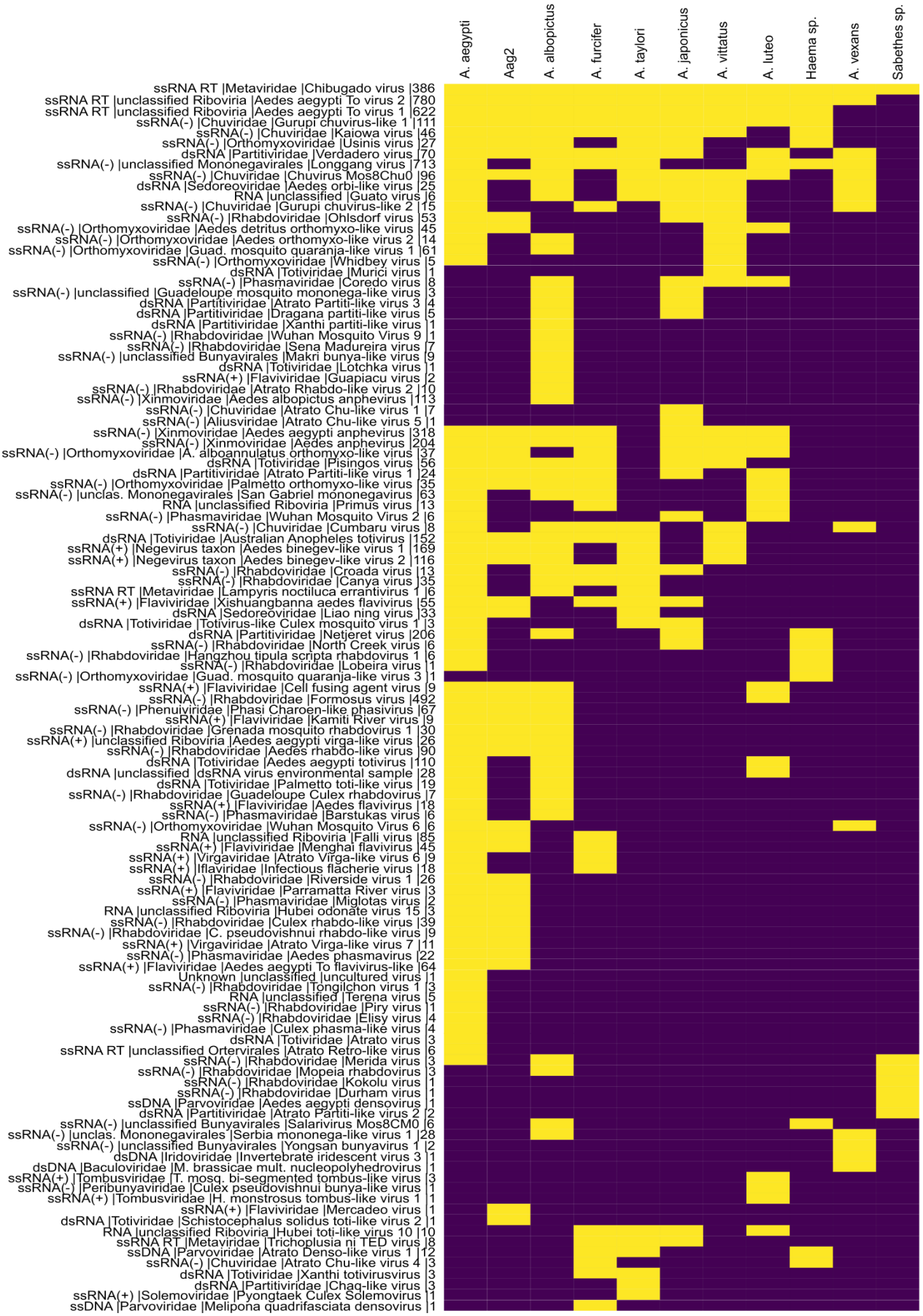


Figura 36 – Coocorrência das espécies virais de EVEs por espécie de mosquito. A cor

amarela marca a presença da espécie viral de EVE na espécie de mosquito (colunas). Cada espécie viral de EVE tem como informações descritas: “Classificação de Baltimore | Família| Spp. | número de contigs não redundantes de EVEs montados para essa Spp. viral”.

Dentre as espécies de mosquitos amostradas em nosso estudo, apenas *A. aegypti* e *A. albopictus* possuem genomas completas sequenciados. Obtivemos 742 contigs não redundantes de EVEs nas bibliotecas de *A. aegypti* e 410 nas bibliotecas de *A. albopictus*. Esses contigs foram alinhados aos genomas de referências dessas espécies. Os alinhamentos dos contigs de EVEs nos genomas de *A. aegypti* representam 5610 inserções virais de 64 vírus. Na **Tabela 10** são mostradas as quantidades de alinhamentos por vírus e as estatísticas de alinhamentos locais das sequências das EVEs. É necessário chamar a tenção para a grande quantidade de alinhamentos da sequência dos retrovírus Aedes *aegypti To virus* 1 e 2 que sozinhos representam quase quatro mil inserções virais no genoma. Destacamos também os alinhamentos genômicos de sequências virais com quase 2Kb e alta identidade com o locus genômico das sequências das EVEs *Formosus vírus* e *Aedes binegev-like virus 1.* Para *A. albopictus*, os alinhamentos dos 410 contigs de EVEs não redundantes representam 1347 inserções virais de 44 vírus **(Tabela 11).** Alinhamentos do retrovírus *Aedes aegypti To virus 2* também são os mais abundantes no genoma de *A. albopictus*, com 555 inserções

Tabela 10 – Alinhamentos dos contigs de EVEs ao genoma de *A. aegypti*.

Virus	Familia	Baltimore	Alinhamenots no genoma				
			Hits	Tamanho médio (nt)	Maior (nt)	Menor (nt)	Identidade média (%)
Aedes aegypti To flavivirus-like	Flaviviridae	ssRNA(+)	12	282.25	354	216	98.65
Aedes aegypti To virus 1	uncla. Riboviria	ssRNA RT	1797	271.44	1149	126	95.72
Aedes aegypti To virus 2	uncla. Riboviria	ssRNA RT	2152	267.62	873	125	96.1
Aedes aegypti anphevirus	Xinmoviridae	ssRNA(-)	208	380.74	1311	141	98.19
Aedes aegypti toti-like virus	Totiviridae	dsRNA	7	222.86	314	151	98.87
Aedes aegypti totivirus	Totiviridae	dsRNA	30	680.27	955	134	99.71
Aedes aegypti totivirus 2	Totiviridae	dsRNA	4	150.5	158	143	99.84
Aedes aegypti virga-like virus	unclassified Riboviria	ssRNA(+)	10	343.1	556	193	94.61
es alboannulatus orthomyxo-like v	Orthomyxoviridae	ssRNA(-)	3	409.33	432	397	97.99
Aedes anphevirus	Xinmoviridae	ssRNA(-)	108	262.06	1201	122	94.01
Aedes binegev-like virus 1	Negevirus taxon	ssRNA(+)	85	424.64	1646	132	98.33
Aedes binegev-like virus 2	Negevirus taxon	ssRNA(+)	7	367.43	754	202	98.23
edes detritus orthomyxo-like viru	Orthomyxoviridae	ssRNA(-)	4	559	773	353	98.77
Aedes orbi-like virus	Sedoreoviridae	dsRNA	2	542	616	468	94.76
Aedes orthomyxo-like virus 2	Orthomyxoviridae	ssRNA(-)	1	173	173	173	98.84
Aedes phasmavirus	Phasmaviridae	ssRNA(-)	1	496	496	496	99.6
Aedes rhabdo-like virus	Rhabdoviridae	ssRNA(-)	17	290.29	973	137	99.1
Atrato Partiti-like virus 1	Partitiviridae	dsRNA	1	316	316	316	98.73
Atrato Retro-like virus	uncla. Ortervirales	ssRNA RT	40	289.35	398	137	92.37
Atrato Virga-like virus 6	Virgaviridae	ssRNA(+)	3	279	387	225	99.15
Atrato Virga-like virus 7	Virgaviridae	ssRNA(+)	3	273	299	260	99.5
Australian Anopheles totivirus	Totiviridae	dsRNA	8	1001.88	1445	297	99.02
Canya virus	Rhabdoviridae	ssRNA(-)	2	437	692	182	99.29
Cell fusing agent virus	Flaviviridae	ssRNA(+)	3	284	378	210	98.66
Chibugado virus	Metaviridae	ssRNA RT	420	304.94	1042	132	97.92
Chuvirus Mos8Chu0	Chuviridae	ssRNA(-)	10	239	754	131	92.38
Croada virus	Rhabdoviridae	ssRNA(-)	20	277.65	537	263	98.69
ulex pseudovishnui rhabdo-like vir	Rhabdoviridae	ssRNA(-)	1	239	239	239	99.58
Culex rhabdo-like virus	Rhabdoviridae	ssRNA(-)	8	284.62	697	155	96.86
Cumbaru virus	Chuviridae	ssRNA(-)	16	152	202	125	98.76
Falli virus	uncla. Riboviria	RNA	28	402.61	742	164	98.88
Formosus virus	Rhabdoviridae	ssRNA(-)	199	233.86	1533	127	96.04
Grenada mosquito rhabdovirus 1	Rhabdoviridae	ssRNA(-)	9	321.33	415	209	97.9
Guadeloupe Culex rhabdovirus	Rhabdoviridae	ssRNA(-)	1	238	238	238	96.64
deloupe mosquito quaranja-like vi	Orthomyxoviridae	ssRNA(-)	8	601.25	1161	196	99.11
Guato virus	unclassified	RNA	9	166	166	166	98.73
Gurupi chuvirus-like 1	Chuviridae	ssRNA(-)	62	407.26	945	167	97.98
Gurupi chuvirus-like 2	Chuviridae	ssRNA(-)	6	564.83	577	504	99.35
ngzhou tipula scripta rhabdovirus	Rhabdoviridae	ssRNA(-)	2	163	171	155	97.02
Hubei odonate virus 15	uncla. Riboviria	RNA	2	196	196	196	98.72
Infectious flacherie virus	Iflaviridae	ssRNA(+)	4	367.5	394	288	96.07
Kaiowa virus	Chuviridae	ssRNA(-)	22	232.95	402	157	99.71
Lampyris noctiluca errantivirus 1	Metaviridae	ssRNA RT	9	191	191	191	92.67
Longgang virus	uncla. Mononegavirales	ssRNA(-)	70	421.71	1063	143	95.4
Menghai flavivirus	Flaviviridae	ssRNA(+)	10	269.5	446	188	96
Merida virus	Rhabdoviridae	ssRNA(-)	2	357.5	358	357	99.58
Netjeret virus	Partitiviridae	dsRNA	37	354.22	849	127	95.15
North Creek virus	Rhabdoviridae	ssRNA(-)	1	156	156	156	99.36
Ochlerotatus scapularis flavivirus	Flaviviridae	ssRNA(+)	7	275.43	344	172	99.42
Ohlsdorf virus	Rhabdoviridae	ssRNA(-)	22	461.68	736	131	98.12
Palm Creek virus	Flaviviridae	ssRNA(+)	4	211.5	294	184	99.15
Palmetto orthomyxo-like virus	Orthomyxoviridae	ssRNA(-)	6	497.67	1354	132	95.48
Phasi Charoen-like phasivirus	Phenuiviridae	ssRNA(-)	10	323.3	621	163	98.91
Pisingos virus	Totiviridae	dsRNA	5	185.4	221	168	99.16
Primus virus	uncla. Riboviria	RNA	1	394	394	394	99.24
Riverside virus 1	Rhabdoviridae	ssRNA(-)	5	198.6	295	159	97.92
San Gabriel mononegavirus	uncla. Mononegavirales	ssRNA(-)	8	429.12	915	190	98.02
Tongilchon virus 1	Rhabdoviridae	ssRNA(-)	5	407	489	226	98.41
Usinis virus	Orthomyxoviridae	ssRNA(-)	25	303.8	451	195	98.62
Verdadero virus	Partitiviridae	dsRNA	22	913.27	997	700	97.76
Xishuangbanna aedes flavivirus	Flaviviridae	ssRNA(+)	20	219.2	530	129	94.99
lsRNA virus environmental sampl	unclassified	dsRNA	4	385.25	604	225	98.32
uncultured virus	unclassified	Unknown	2	217	217	217	99.31

Tabela 11 – Alinhamentos dos contigs de EVEs ao genoma de *A. albopictusi*.

Virus	Familia	Baltimore	Alinhamentos no genoma				
			Hits	Tamanho médio (nt)	Maior (nt)	Menor (nt)	Identidade média (%)
Aedes aegypti To virus 1	unclassified Riboviria	ssRNA RT	124	248.36	409	131	94.96
Aedes aegypti To virus 2	unclassified Riboviria	ssRNA RT	555	229.1	504	130	95.52
Aedes aegypti anphevirus	Xinmoviridae	ssRNA(-)	3	319	350	277	94.75
Aedes aegypti toti-like virus	Totiviridae	dsRNA	3	452	686	304	88.58
Aedes aegypti virga-like virus	uncla. Riboviria	ssRNA(+)	6	342.67	366	260	92.62
Aedes albopictus anphevirus	Xinmoviridae	ssRNA(-)	51	251.22	850	127	91.55
Aedes albopictus cell fusing agent virus	Flaviviridae	ssRNA(+)	5	168.6	203	132	92.34
Aedes anphevirus	Xinmoviridae	ssRNA(-)	1	323	323	323	98.45
Aedes binegev-like virus 1	Negevirus taxon	ssRNA(+)	8	223.12	272	186	96.19
Aedes binegev-like virus 2	Negevirus taxon	ssRNA(+)	63	314.56	631	141	92.22
Aedes flavivirus	Flaviviridae	ssRNA(+)	5	266.2	294	164	98.93
Aedes orbi-like virus	Sedoreoviridae	dsRNA	4	170	208	132	98.61
Aedes orthomyxo-like virus 2	Orthomyxoviridae	ssRNA(-)	5	243.2	338	136	97.27
Atrato Partiti-like virus 1	Partitiviridae	dsRNA	5	239.4	269	166	98.47
Barstukas virus	Phasmaviridae	ssRNA(-)	1	151	151	151	99.34
Canya virus	Rhabdoviridae	ssRNA(-)	10	335.4	663	164	93.99
Chibugado virus	Metaviridae	ssRNA RT	144	221.47	447	124	94.22
Chuvirus Mos8Chu0	Chuviridae	ssRNA(-)	10	293.7	692	131	95.72
Coredo virus	Phasmaviridae	ssRNA(-)	6	224.83	303	161	96.13
Croada virus	Rhabdoviridae	ssRNA(-)	6	161	161	161	91.93
Dragana partiti-like virus	Partitiviridae	dsRNA	3	489	489	489	94.96
Formosus virus	Rhabdoviridae	ssRNA(-)	9	228.67	316	171	92.71
Grenada mosquito rhabdovirus 1	Rhabdoviridae	ssRNA(-)	3	196	229	165	94.84
Guadeloupe Culex rhabdovirus	Rhabdoviridae	ssRNA(-)	1	566	566	566	98.41
Guadeloupe mosquito mononega-like virus	unclassified	ssRNA(-)	1	185	185	185	92.43
Guadeloupe mosquito quaranja-like virus 1	Orthomyxoviridae	ssRNA(-)	7	341.14	492	230	96.74
Guapiacu virus	Flaviviridae	ssRNA(+)	3	449.67	473	403	92.99
Guato virus	unclassified	RNA	167	191.81	220	156	94.06
Gurupi chuvirus-like 1	Chuviridae	ssRNA(-)	10	337.8	567	172	97.71
Kaiowa virus	Chuviridae	ssRNA(-)	26	292.62	476	141	99.75
Longgang virus	uncla. Mononegavirales	ssRNA(-)	23	264.09	539	138	94.33
Merida virus	Rhabdoviridae	ssRNA(-)	1	226	226	226	100
Mopeia rhabdovirus	Rhabdoviridae	ssRNA(-)	1	229	229	229	99.56
Netjeret virus	Partitiviridae	dsRNA	13	314.69	404	153	94.5
Palmetto orthomyxo-like virus	Orthomyxoviridae	ssRNA(-)	4	332.5	462	130	95.74
Phasi Charoen-like phasivirus	Phenuiviridae	ssRNA(-)	2	146	146	146	95.21
Pisingos virus	Totiviridae	dsRNA	4	245.5	292	203	85.89
San Gabriel mononegavirus	uncla.. Mononegavirales	ssRNA(-)	2	278.5	296	261	95.04
Sena Madureira virus	Rhabdoviridae	ssRNA(-)	2	783.5	884	683	98.75
Serbia mononega-like virus 1	uncla. . Mononegavirales	ssRNA(-)	10	193.8	266	162	96.14
Usinis virus	Orthomyxoviridae	ssRNA(-)	4	198.25	247	173	92.62
Verdadero virus	Partitiviridae	dsRNA	34	216.62	359	125	96.52
Xanthi partiti-like virus	Partitiviridae	dsRNA	1	205	205	205	96.58
dsRNA virus environmental sample	unclassified	dsRNA	1	456	456	456	88.38

Tabela 12 – Alinhamentos estatisticamente significativos de contigs não redundantes do EVEroma no Viroma e correlações entre as cargas de pequenos RNAs EVE/vírus.

Mosquito	EVE	gene EVE	Contig EVE (nt)	Alinhamento (nt)	Virus	Contig viral (nt)	Ident(%)	evalue	Cor. piEVE/siViral	Cor. piEVE/piViral
Aalb	Barstukas virus	capsid	367	367	APV	2045	99.455	0.0	0.98	0.99
Aalb	Barstukas virus	capsid	284	289	APV	802	94.118	1.96e-123	0.95	0.99
Aalb	Barstukas virus	capsid	284	284	APV	2045	85.563	7.46e-83	0.91	0.94
Aaeg	Aedes anphevirus	glycoprotein	146	146	AAV	1057	99.315	3.66e-73	0.93	0.99
Aaeg	Aedes anphevirus	capsid	196	177	AAV	267	99.435	2.97e-90	0.94	0.82
Aaeg	Aedes anphevirus	capsid	142	142	AAV	248	100.000	1.27e-72	0.94	0.99
Aaeg	Aedes anphevirus	glycoprotein	165	165	AAV	1057	99.394	1.15e-83	0.93	0.98
Aaeg	Aedes anphevirus	RdRP	100	100	AAV	415	100.000	1.91e-49	0.93	0.99
Aaeg	Aedes anphevirus	glycoprotein	216	149	AAV	575	100.000	2.61e-76	0.93	0.99
Aag2	Cell fusing agent virus	NS2A	120	120	CFAV	1073	98.333	3.90e-57	0.93	0.89
Aaeg	Aedes anphevirus	capsid	250	234	AAV	236	85.897	1.85e-68	0.65	0.89
Aaeg	Aedes anphevirus	capsid	282	143	AAV	236	89.510	7.78e-48	0.48	0.78
Atay	Chaq-like	hypothetical	504	478	AevexCV1	689	80.962	7.88e-105	0.11	0.37
Aalb	Makri bunya-like virus	capsid	222	211	APV	802	99.526	4.27e-109	0.91	0.99
Aalb	Makri bunya-like virus	capsid	222	206	APV	383	99.029	1.20e-104	0.89	0.99
Aalb	Formosus virus	polymerase	216	123	APV	294	91.870	3.52e-45	0.98	0.98
Aalb	Makri bunya-like virus	capsid	234	234	APV	802	98.718	1.60e-118	0.94	0.99
Ajap	Aedes aegypti anphevirus	glycoprotein	393	336	AejapAV1	1420	99.405	1.18e-176	0.99	0.99
Ajap	Aedes aegypti anphevirus	glycoprotein	393	128	AejapAV1	1035	99.219	1.08e-62	0.99	0.99
Ajap	Aedes anphevirus	glycoprotein	363	363	AejapAV1	1420	98.347	0.0	0.99	0.99
Ajap	Aedes anphevirus	glycoprotein	363	363	AejapAV1	1035	98.072	0.0	0.99	0.99
Ajap	Chuvirus Mos8Chu0	glycoprotein	963	293	AejapAV1	299	100.000	1.10e-155	0.99	0.99
Ajap	Chuvirus Mos8Chu0	glycoprotein	963	101	AejapAV1	243	98.020	1.29e-45	0.99	0.99
Ajap	Aedes aegypti anphevirus	glycoprotein	219	189	AejapAV1	266	97.354	3.35e-90	0.98	0.99
Ajap	Aedes anphevirus	glycoprotein	402	268	AejapAV1	343	98.134	7.72e-134	0.99	0.99
Ajap	Aedes aegypti anphevirus	glycoprotein	255	142	AejapAV1	1035	96.479	5.29e-64	0.98	0.99
Ajap	Aedes aegypti anphevirus	glycoprotein	255	145	AejapAV1	1420	95.862	5.29e-64	0.98	0.99
Ajap	[Chuvirus Mos8Chu0	glycoprotein	885	254	AejapAV1	299	100.000	4.83e-134	0.99	0.99
Ajap	[Chuvirus Mos8Chu0	glycoprotein	885	240	AejapAV1	243	100.000	2.93e-126	0.99	0.99
Ajap	Aedes anphevirus	glycoprotein	405	395	AejapAV1	1420	97.215	0.0	0.98	0.99
Ajap	Aedes anphevirus	glycoprotein	333	311	AejapAV1	1035	99.678	1.69e-164	0.98	0.99
Ajap	Aedes anphevirus	glycoprotein	333	207	AejapAV1	779	98.551	2.38e-103	0.98	0.99
Ajap	Aedes anphevirus	glycoprotein	333	152	AejapAV1	1420	99.342	4.12e-76	0.98	0.99

Com o intuito de identificar pares cognatos de Vírus/EVEs que pudessem evidenciar alguma relação imune das EVEs identificadas com os vírus circulantes detectados nesse estudo, alinhamos os 1736 contigs não redundantes do EVEroma aos 267 não redundantes do viroma. Foram obtidos apenas 34 alinhamentos significativos de 24 contigs de EVEs com contigs de 5 espécies virais (**Tabela 12**). Foram calculados os coeficientes de correlação (Pearson) das cargas de piRNAs (24-30nt) com a de siRNA (20-22nt) de cada par cognato de contigs EVE/vírus nas bibliotecas em que os vírus correspondentes foram encontrados. Com exceção de três pares cognatos, todos apresentaram coeficiente de correlação maior que 0.9, indicando um aumento mútuo da carga de siRNA e piRNAs dos pares. Obtivemos os mesmos resultados ao calcular o coeficiente de correlação substituindo a carga viral de siRNA pela de piRNAs virais.

5.4 - Diferenciação de sequências Virais e EVEs utilizando aprendizado de máquina

Para as abordagens de aprendizado de máquina, utilizamos o conjunto de dados mais bem estabelecido por curadoria manual até o momento composto por 1321 sequências virais e 994 sequências de EVEs. Os resultados da execução dos algoritmos não-supervisionados PCA e t-SNE utilizando as sequências virais representadas por 48 atributos quantitativos foram plotados em duas dimensões (**Figura 37**). Há uma evidente separação da maioria dos pontos em dois grandes agrupamentos representando as classes “Viral” e EVEs. No resultado da PCA, a separação dos dois agrupamentos ao longo do eixo da PC1 possui suporte de 54% da variância explicada por essa componente. Esses resultados evidenciam a alta capacidade distintiva dos atributos representativos para as classes “viral” e “EVE”.

Os resultados das métricas de desempenho dos melhores modelos com hiper parâmetros refinados para cada algoritmo são mostrados na **Tabela 13**. Com exceção do modelo *Naive Bayes* (Acurácia = 0.88) todos os modelos apresentam acurácia maior que 0.9. As demais métricas também evidenciam um bom desempenho do modelo na diferenciação das duas classes para todos os algoritmos testados. Os. modelos do tipo ensemble apresentaram as maiores acurácias (Acurácia. = 0.93).

Figura 37 – Reduções dimensionais com PCA e t-SNE utilizando as representações de atributos baseados em quantificações de pequenos RNAs. As reduções dimensionais foram realizadas com o conjunto de dados de sequências virais representadas por 48 atributos de pequenos RNAs. O resultado da análise de componentes principais (PCA) foi plotado em duas dimensões. PC1 explica 54% da variância e a PC2 18%. Resultado da redução dimensional com t-SNE foi plotado em duas dimensões. Pontos vermelhos representam contigs virais e azuis contigs de EVEs.

Tabela 13 – Resultados das métricas de desempenho dos melhores modelos de cada algoritmo testado para distinção de sequências virais de EVEs após hiper parâmetros refinados em função da acurácia.

Modelo	Acur.	Precisão				Revocação				f1-score			
		EVE	Viral	macro	balan.	EVE	Viral	macro	balan.	EVE	Viral	macro	balan.
Naive Bayes	0.88	0.86	0.89	0.88	0.88	0.85	0.90	0.87	0.88	0.86	0.89	0.88	0.88
Decision Tree	0.92	0.95	0.89	0.92	0.92	0.85	0.96	0.91	0.92	0.90	0.93	0.91	0.91
SVM	0.91	0.95	0.89	0.92	0.92	0.84	0.97	0.90	0.91	0.89	0.93	0.91	0.91
KNN	0.92	0.94	0.91	0.92	0.92	0.87	0.96	0.91	0.92	0.90	0.93	0.92	0.92
Random Forest	0.93	0.96	0.91	0.93	0.93	0.87	0.97	0.92	0.93	0.91	0.94	0.92	0.93
Adap. Boosting	0.92	0.94	0.91	0.92	0.92	0.87	0.96	0.91	0.92	0.90	0.93	0.92	0.92
Grad. Tree Boost.	0.93	0.94	0.92	0.93	0.93	0.89	0.96	0.92	0.93	0.91	0.94	0.92	0.93
CatBoost	0.93	0.96	0.91	0.93	0.93	0.87	0.97	0.92	0.93	0.91	0.94	0.92	0.93
XGBoost	0.93	0.95	0.92	0.93	0.93	0.88	0.96	0.92	0.93	0.91	0.94	0.93	0.93
LightGBM	0.93	0.95	0.91	0.93	0.93	0.88	0.97	0.92	0.93	0.91	0.94	0.93	0.93

Na **Figura 38** são mostradas as curvas ROC e DET do desempenho dos modelos. Ambos resultados evidnciam alto desempenho de classificação, com os

modelos do tipo ensemble apresentando os maiores valores de AUC, *Random Forest* (tipo *bagging*) e *CatBoost* (tipo *boosting*) com AUC = 0.98.

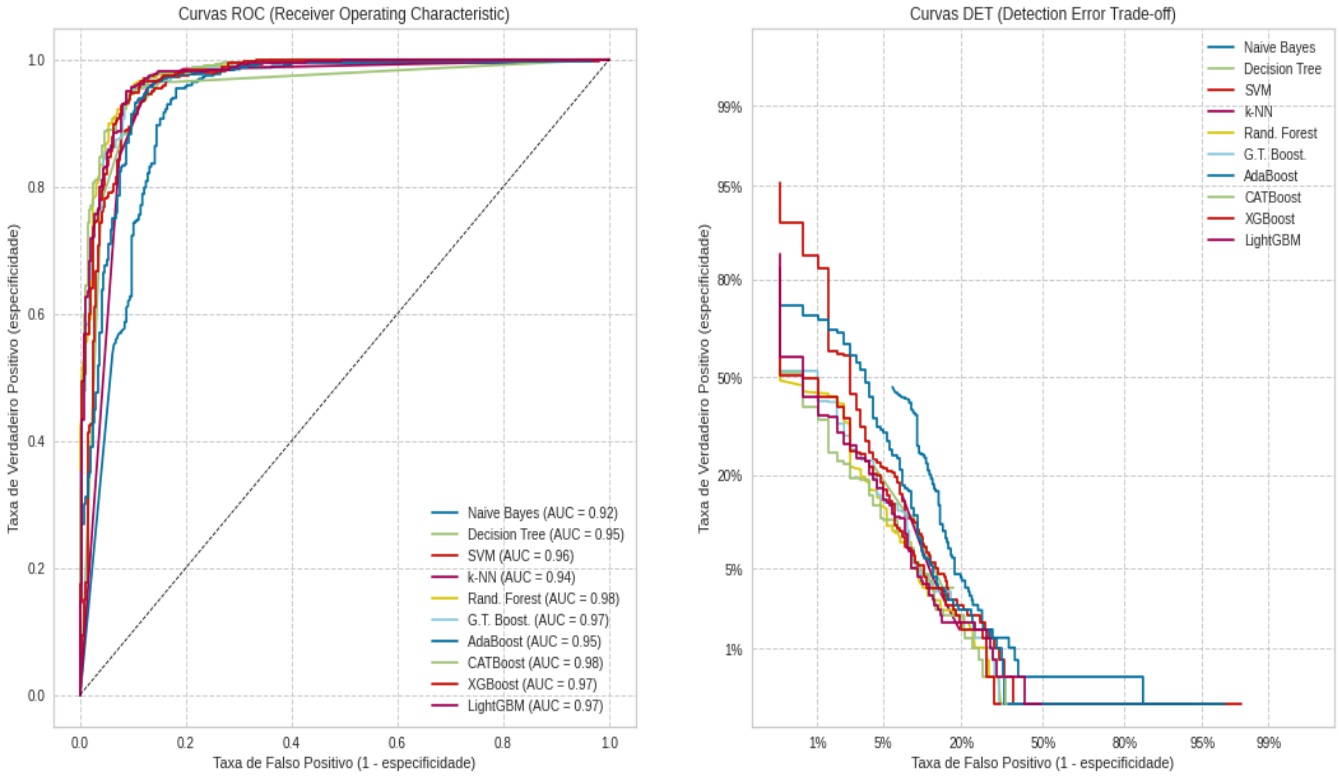


Figura 38 – Avaliação de desempenho dos classificadores treinados para distinguir sequências Virais de EVEs. Os hiperparâmetros dos modelos classificadores obtidos para cada um dos dez diferentes algoritmos foram otimizados em função da acurácia. Os gráficos mostram as curvas ROC e DET dos modelos com melhor acurácia obtidos para cada algoritmo. Todos os modelos apresentam ´ótimo desempenho da distinção das duas classes com valores de AUC maiores que 0.90, com destaque para os modelos do tipo ensemble, Random Forest e CATBoost com valores de AUC = 0.98.

5.5 - Ferramenta para análise de viromas com bibliotecas de pequenos RNAs

Concluímos com sucesso a conteinerização de todos os scripts e dependências necessárias para execução do pipeline de metagenômica utilizando bibliotecas de pequenos RNAs (**Figura 39**). A ferramenta produzida recebe o nome de Small RNA Metavir e já disponibilizamos uma versão funcional do contêiner no GitHub https://github.com/rnai-bioinfo/small-rna-metavir .

O pipeline é um programa de computador encapsulado em um container Docker inteiramente preparado para sua execução sem a necessidade de qualquer tipo de gestão de dependências ou demais configurações exceto a disponibilidade de um container runtime. A codificação do pipeline consiste num

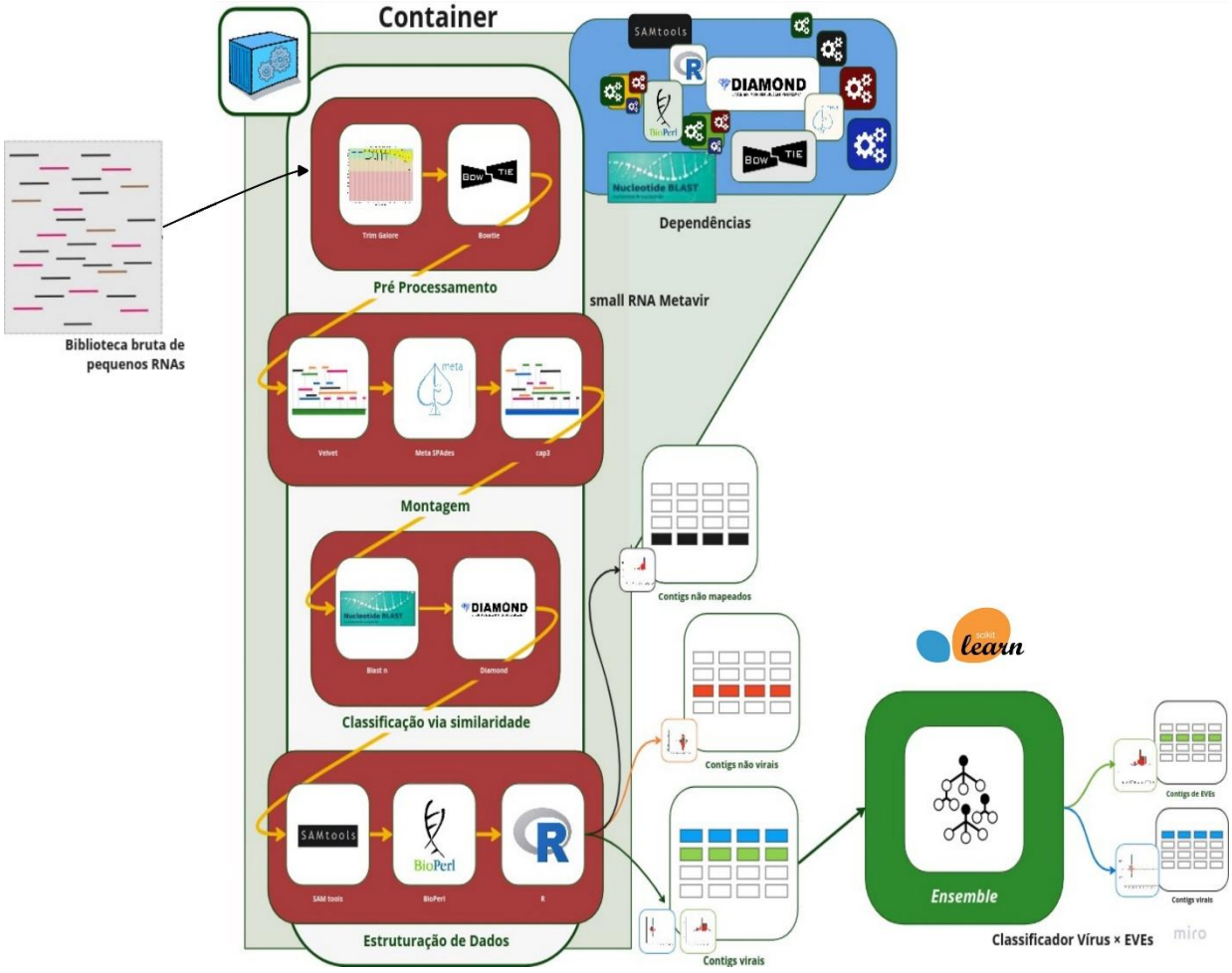


Figura 39 – Representação esquemática da ferramenta *small RNA Metavir* com suas principais dependências e recursos conteinerizados.

script principal *main.pl* e um conjunto de scripts auxiliares escritos nas linguagens Perl, Python, R e Shell Scripts.

O pipeline é preparado para funcionar dentro de uma organização específica de diretórios a qual é assegurada por seu container de execução:

/small_rna_metavir/: Diretório principal que comporta códigos fontes e demais arquivos atrelados e / ou gerados pela execução do pipeline;

/small_rna_metavir/src: Localização dos arquivos fonte executados dentro do pipeline;

/small_rna_metavir/asset: Localização interna de todas as dependências do pipeline, tais como os arquivos de entrada das bibliotecas a serem processadas;

A partir de uma biblioteca de pequenos RNAs como entrada a ferramenta tem como principais arquivos de saída:

1- Arquivos de texto com informações de reads alinhadas ao genoma do hospedeiro escolhido e a referência de genoma de bactérias e figura com a distribuição e tamanho dos reads alinhados ao hospedeiro;
2- Arquivos fasta com contigs montados a partir da combinação de diferentes *k-mers* e intervalos de tamanhos de reads;
3- Arquivos tabulares com resultados de alinhamentos locais dos contigs maiores que 200nt com os programas Blastn e DIAMOND;
4- Arquivos fasta com contigs classificadas como “Virais”; “Não-virais” e “Desconhecidas” contendo a descrição do melhor alinhamento no cabeçalho de cada sequência;
5- Tabelas contendo os contigs classificados como “Virais”; “Não-virais” e “Desconhecidos” representados por 48 atributos quantitativos gerados a partir do processamento dos alinhamentos de pequenos RNAs a cada contig;
6- Para cada contig das três classes são geradas figuras do perfil de distribuição de tamanho de pequenos RNAs, cobertura de reads totais ou separadas por intervalos representativos das vias de pequenos RNAs, *e.g* 21nt, 24-30 nt.
7- Arquivo tabular com classificação dos contigs virais em “Viral” ou “EVE”.

Durante a execução da tese a ferramenta passou por etapas de refinamento ao longo das análises das 122 bibliotecas de pequenos RNAs do projeto ZikaAlliance geradas pelo nosso grupo. Após a criação do primeiro container a ferramenta foi testada com 99 bibliotecas públicas de pequenos RNAs obtidas do SRA. Um resumo dos resultados das execuções é apresentado na **Tabela Suplementar 3.**

5.6 - Detecção de pequenos RNAs de *Wolbachia*

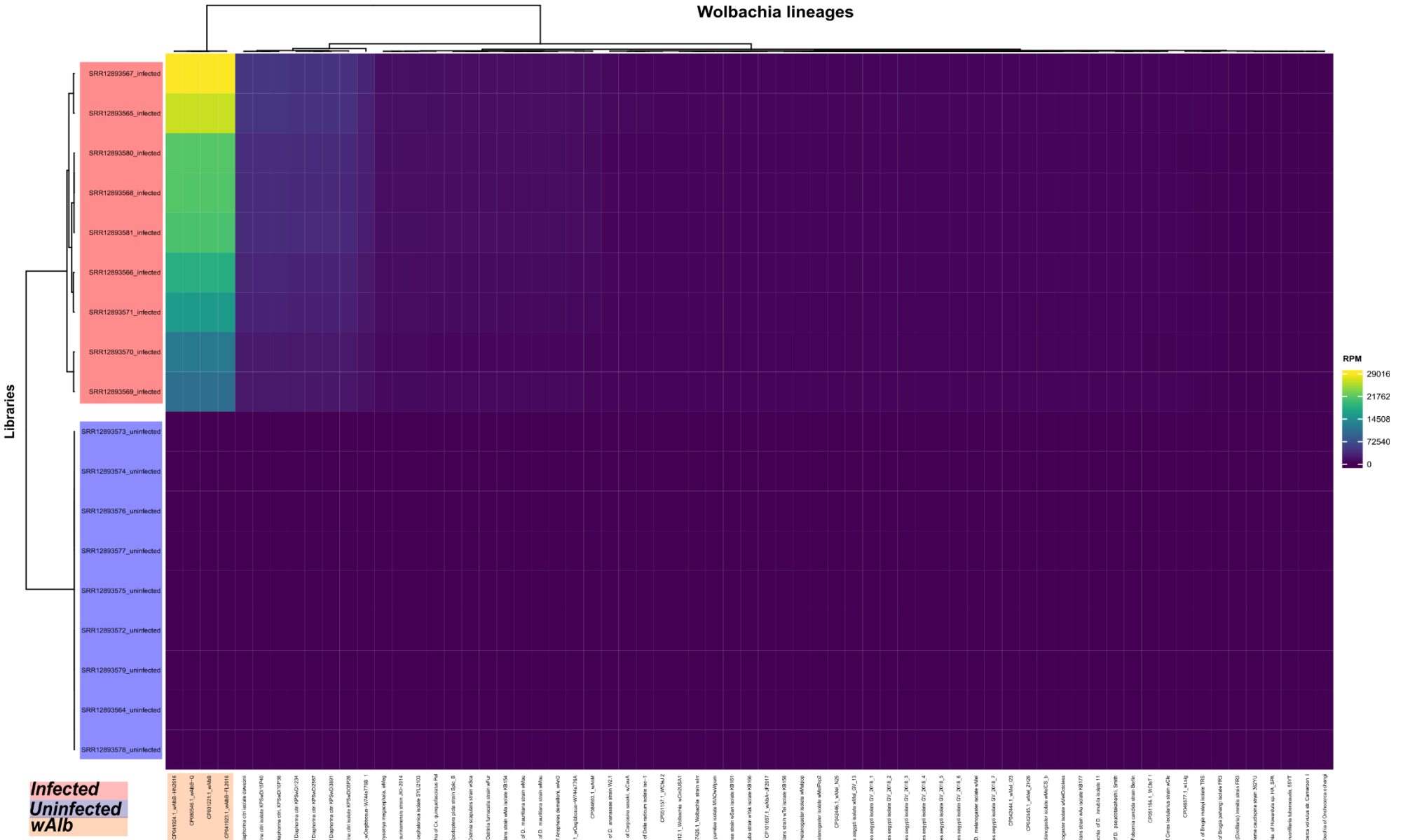
5.6.1 - Detecção de pequenos RNAs de *Wolbachia* em dados públicos de experimento controle com tetraciclina

Para testar se nossa estratégia de análise de pequenos RNAs pode ser usada para detecção e quantificação de pequenos RNAs de *Wolbachia* em bibliotecas de

mosquitos, analisamos 18 bibliotecas de RNA-seq de pequenos RNAs públicas de mosquitos *Ae. aegypti* infectados e não-infectados por *Wolbachia* (BISHOP et al., 2022). No experimento que deu origem a essas bibliotecas, a infecção por *Wolbachia* foi removida com o uso do antibiótico tetraciclina. Na **Figura 40**, comparamos a quantidade de pequenos RNAs alinhados a genomas de *Wolbachia* entre as bibliotecas oriundas de mosquitos infectados e não-infectados por *Wolbachia*, evidenciado que os pequenos RNAs podem ser usados para detecção e quantificação da carga de *Wolbachia* em uma amostra via RNA-seq. Há alta especificidade de detecção da linhagem de *Wolbachia* utilizando reads de pequenos RNAs. Os alinhamentos de reads oriundas das bibliotecas de mosquitos infectados alinharam majoritariamente aos genomas da linhagem *wAlb,* a linhagem utilizada para infecção dos mosquitos no estudo do qual as bibliotecas de RNA-seq foram obtidas.

Para a **Figura 40** a legenda é mostrada antes da Figura, que se encontra na próxima página

Figura 40 – Detecção de pequenos RNAs de *Wolbachia* em bibliotecas de *A. aegypti* infectados artificialmente com o endossimbionte em laboratório de laboratório. Bibliotecas de pequenos RNAs geradas de amostras de *A. aegypti* infectados com *wAlb* em laboratório e divididos em tratados e não-tratados com tetraciclina para remoção da infecção pelo endossimbionte foram obtidas do SRA e são oriundas do experimento de BISHOP et al. 2022. Após filtragem de reads alinhadas ao genoma de *A. aegypti as reads* foram alinhadas a uma referência contendo 67 genomas de linhagens de *Wolbachia* (colunas). Bibliotecas de mosquitos não tratados com tetraciclina estão destacadas em vermelho e de mosquitos tratados com tetraciclina de azul. Os genomas de *Wolbachia*s da linhagem *wAlb* estão destacadas de laranja. A quantidade de reads alinhadas aos genomas de *Wolbachia*s foram normalizadas por RPM.


Wolbachia lineages
Libraries
SRR12893567_infected
SRR12893565_infected
SRR12893580_infected
SRR12893568_infected
SRR12893581_infected
SRR12893566_infected
SRR12893571_infected
SRR12893570_infected
SRR12893569_infected
SRR12893573_uninfected
SRR12893574_uninfected
SRR12893576_uninfected
SRR12893577_uninfected
SRR12893575_uninfected
SRR12893572_uninfected
SRR12893579_uninfected
SRR12893564_uninfected
SRR12893578_uninfected
RPM
29016
21762
14508
72540
0
Infected
Uninfected
wAlb

Na **Figura 41** é mostrada a distribuição de tamanho dos pequenos RNAs das nove bibliotecas de mosquitos infectados alinhados ao genoma de *wAlb*. Podemos notar uma distribuição de tamanho de pequenos RNAs de *wAlb* similar entre as bibliotecas, com os tamanhos de 20 e 21 nt mais frequentes e com proporções similares em todas as bibliotecas. Na **Figura 42** são mostrados os 40 genes de *wAlb* com maiores contagens de pequenos RNAs alinhados. Pequenos RNAs de tRNAs são os mais abundantes. Também podemos notar uma grande quantidade de pequenos RNAs de outros "*house keepings*" ncRNAs sabidamente abundantes em células procarióticas como: 6S, signal recognition particle sRNA, 16S rRNA, 23S rRNA, RNAse P RNA. Pequenos RNAs de genes codificadores de proteínas ribossomais também são abundantes.

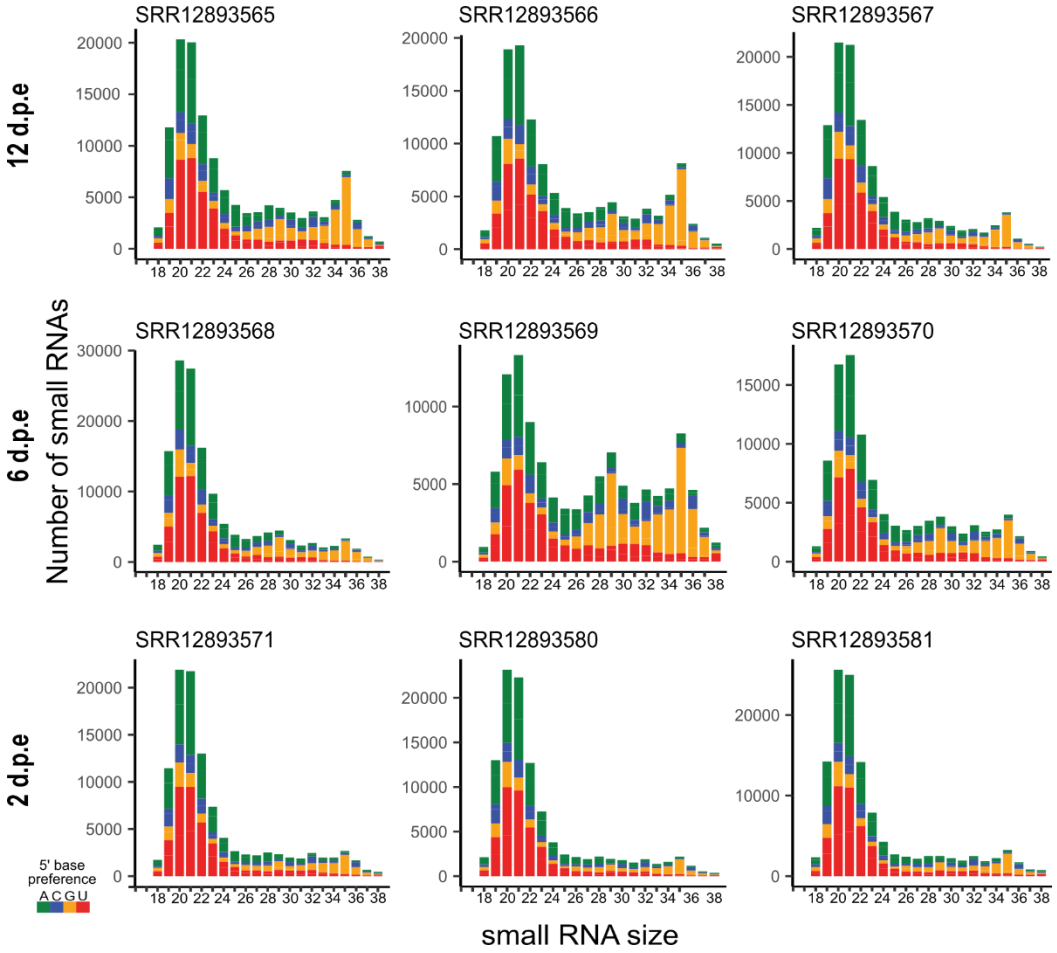


Figura 41 – Distribuição de tamanho dos pequenos RNAs de *wAlb* artificialmente infectando *A. aegypti*. A distribuição de tamanhos dos pequenos RNAs alinhados ao genoma de *wAlb* foi plotada para as 9 bibliotecas não tratadas com tetraciclina. Cores nas barras representam a frequência de nucleotídeos na primeira base 5' dos reads. d.p.e (dias pós emergência dos mosquitos usados no experimento de BISHOP et al., 2022).

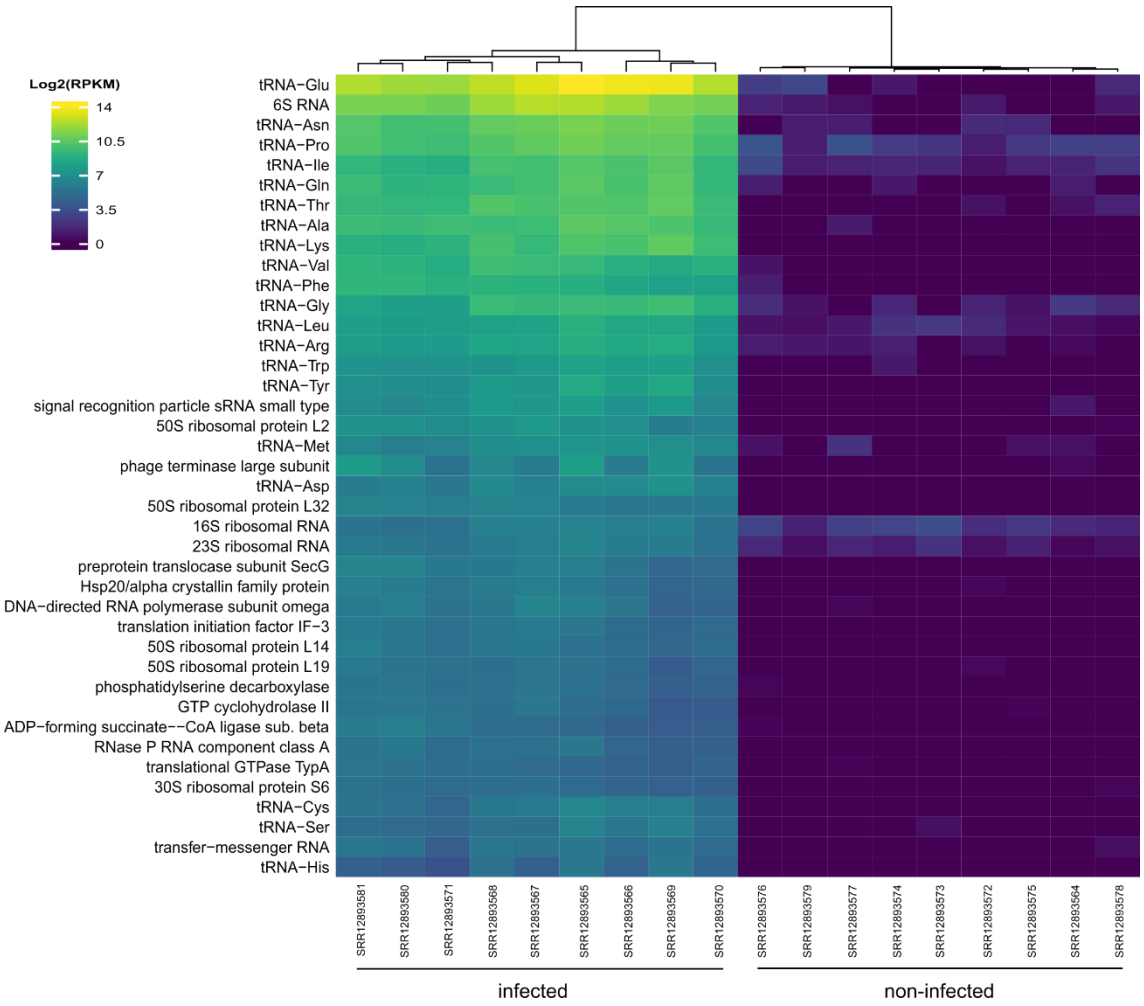


Figura 42 – 40 genes com as maiores contagens de pequenos RNAs de *wAlb* artificialmente infectando *A. aegypti.* Utilizando a anotação do genoma de *wAlb* as reads alinhadas ao genoma foram quantificadas por gene com a ferramenta HTSeq. A contagem de reads por gene foi normalizada por RPKM e os 40 genes com mais alinhamentos nas bibliotecas não tratadas com tetraciclina foram plotados no heatmap.

5.6.2 - Detecção de pequenos RNAs de *Wolbachia* nas bibliotecas de *A. albopictus* de campo

Para avaliar a estratégia de detecção de pequenos RNAs de *Wolbachia* em bibliotecas geradas a partir de mosquitos de campo, começamos analisando as 22 bibliotecas *de A. albopictus* geradas por nosso grupo devido ao fato de que essa espécie é sabidamente infectada por *Wolbachia* naturalmente. Apesar de um padrão de alinhamentos mais ruidosos comparadas as bibliotecas de *A. aegypti* infectadas em laboratório, podemos observar um agrupamento de 11 bibliotecas com uma grande quantidade de reads alinhadas especificamente aos genomas da linhagem da *wAlb* coerente com a infecção natural esperada em mosquitos de campo (**Figura 43**). As bibliotecas de *A. albopictus* do Gabão e

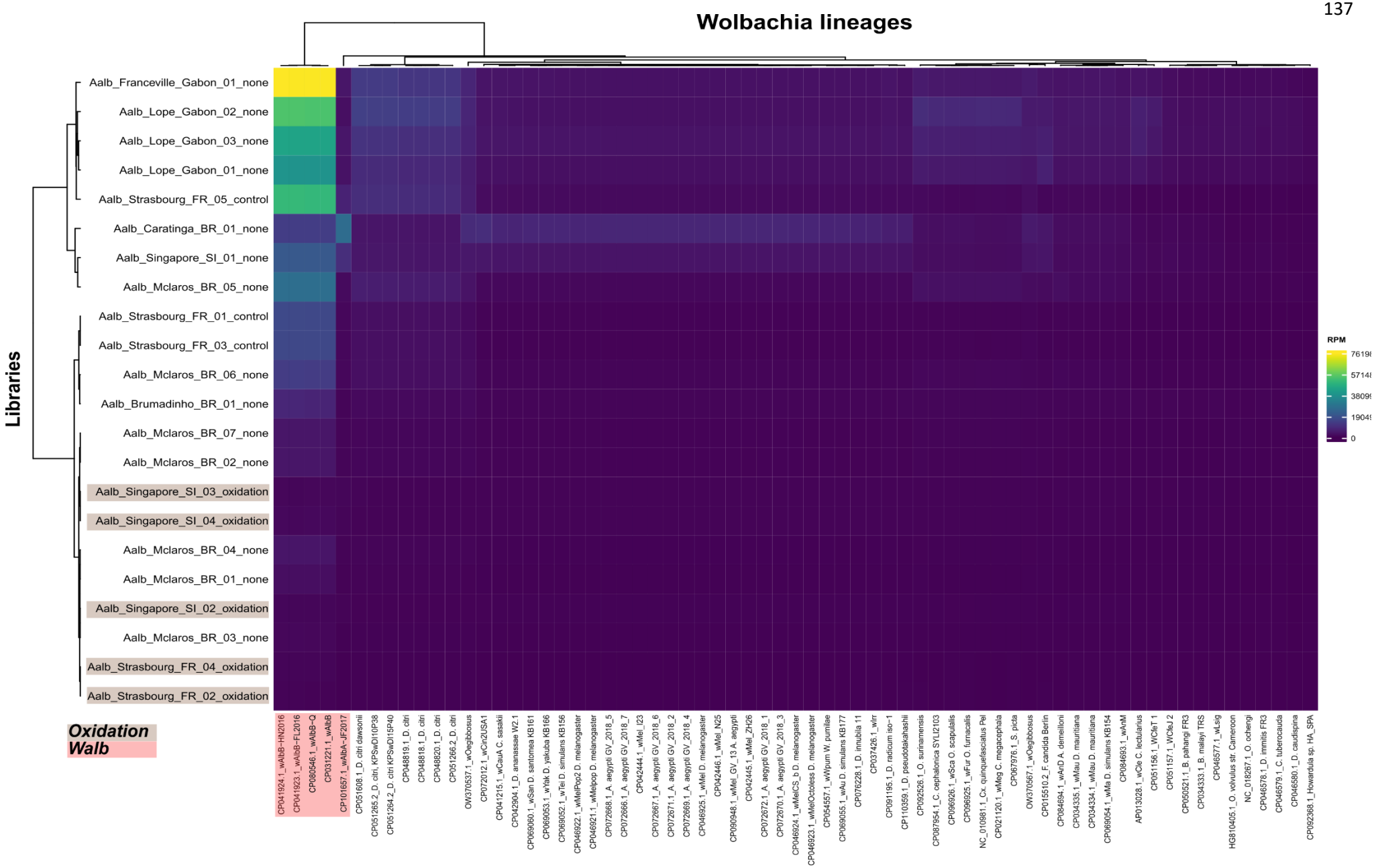


Figura 43 - Detecção de pequenos RNAs de *Wolbachia* em bibliotecas de campo de *A. albopictus*. As 22 bibliotecas de pequenos RNAs de *A. albopictus* coletados em campo no projeto ZIkaAlliance foram alinhadas a referência contendo 67 genomas de linhagens de *Wolbachia* (colunas). Bibliotecas que passaram por oxidação de RNA antes do sequenciamento estão destacadas em marrom. As bibliotecas Aalb_Strasbourg_FR_01, Aalb_Strasbourg_FR_03, Aalb_Strasbourg_FR_05 são amostras controles do processo de oxidação aplicado no preparo das bibliotecas Aalb_Strasbourg_FR_02 e Aalb_Strasbourg_FR_04. Os genomas de *Wolbachia*s da linhagem *wAlb* estão destacadas de vermelho. A quantidade de reads alinhadas aos genomas de *Wolbachia*s foram normalizadas por RPM.

uma biblioteca da França formaram um agrupamento de mosquitos com uma alta carga de pequenos RNAs de *wAlb*.

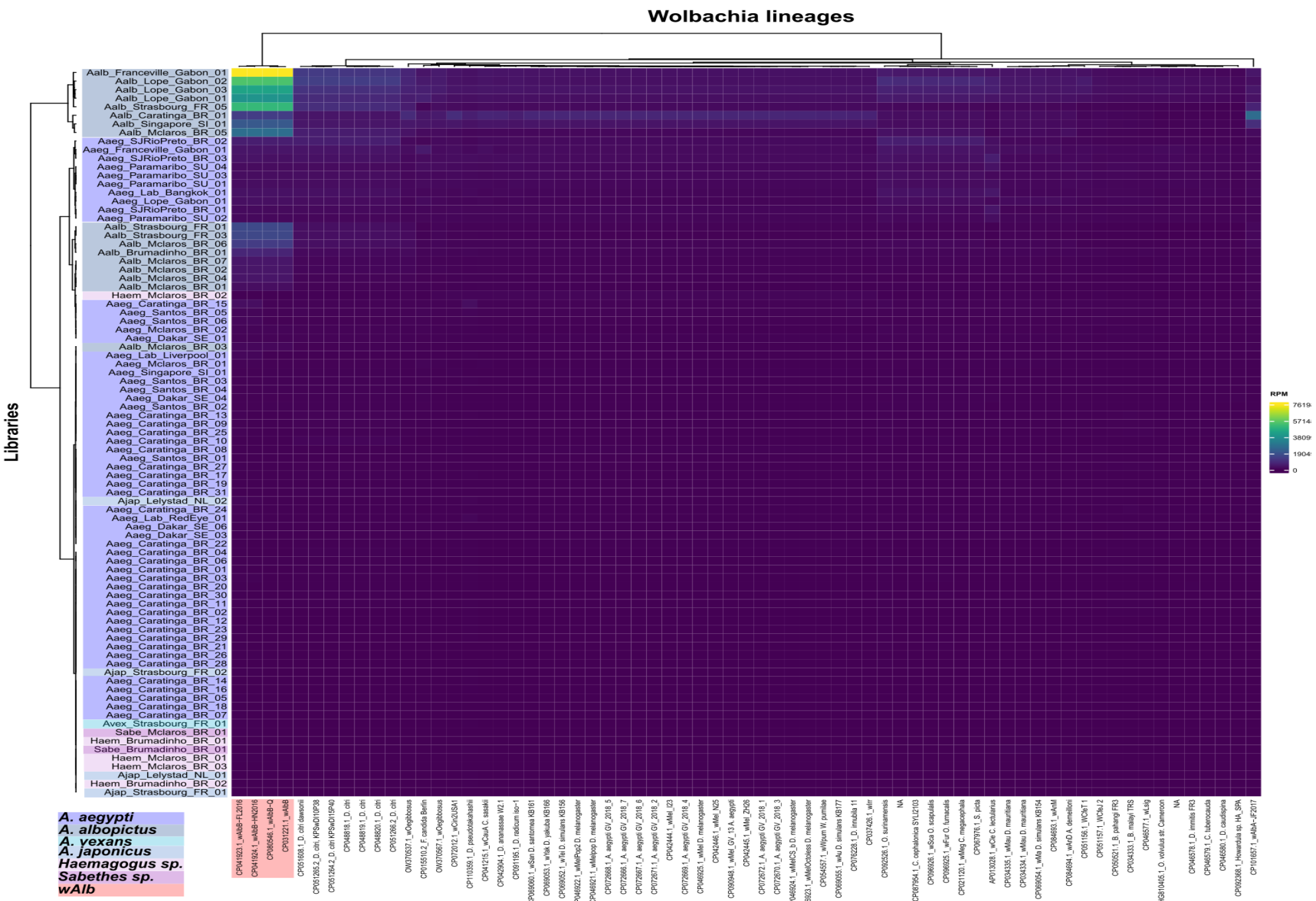
Cinco amostras de RNA de *A. albopictus* passaram por um protocolo de oxidação por periodato de sódio ($NaIO_4$) antes do sequenciamento. Tal tratamento enriquece as bibliotecas para pequenos RNAs com a modificação metil 2'O comumente associados a proteína Argonauta. A reação de beta-eliminação irá promover a degradação de RNAs com a terminação 2'OH livre. Consequentemente, esse tratamento reduz a quantidade de pequenos RNAs derivados de processos de degradação enzimática na amostra. Ao compararmos bibliotecas oxidadas e as controle não-oxidadas de amostras da França, podemos observar que a oxidação removeu quase por completo as reads alinhadas aos genomas de *wAlb* (**Figura 43**).

5.6.3 - Detecção de pequenos RNAs de *Wolbachia* nas bibliotecas não oxidadas de mosquitos de campo

Para detecção de pequenos RNAs de *Wolbachia* nas demais bibliotecas de mosquitos de campo, utilizamos apenas as bibliotecas que não passaram por processo de oxidação, restando 85 bibliotecas de amostras de seis espécies (**Figura 44**). Apenas bibliotecas de *A. albopictus* apresentam quantidades expressivas de alinhamentos contra a referência de genomas de *Wolbachia*s.

Para a **Figura 44** a legenda é mostrada antes da Figura, que se encontra na próxima página

Figura 44 - Detecção de pequenos RNAs de *Wolbachia* nas bibliotecas não oxidadas de mosquitos de campo. 85 bibliotecas de pequenos RNAs preparadas de amostras de 6 espécies de mosquitos que não passaram por processo de oxidação prévia ao sequenciamento foram alinhadas a referência contendo 67 genomas de *Wolbachia*s. Bibliotecas das mesmas espécies de mosquito foram destacadas com as mesmas cores. Os genomas de *Wolbachia*s da linhagem *wAlb* estão destacadas de vermelho. A quantidade de reads alinhadas aos genomas de *Wolbachia*s foram normalizadas por RPM.


Wolbachia lineages
Libraries
Aalb_Franceville_Gabon_01
Aalb_Lope_Gabon_02
Aalb_Lope_Gabon_03
Aalb_Lope_Gabon_01
Aalb_Strasbourg_FR_05
Aalb_Caratinga_BR_01
Aalb_Singapore_SI_01
Aalb_Mclaros_BR_05
Aaeg_SJRioPreto_BR_02
Aaeg_Franceville_Gabon_01
Aaeg_SJRioPreto_BR_03
Aaeg_Paramaribo_SU_04
Aaeg_Paramaribo_SU_03
Aaeg_Paramaribo_SU_01
Aaeg_Lab_Bangkok_01
Aaeg_Lope_Gabon_01
Aaeg_SJRioPreto_BR_01
Aaeg_Paramaribo_SU_02
Aalb_Strasbourg_FR_01
Aalb_Strasbourg_FR_03
Aalb_Mclaros_BR_06
Aalb_Brumadinho_BR_01
Aalb_Mclaros_BR_07
Aalb_Mclaros_BR_02
Aalb_Mclaros_BR_04
Aalb_Mclaros_BR_01
Haem_Mclaros_BR_02
Aaeg_Caratinga_BR_15
Aaeg_Santos_BR_05
Aaeg_Santos_BR_06
Aaeg_Mclaros_BR_02
Aaeg_Dakar_SE_01
Aalb_Mclaros_BR_03
Aaeg_Lab_Liverpool_01
Aaeg_Mclaros_BR_01
Aaeg_Singapore_SI_01
Aaeg_Santos_BR_03
Aaeg_Santos_BR_04
Aaeg_Dakar_SE_04
Aaeg_Santos_BR_02
Aaeg_Caratinga_BR_13
Aaeg_Caratinga_BR_09
Aaeg_Caratinga_BR_25
Aaeg_Caratinga_BR_10
Aaeg_Caratinga_BR_08
Aaeg_Santos_BR_01
Aaeg_Caratinga_BR_27
Aaeg_Caratinga_BR_17
Aaeg_Caratinga_BR_19
Aaeg_Caratinga_BR_31
Ajap_Lelystad_NL_02
Aaeg_Caratinga_BR_24
Aaeg_Lab_RedEye_01
Aaeg_Dakar_SE_06
Aaeg_Dakar_SE_03
Aaeg_Caratinga_BR_22
Aaeg_Caratinga_BR_04
Aaeg_Caratinga_BR_06
Aaeg_Caratinga_BR_01
Aaeg_Caratinga_BR_03
Aaeg_Caratinga_BR_20
Aaeg_Caratinga_BR_30
Aaeg_Caratinga_BR_11
Aaeg_Caratinga_BR_02
Aaeg_Caratinga_BR_12
Aaeg_Caratinga_BR_23
Aaeg_Caratinga_BR_29
Aaeg_Caratinga_BR_21
Aaeg_Caratinga_BR_26
Aaeg_Caratinga_BR_28
Ajap_Strasbourg_FR_02
Aaeg_Caratinga_BR_14
Aaeg_Caratinga_BR_16
Aaeg_Caratinga_BR_05
Aaeg_Caratinga_BR_18
Aaeg_Caratinga_BR_07
Avex_Strasbourg_FR_01
Sabe_Mclaros_BR_01
Haem_Brumadinho_BR_01
Sabe_Brumadinho_BR_01
Haem_Mclaros_BR_01
Haem_Mclaros_BR_03
Ajap_Lelystad_NL_01
Haem_Brumadinho_BR_02
Ajap_Strasbourg_FR_01
RPM
7619
5714
3809
1904
0
A. aegypti
A. albopictus
A. vexans
A. japonicus
Haemagogus sp.
Sabethes sp.
wAlb
CP041923.1_wAlbB-FL2016
CP041924.1_wAlbB-HN2016
CP080546.1_wAlbB-Q
CP031221.1_wAlbB
CP051608.1_D. citri dawsonii
CP051265.2_D. citri, KPSwDi10P38
CP051264.2_D. citri KPSwDi15P40
CP048818.1_D. citri
CP048819.1_D. citri
CP048820.1_D. citri
CP051266.2_D. citri
OW376537.1_wOegibbosus
CP015510.2_F. candida Berlin
OW370667.1_wOegibbosus
CP072012.1_wCin2USA1
CP110359.1_D. pseudotakahashii
CP041215.1_wCauA C. sasakii
CP042904.1_D. ananassae W2.1
CP091195.1_D. radicum iso-1
CP069060.1_wSan D. santomea KB161
CP069053.1_wYak D. yakuba KB166
CP069032.1_wTai D. simulans KB156
CP046922.1_wMelPop2 D. melanogaster
CP046921.1_wMelpop D. melanogaster
CP072668.1_A. aegypti GV_2018_5
CP072666.1_A. aegypti GV_2018_7
CP072667.1_A. aegypti GV_2018_6
CP072671.1_A. aegypti GV_2018_2
CP042444.1_wMel_I23
CP072669.1_A. aegypti GV_2018_4
CP046925.1_wMel D. melanogaster
CP042446.1_wMel_N25
CP090948.1_wMel_GV_13 A. aegypti
CP042445.1_wMel_ZH26
CP072672.1_A. aegypti GV_2018_1
CP072670.1_A. aegypti GV_2018_3
CP046924.1_wMelCS_b D. melanogaster
CP046923.1_wMelOctoless D. melanogaster
CP054557.1_wWpum W. pumilae
CP069055.1_wAu D. simulans KB177
CP076228.1_D. innubila 11
CP037426.1_wIrr
CP092526.1_O. suriamensis
NA
CP087954.1_C. cephalonica SYLZ103
CP096926.1_wSca O. scapularis
CP096925.1_wFur O. furcatalis
CP021120.1_wMeg C. megacephala
CP087976.1_S. picta
AP013028.1_wCle C. lectularius
CP034335.1_wMau D. mauritiana
CP034334.1_wMau D. mauritiana
CP069054.1_wMa D. simulans KB154
CP084693.1_wAnM
CP084694.1_wAnD A. demeilloni
CP051156.1_WCfeT 1
CP051157.1_WCfeJ 2
CP050521.1_B. pahangi FR3
CP034333.1_B. malayi TRS
CP046577.1_wLsig
HG810405.1_O. volvulus str. Cameroon
NA
CP046578.1_D. immitis FR3
CP046579.1_C. tuberocauda
CP046580.1_D. cautispina
CP092368.1_Howardula sp. HA_SPA
CP101657.1_wAbA-JF2017

5.6.4 - Detecção de pequenos RNAs de *Wolbachia* nas bibliotecas de mosquitos de Niterói – RJ

Concluímos o sequenciamento de 16 bibliotecas de *A. aegypti* e três de *A. albopictus* coletados na cidade de Niterói-RJ onde o projeto de Controle de *A.* aegypti por *Wolbachia* (**Tabela 14**) está em execução. Para a detecção de pequenos RNAs de *Wolbachia* nas bibliotecas de mosquitos de Niterói, alinhamos as bibliotecas com a referência de 67 genomas de *Wolbachia* (**Figura 45**). Podemos observar para as bibliotecas de *A. aegypti* da região controle (Zona5), Aaeg_16 e Aaeg_11 (cinza) uma quantidade vestigial de alinhamentos para a primeira e ausência de alinhamentos para a segunda. É possível também notar um padrão temporal de estabelecimento da infecção por *Wolbachia*s na população de *A. aegypti* de Niterói ao observarmos a quantificação de reads alinhados aos genomas de *Wolbachia* nas bibliotecas de 2018 comparado as bibliotecas de 2020 e 2021. Quanto a especificidade de linhagens de *Wolbachia* com quantidade expressiva de alinhamentos, apesar de alinhamentos ruidosos em outras linhagens, a quantidade de alinhamentos das bibliotecas de *A. aegypti* é expressivamente maior nas linhagens de *wMel* (verde) e de *Wolbachia*s similares que infectam outros Drosofilideos. As três bibliotecas de *A. albopictus* possuem um perfil claro de enriquecimento de alinhamentos nos genomas da linhagem *wAlb* (vermelho) e ausência de alinhamentos nos genomas das linhagens de *wMel*, incluindo a linhagem utilizada para infectar *A. aegypti* artificialmente.

Para a **Figura 45** a legenda é mostrada antes da Figura, que se encontra na próxima página

Figura 45 - Detecção de pequenos RNAs de *Wolbachia* em bibliotecas de campo de *A. aegypti* e *A. albopictus* de Niterói-RJ. 16 bibliotecas de *A. aegypti* e 3 bibliotecas de *A. albopictus* foram alinhadas contra a referência contendo 67 genomas de *Wolbachia*s. Nomes das bibliotecas possuem informações das “espécies | Zona | data de coleta | nro. de fêmeas na amostra”. Bibliotecas foram destacadas de acordo com a Zona de Niterói na qual a amostra foi obtida (Zona1 verde, Zona2 laranja, Zona3 amarelo, Zona4 azul, Zona5 cinza). A Zona 5 corresponde a região controle na qual não foram soltos *A. aegypti* artificialmente infectados por *wMel*. Os genomas de *Wolbachia*s da linhagem *wAlb* estão destacadas de vermelho e *wMel* em verde. A quantidade de reads alinhadas aos genomas de *Wolbachia*s foram normalizadas por RPM.

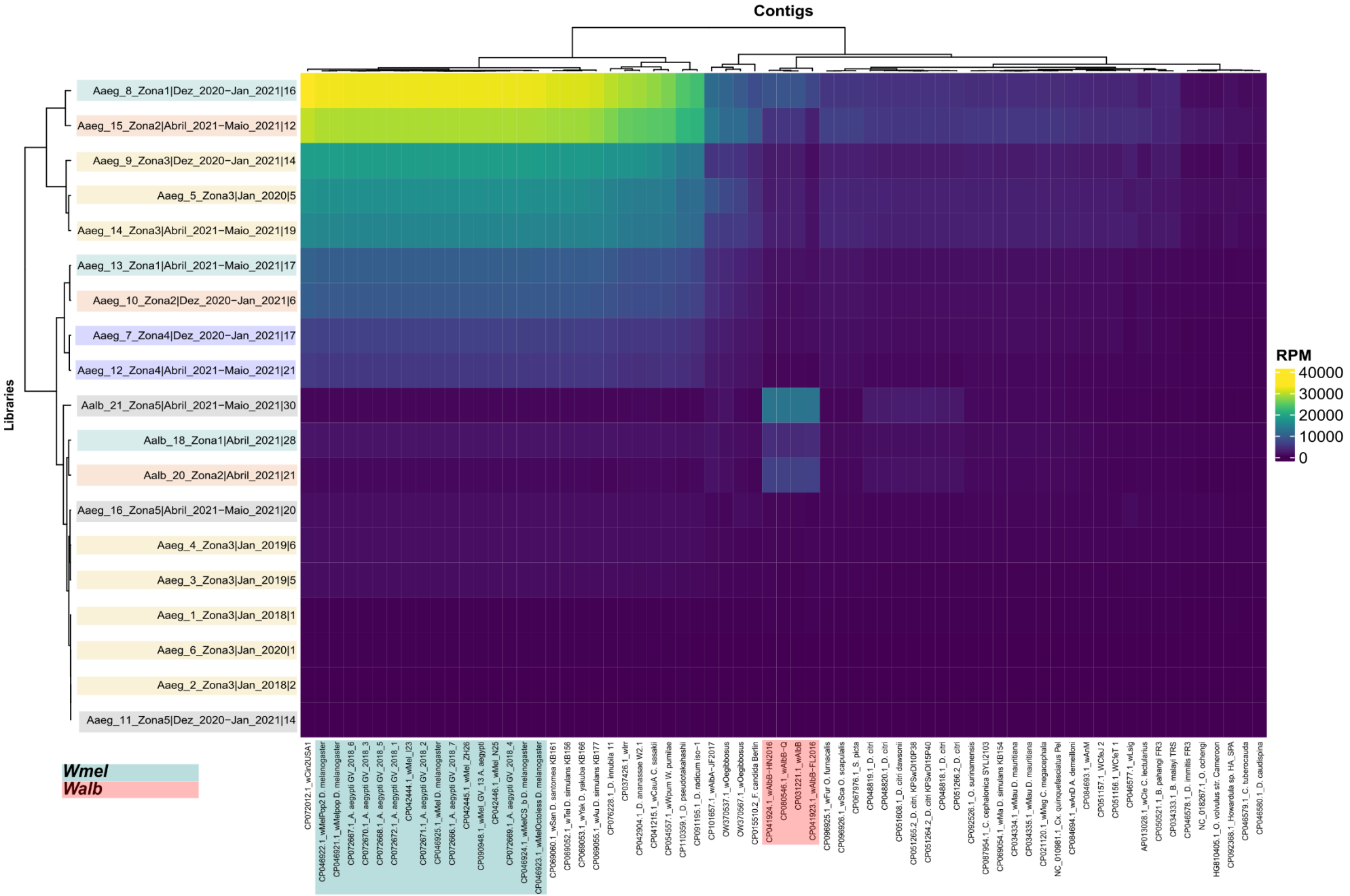

Contigs
Libraries
Aaeg_8_Zona1|Dez_2020−Jan_2021|16
Aaeg_15_Zona2|Abril_2021−Maio_2021|12
Aaeg_9_Zona3|Dez_2020−Jan_2021|14
Aaeg_5_Zona3|Jan_2020|5
Aaeg_14_Zona3|Abril_2021−Maio_2021|19
Aaeg_13_Zona1|Abril_2021−Maio_2021|17
Aaeg_10_Zona2|Dez_2020−Jan_2021|6
Aaeg_7_Zona4|Dez_2020−Jan_2021|17
Aaeg_12_Zona4|Abril_2021−Maio_2021|21
Aalb_21_Zona5|Abril_2021−Maio_2021|30
Aalb_18_Zona1|Abril_2021|28
Aalb_20_Zona2|Abril_2021|21
Aaeg_16_Zona5|Abril_2021−Maio_2021|20
Aaeg_4_Zona3|Jan_2019|6
Aaeg_3_Zona3|Jan_2019|5
Aaeg_1_Zona3|Jan_2018|1
Aaeg_6_Zona3|Jan_2020|1
Aaeg_2_Zona3|Jan_2018|2
Aaeg_11_Zona5|Dez_2020−Jan_2021|14
RPM
40000
30000
20000
10000
0
Wmel
Walb
CP072012.1_wCin2USA1
CP046922.1_wMelPop2 D. melanogaster
CP046921.1_wMelpop D. melanogaster
CP072667.1_A. aegypti GV_2018_6
CP072670.1_A. aegypti GV_2018_3
CP072668.1_A. aegypti GV_2018_5
CP072672.1_A. aegypti GV_2018_1
CP042444.1_wMel_I23
CP072671.1_A. aegypti GV_2018_2
CP046925.1_wMel D. melanogaster
CP072666.1_A. aegypti GV_2018_7
CP042445.1_wMel_ZH26
CP090948.1_wMel_GV_13 A. aegypti
CP042446.1_wMel_N25
CP072669.1_A. aegypti GV_2018_4
CP046924.1_wMelCS_b D. melanogaster
CP046923.1_wMelOctoless D. melanogaster
CP069060.1_wSan D. santomea KB161
CP069052.1_wTei D. simulans KB156
CP069053.1_wYak D. yakuba KB166
CP069055.1_wAu D. simulans KB177
CP076228.1_D. innubila 11
CP037426.1_wHr
CP042904.1_D. ananassae W2.1
CP041215.1_wCauA C. sasakii
CP054557.1_wWpum W. pumilae
CP110359.1_D. pseudotakahashii
CP091195.1_D. radicum iso−1
CP101657.1_wAlbA−JF2017
OW370537.1_wOegibosus
OW370567.1_wOegibbosus
CP015510.2_F. candida Berlin
CP041924.1_wAlbB−HN2016
CP080546.1_wAlbB−Q
CP031221.1_wAlbB
CP041923.1_wAlbB−FL2016
CP096925.1_wFur O. furnacalis
CP096926.1_wSca O. scapulalis
CP067976.1_S. picta
CP048819.1_D. citri
CP048820.1_D. citri
CP051608.1_D. citri dawsonii
CP051265.2_D. citri, KPSwDI10P38
CP051264.2_D. citri KPSwDI15P40
CP048818.1_D. citri
CP051266.2_D. citri
CP092526.1_O. surinamensis
CP087954.1_C. cephalonica SYLi2103
CP069054.1_wMa D. simulans KB154
CP034334.1_wMau D. mauritiana
CP034335.1_wMau D. mauritiana
CP021120.1_wMeg C. megacephala
NC_010981.1_Cx. quinquefasciatus Pel
CP084694.1_wAnD A. demeilloni
CP084693.1_wAnM
CP051157.1_WCleJ 2
CP051156.1_WCleT 1
CP046577.1_wLsig
AP013028.1_wCle C. lectularius
CP050521.1_B. pahangi FR3
CP034333.1_B. malayi TRS
CP046578.1_D. immitis FR3
NC_018267.1_O. ochengi
HG810405.1_O. volvulus str. Cameroon
CP092368.1_Howardula sp. HA_SPA
CP046579.1_C. tuberocauda
CP046580.1_D. caudispina

Tabela 14 - Dados de coletas, sequenciamento e contigs montados das bibliotecas de pequenos RNAs de mosquitos coletados em Niterói-RJ nas regiões de avaliação do projeto de controle de *A. aegypti* com *Wolbachia*;

Id_biblioteca	Zona	Periodo_coleta	Nro. de Femeas	reads		Contigs montados			
				brutas	trim.	não viral	viral	EVEs	Descon.
Aaeg_1	3	Jan_2018	1	2E+07	1E+07	28	18	2	5
Aaeg_2	3	Jan_2018	2	2E+07	2E+07	36	20	6	28
Aaeg_3	3	Jan_2019	5	2E+07	1E+07	14	21	1	4
Aaeg_4	3	Jan_2019	6	1E+07	7E+06	12	24	0	0
Aaeg_5	3	Jan_2020	5	2E+07	1E+07	14	46	2	1
Aaeg_6	3	Jan_2020	1	1E+07	1E+07	15	14	0	4
Aaeg_7	4	Dez_2020-Jan_2021	17	1E+07	1E+07	16	28	0	7
Aaeg_8	1	Dez_2020-Jan_2021	16	4E+06	4E+06	3	19	0	0
Aaeg_9	3	Dez_2020-Jan_2021	14	2E+07	8E+06	11	16	0	0
Aaeg_10	2	Dez_2020-Jan_2021	6	2E+07	1E+07	35	20	5	10
Aaeg_11	5	Dez_2020-Jan_2021	14	2E+07	8E+06	15	34	9	0
Aaeg_12	4	Abril_2021-Maio_202	21	1E+07	9E+06	17	15	0	3
Aaeg_13	1	Abril_2021-Maio_202	17	2E+07	1E+07	20	20	1	4
Aaeg_14	3	Abril_2021-Maio_202	19	2E+07	1E+07	23	34	7	5
Aaeg_15	2	Abril_2021-Maio_202	12	1E+07	1E+07	19	12	2	7
Aaeg_16	5	Abril_2021-Maio_202	20	2E+07	1E+07	18	30	7	0
Aalb_18	1	Abril_2021-Maio_202	28	2E+07	2E+07	32	2	0	19
Aalb_20	2	Abril_2021-Maio_202	21	1E+07	1E+07	32	8	8	23
Aalb_21	5	Abril_2021-Maio_202	30	1E+07	1E+07	35	4	4	27

5.6.5 - Análise do viroma dos mosquitos de Niterói – RJ

Após execução do nosso pipeline de análises de viroma, foi montado um total de 927 sequências maiores que 200nt divididas em 385 virais, 395 não virais e 147 desconhecidas (**Tabela 14**). Analisando o perfil de pequenos RNAs obtivemos 331 sequências oriundas de vírus exógenos e 54 de EVEs. Após remoção de redundâncias das sequências de vírus exógenos com o CDHIT, obtivemos 146 contigs não redundantes que foram utilizados para análise de coocorrência com a quantificação de pequenos RNAs de 20-22nt (siRNAs) de todas as bibliotecas (**Figura 46**). Podemos observar uma prevalência de 100% de sequências do vírus PCLV em todas as bibliotecas de *A. aegypti*. O vírus HTV também é altamente prevalente, ausente em apenas uma biblioteca, Aaeg_10_Zona3. Podemos determinar a presença de outros quatro vírus nas bibliotecas de *A. aegypti*, dois previamente detectados em bibliotecas de *A. aegypti* do Brasil nas bibliotecas do projeto ZikaAlliance: *Orbis virgavirus* (OVV) e *Guadeloupe mosquito vírus* (GMV) e duas novas detecções: *Guadeloupe mosquito quaranja-like virus* (GMqlV) e *Rio Chico vírus* (RCV). Nas três bibliotecas de *A. albopictus,* sequências isoladas de vírus de *A. aegypti* tiveram

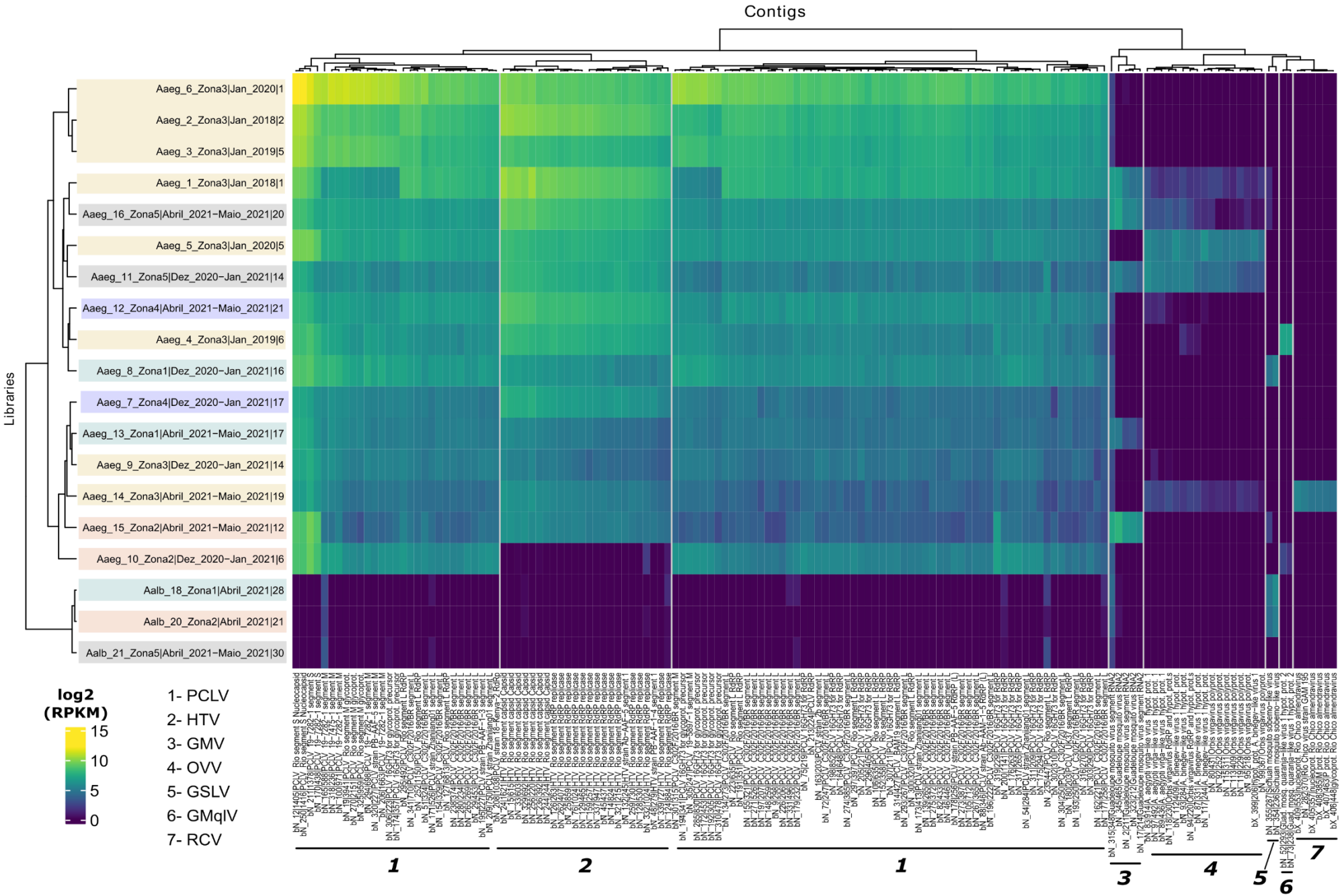

Contigs
Libraries
Aaeg_6_Zona3|Jan_2020|1
Aaeg_2_Zona3|Jan_2018|2
Aaeg_3_Zona3|Jan_2019|5
Aaeg_1_Zona3|Jan_2018|1
Aaeg_16_Zona5|Abril_2021-Maio_2021|20
Aaeg_5_Zona3|Jan_2020|5
Aaeg_11_Zona5|Dez_2020-Jan_2021|14
Aaeg_12_Zona4|Abril_2021-Maio_2021|21
Aaeg_4_Zona3|Jan_2019|6
Aaeg_8_Zona1|Dez_2020-Jan_2021|16
Aaeg_7_Zona4|Dez_2020-Jan_2021|17
Aaeg_13_Zona1|Abril_2021-Maio_2021|17
Aaeg_9_Zona3|Dez_2020-Jan_2021|14
Aaeg_14_Zona3|Abril_2021-Maio_2021|19
Aaeg_15_Zona2|Abril_2021-Maio_2021|12
Aaeg_10_Zona2|Dez_2020-Jan_2021|6
Aalb_18_Zona1|Abril_2021|28
Aalb_20_Zona2|Abril_2021|21
Aalb_21_Zona5|Abril_2021-Maio_2021|30
log2 (RPKM)
15
10
5
0
1- PCLV
2- HTV
3- GMV
4- OVV
5- GSLV
6- GMqlV
7- RCV
1
2
1
3
4
5
6
7

Figura 46 - Coocorrência dos 146 contigs virais não-redundantes nas bibliotecas de *A. aegypti* e *A. albopictus* de Niterói-RJ. O heatmap representa a contagem normalizada (RPKM) de pequenos RNAs de 20-22 nt alinhados a cada um dos 146 contigs virais não-redundantes (colunas) nas 19 bibliotecas analisadas (linhas). Nomes das bibliotecas possuem informações das “espécies | Zona | data de coleta | nro. de fêmeas na amostra”. Bibliotecas foram destacadas de acordo com a Zona de Niterói na qual a amostra foi obtida (Zona1 verde, Zona2 laranja, Zona3 amarelo, Zona4 azul, Zona5 cinza). A coocorrência dos contigs no mesmo agrupamento hierárquico em conjunto a curadoria dos resultados de alinhamentos locais de cada contig permitiram a definição de sete vírus únicos. Os dendrogramas apresentados tem como combinação de método e distância para agrupamentos das linhas e colunas "complete" com “euclidean”.

alinhamentos, porém, excluímos a hipótese de vírus compartilhados entre as duas espécies por termos posteriormente detectado a contaminação de 1 mosquito *A. aegypti* em meio a cada pool de *A. albopictus* ao realizarmos qPCR para diferenciação dessas duas especificas nas amostras de RNA separado de cada mosquito que compôs o pool sequenciado. Nossa hipótese é de que o único vírus detectado para *A. albopictus* é o *Guangzhou sobemo-like vírus* (GSLV), consistente com a detecção prévia das bibliotecas de *A. albopictus* do Brasil no projeto ZikaAlliance. Também detectamos posteriormente por qPCR das amostras de RNA restante dos mosquitos únicos que compõe o pool da biblioteca Aaeg_8_Zona1 um mosquito *A. albopictus* em meio ao pool de *A. aegypti*. O que explica a detecção de GSLV nessas amostras.

Todas as bibliotecas foram alinhadas contra uma referência de sequências de arbovírus e não foi possível detectar reads dos vírus presentas na referência.

Na **Figura 47** são mostrados os valores de RPKM das reads de 20-22nt (siRNAs) alinhadas aos contigs virais dos seis vírus detectados nas bibliotecas de *A. aegypti* agrupados por data e coloridos pela zona de coleta. Para os vírus PCLV e HTV podemos observar uma diminuição na média da carga de pequenos RNAs virais ao longo do tempo.

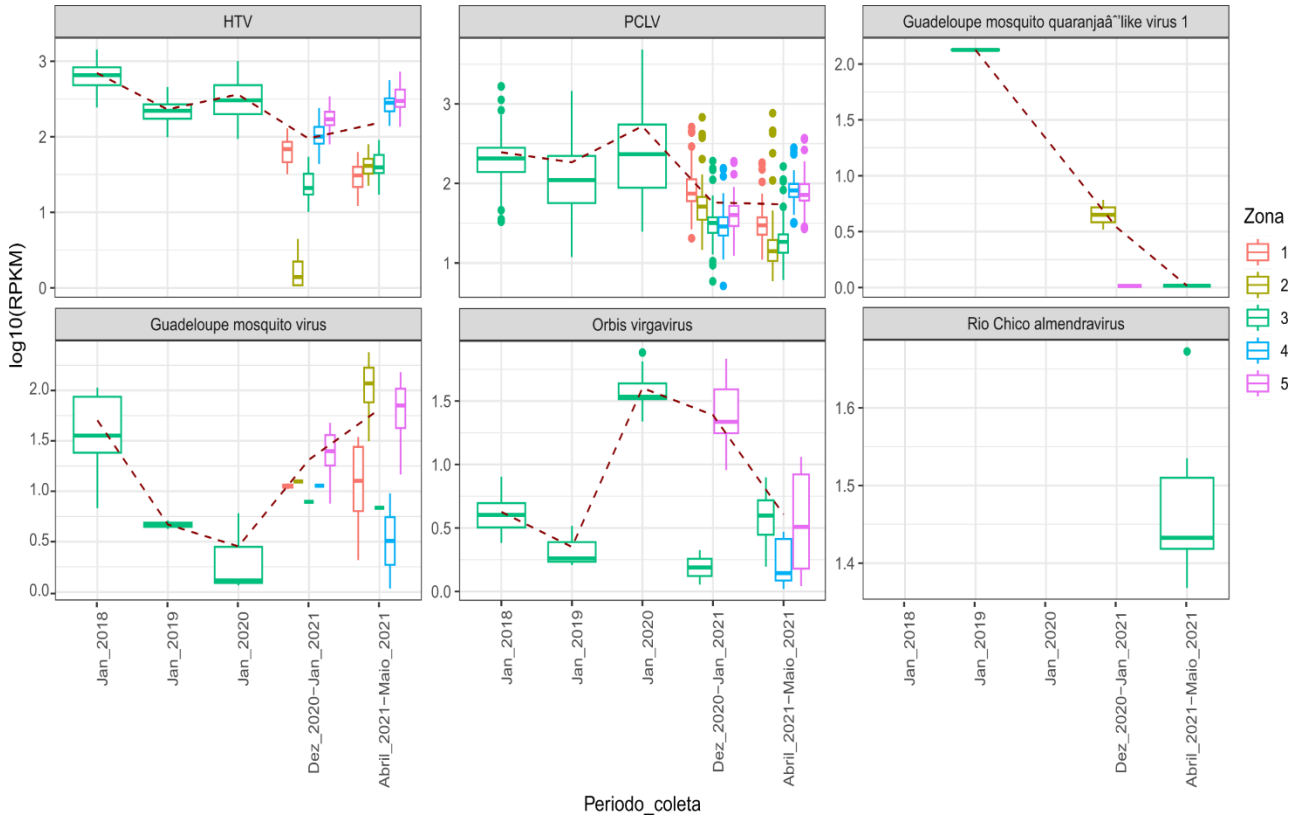


Figura 47 – Carga de pequenos RNAs dos seis vírus detectados em *A. aegypti* ao longo das datas de coletas em Niteroi-RJ. Valores de RPKM da quantidade de reads de 20-22nt alinhados aos contigs virais foram plotados como box-plots ao longo das datas de coleta. Linhas tracejadas indicam a tendência dos da média de RPKM da carga de pequenos RNAs virais ao longo das coletas.

5.6.6 - Análise do efeito da infecção por *wMel* na carga viral de ISVs

Na **Figura 48**, foram plotadas as médias das cargas de pequenos RNAs virais e a carga normalizada de pequenos RNAs de *wMel* ao longo do tempo e separados por Zona de coleta. Na Zona3, em que foram obtidos 5 pontos temporais de coleta, podemos notar um aumento na carga de pequenos RNAs de *wMel* concomitante a diminuição da carga de pequenos RNAs virais de HTV e PCLV.

Foram calculadas correlações entre as médias de RPKM das cargas de pequenos RNAs dos contigs virais de cada vírus e a carga de pequenos RNAs de *wMel* para cada biblioteca. Para os vírus PCLV e HTV, obtivemos uma correlação negativa e estatisticamente significativa (**Figura 49**) evidenciando que o aumento da carga de pequenos RNAs de *wMel* leva a uma redução na carga de pequenos RNAs desses ISVs.

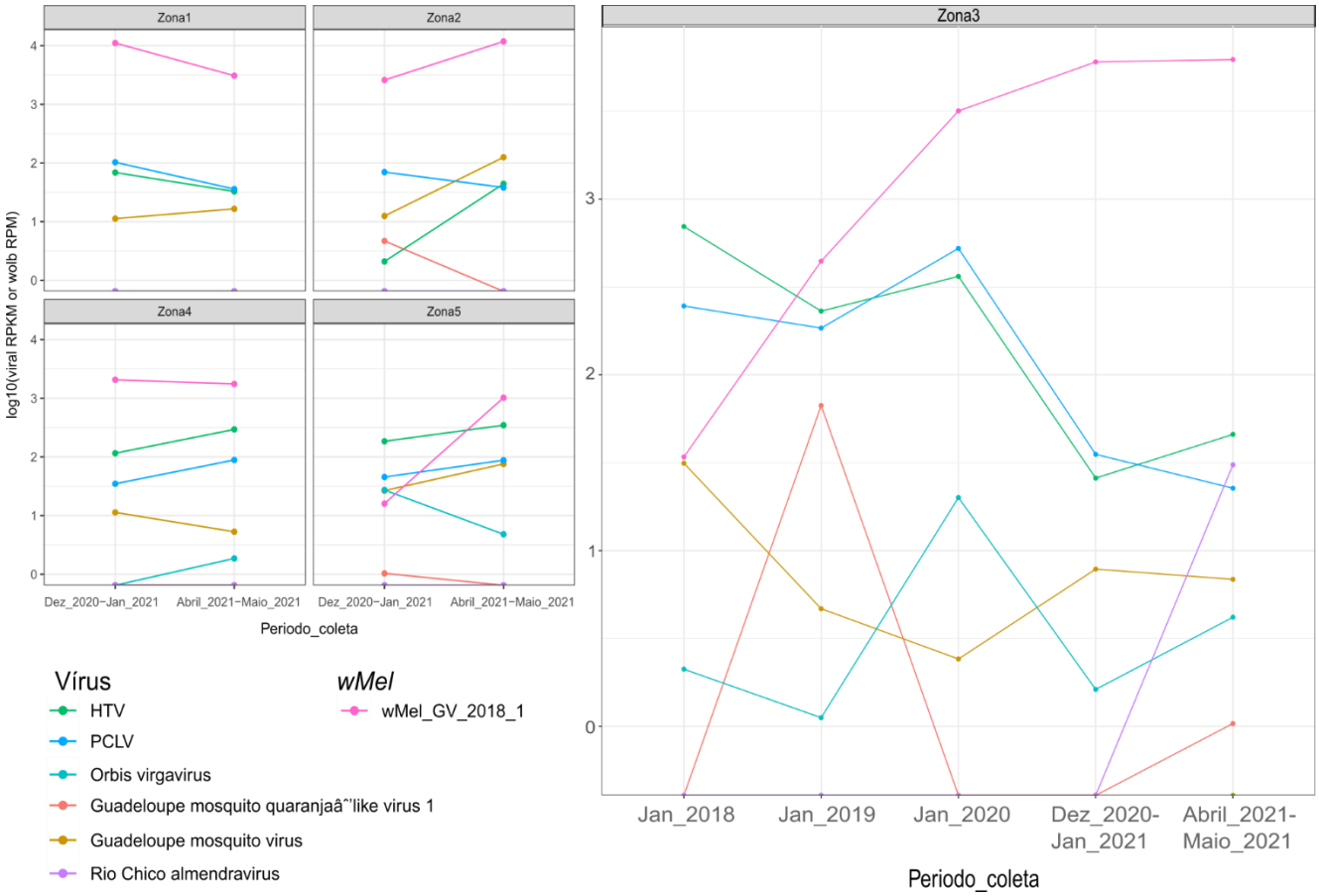


Figura 48 – Análise temporal da carga de pequenos RNA virais e de *Wolbachia* por Zona de coleta da cidade de Niterói-RJ. As cargas de pequenos RNAs virais e de *wMel* foram separadas por Zonas e as médias das cargas plotadas por data de coleta. A Zona3 é a única com mais de duas datas de coleta amostradas e permitiu visualizar uma tendência no aumento da carga de *wMel* (rosa) acompanhado por uma redução na carga dos ISVs PCLV (azul) e HTV (verde).

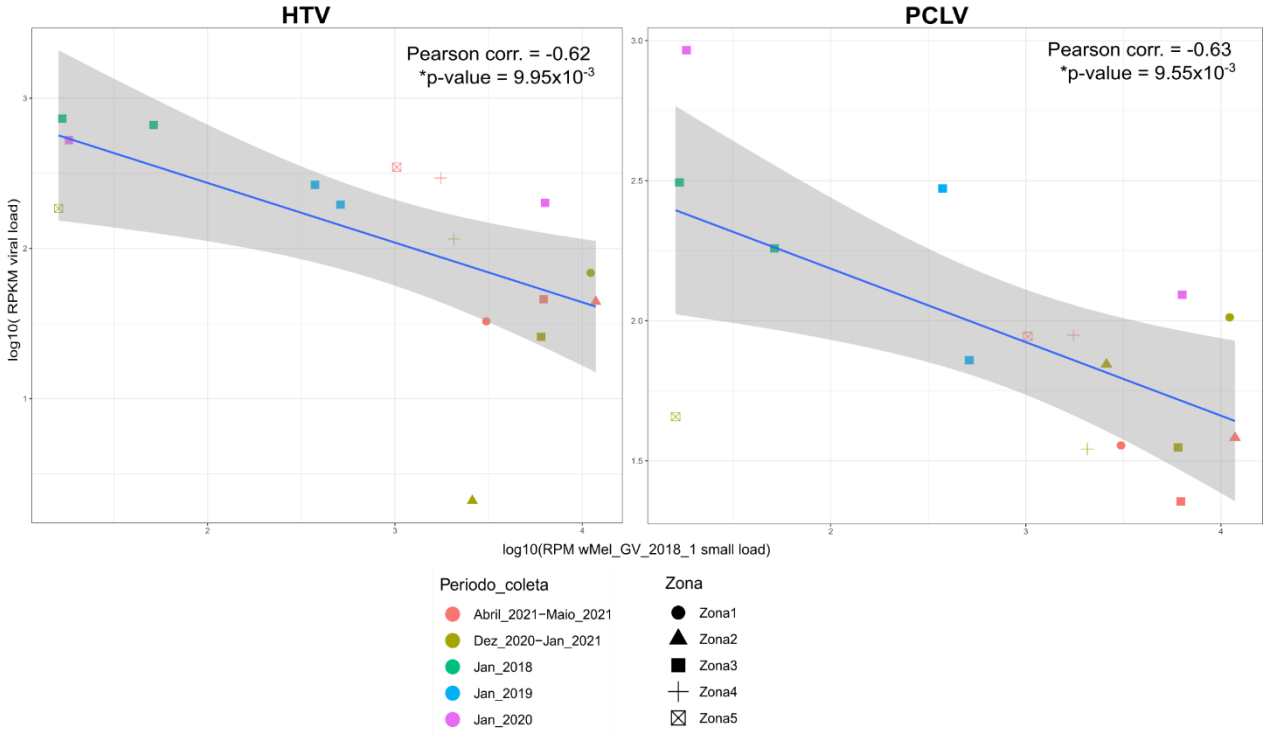


Figura 49 – Correlação negativa entre a carga de pequenos RNAs viral dos ISVs PCLV e HTV e de pequenos RNAs de *Wolbachia*. A relação das cargas de pequenos RNAs dos vírus (Y) com a carga de pequenos RNAs de *Wolbachia* foi plotada para todas as bibliotecas. Há uma correlação negativa e significativa (PCLV Pearson.corr = −0.63, *p−value = 0.00955 e HTV Pearson.corr = −0.62, *p−value = 0.00995) entre a carga de pequenos RNA virais dos ISVs PCLV e HTV com a carga de pequenos RNAs de *Wolbachia*

6 - Discussão

As análises das bibliotecas de pequenos RNAs das 10 espécies amostradas nesse estudo revelaram um cenário de alta complexidade e diversidade do viroma e EVEroma de mosquitos vetores. Com exceção de duas sequências com similaridade com vírus de DNA montadas em bibliotecas de *A. furcifer*, todas sequências virais detectadas nas bibliotecas de mosquitos desse estudo são oriundas de vírus com genomas de RNA. A detecção de vírus de DNA usando abordagens de RNA-seq não é uma limitação pois, vírus de DNA produzem diferentes moléculas de RNA durante seu ciclo de replicação (DE FARIA et al., 2022). Vírus de DNA têm sido identificados com sucesso em mosquitos por meio de abordagens de metatranscriptômica baseadas em sequenciamento de longos e pequenos RNAs (FENG et al., 2022; MA et al., 2011). De uma perspectiva mais ampla, os vírus de DNA parecem estar sub-representados nos viromas de mosquitos e em eucariotos como um todo (DE ALMEIDA et al., 2021; KRUPOVIC; DOLJA; KOONIN, 2023). Portanto, é provável que vírus de DNA estejam, de fato, ausentes nos mosquitos analisados nesse estudo e tal resultado não é devido às limitações do método empregado.

A detecção de arbovírus em estudos metagenômicos com mosquitos de campo sem enriquecimento prévio de amplicons alvo por PCRs se trata de um evento raro. A dificuldade na detecção é atribuída a baixa prevalência e carga viral de arbovírus em mosquitos de campo (OLMO et al., 2023). O único arbovírus detectado nesse estudo, YFV em uma biblioteca de mosquitos *Haemagogus sp*. coletados em Brumadinho, apresentou baixa carga viral na detecção confirmada por qPCR no RNA que deu origem a biblioteca. A baixa carga viral, provavelmente, influenciou em uma montagem apenas de contigs pequenos para esse vírus. O perfil de pequenos RNAs dos contigs de YFV não apresentou características da siRNAs, mas sim um perfil de cobertura assimétrica na fita senso com pequenos RNAs de vários tamanhos, uma evidência de degradação de RNA ou inibição da via de siRNA (AGUIAR; OLMO; MARQUES, 2016). A possibilidade de se tratar de degradação de RNA da amostra é baixa, pois os resultados das corridas de Bioanalyzer das amostras de RNAs dos mosquitos que compõe a amostra sequenciada foram revisitados e todos apresentam alto índice RIN (*RNA Integrity Number*). Há evidências de que a proteína do capsídeo de YFV pode interferir na atividade da Dicer em *A. aegypti*, inibindo a via de RNAi (SAMUEL et al., 2016). O perfil de pequenos RNAs que observamos para o contig

representativo de YFV é similar ao gerado pelo *Flock House vírus* que codifica uma proteína chamada B2 que inibibi a via de RNAi em drosófilas (AGUIAR; OLMO; MARQUES, 2016). Se estamos observando um fenômeno de bloqueio da via de RNAi por YFV em *Haemagogus spp.* se trata de uma especulação pois temos apenas uma observação. Porém, esse resultado abre precedentes para estudos da via de RNAi em resposta a arbovírus nesse importante vetor nativo de YFV.

A baixa prevalência de arbovírus em populações de mosquitos nos permite inferir que esses vírus não geram a principal força evolutiva que molda o sistema imune antiviral desses insetos. Por outro lado, a alta prevalência e carga viral dos ISVs fazem desses vírus residentes componentes biológicos relevantes para compreendermos aspectos evolutivos da resposta imune de mosquitos. Todos os demais vírus além de YFV detectados nesse estudo são potenciais ISVs baseado nos resultados de similaridade de sequência e dados da literatura de vírus próximos. Para PCLV, HTV, CFAV e AAV temos evidências definitivas de que se tratam de ISVs por esses vírus não replicarem em hospedeiros vertebrados ou culturas de células de hospedeiros vertebrados (OLMO et al., 2023; PARRY; ASGARI, 2018; STOLLAR; THOMAS, 1975). Para os demais vírus, a classificação definitiva como ISVs dependerá de evidências experimentais. Nossas análises de coocorrência de sequências virais evidenciam que todos os vírus detectados nesse estudo são espécie-específicos, resultado coerente com o fato de que a transmissão de ISVs provavelmente ocorre apenas na mesma espécie de forma vertical, da fêmea para a prole (BOLLING et al., 2012) ou venérea, entre machos e fêmeas (MAVALE et al., 2005).

Ae. aegypti, o principal vetor de arbovírus no planeta, foi a espécie em que detectamos mais vírus. Foram detectados 10 vírus pertencendo a no mínimo 9 famílias virais diferentes. O viés de amostragem para essa espécie deve ser considerado ao interpretar esse resultado, foram sequenciadas 76 bibliotecas dessa espécie. Em *A. albopictus*, também considerado um dos vetores mais importantes de arbovírus no planeta, detectamos apenas dois vírus. Mesmo considerando a discrepância de amostragem*, A. albopictus* (22 bibliotecas) parece ter um viroma menos diverso que *A. aegypti*. Uma única espécie de vírus foi detectada em cada população de *A. albopictus* se considerarmos a divisão América do Sul *e* África/Europa, enquanto 4–6 vírus foram detectados nessas regiões em *A. aegypti*. Também podemos comparar com a espécie invasora *A. japonicus*, que com apenas quatro bibliotecas de 2

populações encontramos 6 vírus. Uma hipótese para explicar o viroma residente menos diverso de *A. albopictus* é o fato de essa ser uma das poucas espécies de mosquitos naturalmente infectadas por *Wolbachia* que pode conferir algum tipo de proteção antiviral. De fato, em nossas análises de detecção de pequenos RNAs de Wolbachia, *A. albopictus* é a única espécie em que detectamos a infecções naturais da bactéria em mosquitos de campo. Todos os vírus encontrados em *A. aegypti* e *A. albopictus* foram confirmados por PCR nos experimentos da nossa recente publicação envolvendo o viroma desses dois mosquitos (OLMO et al., 2023). Além dessas validações, outro resultado experimental importante para o estabelecimento da nossa estratégia metagenômica com pequenos RNAs foi o de que a carga viral determinada pela abundância de pequenos RNAs correspondeu à detecção por qRT–PCR para os vírus detectados em *A. aegypti* e *A. albopictus*.

Dentre os vírus detectados em *A. aegypti* os ISVs HTV e PCLV são os vírus mais prevalentes e que atingem as maiores cargas de pequenos RNAs virais nas bibliotecas analisadas. Nosso grupo mostrou que altas cargas de HTV e PCLV são encontradas em mosquitos coletados em áreas com alta incidência de DENV/ZIKV (OLMO et al., 2023). Em experimentos de laboratório, a infecção por HTV e PCLV levou a um aumento da replicação de arbovírus em mosquitos. Além disso, a infecção por esses ISVs também reduziu o período de incubação extrínseca dos mosquitos para arbovírus, o que faz com que os mosquitos sejam capazes de transmitir os vírus mais cedo após adquiri-los do sangue de camundongos infectados. Esses resultados foram surpreendentes, pois um estudo prévio mostrou que CFAV reduz a disseminação de arbovírus em *A. aegypti*, fenômeno provavelmente associado a um efeito chamado de “exclusão por superinfecção” (BAIDALIUK et al., 2019). A descoberta do nosso grupo sobre os efeitos de HTV e PCLV na transmissão de arbovírus evidencia a importância dos ISVs e da estratégia de análise de viroma em larga escala para a prospecção de vírus que podem impactar a competência vetorial de mosquitos vetores.

No trabalho de AGUIAR et al., 2015, os autores mostraram que a estratégia de pequenos RNAs para análise de viromas tem como principal vantagem o fato do sequenciamento ser concentrado em uma porção de RNAs naturalmente enriquecida para sequências virais. Isso torna o preparo das bibliotecas menos custoso e laborioso, pois não necessita protocolos de filtragem para concentração de partículas virais ou depleção de RNAs ribossomais. Os autores também mostram que para os

vírus detectados em insetos no trabalho a montagem e cobertura de contigs com pequenos RNAs é melhor que a de longos RNAs. Porém, é preciso pontuar que uma potencial desvantagem dos pequenos RNAs para montagem de sequências é a sua limitação intrínseca em relação a extensão de montagem de contigs quando os algoritmos heurísticos com grafos de *de Bruijn* são usados (COMPEAU; PEVZNER; TESLER, 2011). O intervalo de tamanho de reads 15-50nt limita os possíveis *k-mers* a serem utilizados para estabelecimentos dos “nós” dos grafos que são determinantes da precisão e extensão das montagens. Essa característica intrínseca de nossos dados brutos pode explicar a grande quantidade de pequenos contigs obtidos. Em nossa ferramenta *small RNA Metavir*, tomamos o cuidado de combinar dois montadores, *SPAdes* e *Velvet* (BANKEVICH et al., 2012; ZERBINO; BIRNEY, 2008), que executam várias rodadas de montagens combinando diferentes intervalos de tamanho de reads e *k-mers* e, ao final, consolidamos os resultados de diversas combinações de montagens com o programa CAP3 (HUANG; MADAN, 1999). Além disso a avaliação dos gráficos de cobertura por pequenos RNAs que representam siRNAs (20-22nt) nos permite conferir potenciais sequências “quimeras” montadas. Assim como no trabalho de AGUIAR et al., 2015, a potencial limitação da extensão de contigs não nos impediu de obter genomas virais completos como o de AejapTV1, no qual a qualidade da montagem permitiu a análise de detalhes estruturais relacionados a estratégia de f*rameshifting* ribossomal desse vírus. É importante ressaltar que o uso de RNAs longos, e até mesmo de sequenciamento de amplicons virais previamente enriquecidos por PCR não são garantias de obtenção de genomas virais completos. A carga viral na amostra será somada a profundidade do sequenciamento serão fatores determinantes do sucesso da montagem de genomas virais. Em um cenário ideal, a combinação do sequenciamento de pequenos RNAs para inferência da resposta imune de RNAi do mosquito hospedeiro com o sequenciamento de bibliotecas *paired-end* de RNAs longos para conferir e garantir a extensão das montagens, sem dúvidas, renderia resultados ótimos para explorar o viroma de mosquitos. Porém a realidade dos custos de sequenciamento impossibilita esse cenário para muitos grupos de pesquisa. Em nossos trabalhos, temos mostrado que a estratégia de sequenciamento de pequenos RNAs oferece bons custos-benefícios em relação ao preparo de bibliotecas e processamento de dados para o estudo do viroma de mosquitos vetores (ABBO et al., 2023; AGUIAR et al., 2015; OLMO et al., 2023).

Dentre os vírus identificados nesse estudo, os Narnavirus são exemplos do poder de resolução da nossa estratégia de metagenômica viral com pequenos RNAs. Vírus da família *Narnaviridae* (https://ictv.global/report_9th/RNApos/Narnaviridae) foram inicialmente identificados em fungos. Esses vírus não codificam proteínas de capsídeo, porém, possuem RdRPs homólogas a vírus. Recentemente, os relatos da presença desses vírus em outros organismos aumentaram com o emprego de estratégias metagenômicas (KOONIN et al., 2021). Porém, poucos autores tomam o cuidado de salientar algo importante, em dados metagenômicos não é possível saber se esses vírus estão infectando o organismo estudado ou são oriundos de contaminações por fungos nas amostras sequenciadas. O perfil de pequenos RNAs gerados pelas vias de RNAi de fungos é diferente do padrão de 21nt dsRNAs observados em mosquitos (DRINNENBERG et al., 2009; VAINIO et al., 2015). A diferença em relação as vias de RNAi chega ao extremo de a levedura *Saccharomyces cerevisae,* que dá o nome a espécie tipo de Narnavirus no ICTV (*Saccharomyces 20S RNA narnavirus*) não possuir vias de RNAi (DRINNENBERG et al., 2009). Portanto, o claro perfil de siRNAs com 21nt dsRNAs para os vírus *Aslam narnavirus* infectando *Ae. aegypti* e AejapNV1 infectando *Ae. japonicus,* são fortes evidências de que esses vírus estão replicando nesses mosquitos.

Para o Narnavirus, AejapNV1, a montagem *de novo* utilizando pequenos RNAs não apenas resultou no segmento primário completo do genoma de aproximadamente 3 kb que codifica ORFs ambigramáticas sendo a da fita senso a proteína RdRP de AejapNV1, mas também revelou a presença de um segmento genômico secundário com cerca de 1 kb, também ambigramático, mas sem similaridade significativa com sequências presentes no Genbank. Podemos associar com alta confiança o novo segmentoi descoberto a AejapNV1. Os perfis de abundância e distribuição de tamanho de pequenos RNAs de ambos os segmentos é similar. Evidências adicionais de que esses segmentos pertencem ao mesmo vírus são as ORFs ambigramáticas e as estruturas de RNA conservadas nas extremidades 3', ambas características genômicas peculiares de Narnavirus. Narnavírus bisegmentados também foram descobertos em *Plasmodium* (CHARON et al., 2019), em um tripanossomatídeo (GRYBCHUK et al., 2018; LYE et al., 2016), e recentemente, em mosquitos (BATSON et al., 2021; DUDAS et al., 2021). Para o CxNV1, um narnavírus bisegmentado e ambigramático associado a mosquitos do gênero *Culex*, descobriu-se que a presença do segmento primário RdRp era necessária para a replicação do segmento secundário

em cultura de células, enquanto o segmento primário poderia persistir na mesma sem a presença do segmento secundário (RETALLACK et al., 2021). Em nosso estudo, os segmentos primário e secundário do genoma do AejapNV1 coocorreram nas quatro bibliotecas de *Ae. japonicus*. Os resultados dos PCRs realizados pela nossa colaboradora, Dra. Sandra Abbo (Universidade de Wageningen, Holanda), mostram que os dois segmentos coocorrem com alta prevalência em fêmeas, machos, larvas e ovos de *Ae. japonicus* coletados em campo (ABBO et al., 2023). Tal resultado experimental é coerente com descobertas anteriores para CxNV1 em mosquitos capturados na natureza (BATSON et al., 2021), sugerindo que a manutenção do vírus em mosquitos de campo pode depender da presença de ambos os segmentos desses Narnavírus.

Apesar de algumas características conservadas entre AejapNV1 S2 e CxNV1 S2, como as regiões UTR e a estrutura ambigramática das ORFs, essas sequências são altamente divergentes em nível de nt e aa. Essa divergência que beira a impossibilidade de inferência de homologia levanta questões sobre a origem e função desse segmento genômico secundário nesses Narnavirus. Analisamos a conservação entre as proteínas codificadas por AejapNV1 S2 e CxNV1 S2 a nível de estruturas. Foram observados pequenos motivos conservados nos níveis de estruturas secundárias e terciárias entre as proteínas codificadas pelas fORFs ao comparar a presença e sintenia das regiões preditas de α-hélice e as regiões estruturadas preditas pelo AlphaFold. No entanto, a função do segmento secundário do genoma de AejapNV1 e CxNV1 permanece desconhecida, e estudos futuros são necessários para elucidar seu papel e origem evolutiva. Esses segmentos secundários de Narnavírus pertencem à chamada "matéria escura" da metagenômica, visto que métodos de similaridade de sequência, como alinhamentos locais, não foram capazes de associar a sequência de AejapNV1 S2 à referência CxNV1 S2 no GenBank ou a qualquer outra sequência dessa base de dados. Mesmo a aplicação da poderosa ferramenta AlphaFold (JUMPER et al., 2021) para prever estruturas terciárias e inferir função proteica foi limitada pela falta de sequências similares as ORFs em AejapNV1 S2. O AlphaFold depende de alinhamentos iniciais com bancos de dados de sequência e estrutura de proteínas, como Uniref90 e o PDB, para adquirir dados tridimensionais atômicos e realizar avaliações estruturais. Isso pode explicar a falta de predições confiáveis para as fORFs dos segmentos 2 e para qualquer tentativa de prever estruturas para sequências altamente divergentes. Este é um exemplo de como

as ferramentas de bioinformática para identificação de sequências virais em dados metagenômicos ainda têm espaço para melhorias. Estudos que abordem a "matéria escura" da metagenômica viral são fundamentais para permitir generalizações automatizadas em larga escala sobre a evolução e função de sequências virais divergentes. Nossa abordagem de agrupamento baseada em perfis de pequenos RNAs e coocorrência foi crucial para identificar o contig desconhecido correspondente ao segmento 2 de AejapNV1. Isso reforça o potencial da nossa estratégia baseada em pequenos RNAs para recuperar sequências virais da "matéria escura" da metagenômica viral.

A origem de uma parte substancial de contigs “desconhecidos” presentes nas bibliotecas de *Ae. japonicus* e nas demais bibliotecas de mosquitos permanece enigmática. Apesar da presença de um perfil de siRNA, não pudemos associar claramente 21 contigs desconhecidos de *Ae. japonicus* a nenhuma espécie viral definida neste estudo. Dentre os 14645 contigs desconhecidos obtidos em todas as 122 bibliotecas do projeto ZikaAlliance, 736 apresentam perfil de siRNA e foram classificados como virais pela nossa ferramenta *small RNA Metavir*. Porém, diferente das evidências que conseguimos acumular para afirmar que AejapNV1 S2 é um exemplo da matéria escura viral, as análises iniciais realizadas até o momento não evidenciam que há mais contigs virais nesse grupo de sequências. Porém, comparado ao aprofundamento de análises no viroma de *Ae. japonicus com* apenas quatro bibliotecas, pouca atenção foi dada a essas 14 mil sequência e analisá-las melhor faz parte dos próximos passos dos projetos em andamento no laboratório. Também serão necessárias mais análises para determinar se sequências desconhecidas com perfil de siRNA são sempre de origem viral. Uma explicação alternativa para sequências desconhecidas gerando siRNAs deriva da falta de genomas de referência disponível para a maioria das espécies de mosquitos, implicando que alguns desses contigs desconhecidos podem ter se originado de elementos repetitivos desconhecidos no genoma e/ou fragmentos de mRNA sobrepostos que podem produzir siRNAs endógenos (GHILDIYAL et al., 2008). Até o momento, o classificador embarcado em nossa ferramenta *small RNA Metavir* foi treinado com um conjunto de dados de sequências que inicialmente tiveram similaridade significativa com sequências virais e posteriormente foram separadas em “EVEs” e “Virais”. Para que a ferramenta possa ser efetivamente usada para explorar sequências desconhecidas em bibliotecas de pequenos RNAs iremos incluir no treino dos próximos classificadores sequências de

elementos transponíveis e repetitivos gerando perfis similares aos de siRNAs virais como um dos próximos passos de desenvolvimento. Nessa tese, foram processadas mais de 122 bibliotecas de pequenos RNAs para, até o momento, descobrirmos um exemplo claro de sequência viral em meio a matéria escura da metagenômica de mosquitos, que foi AejapNV1 S2 (ABBO et al., 2023). Tudo indica que explorar essas sequências virais altamente divergentes será um trabalho árduo de mineração de dados. Aos dispostos a aceitar o desafio, algo é certo, nesses tempos em que sequências virais tem sido exploradas em escala de Petabase de dados (EDGAR et al., 2022), não faltará dados brutos e o foco poderá ser no desenvolvimento de estratégias computacionais que otimizem recursos para tal mineração de novas sequências virais que poderão contribuir imensamente para nosso conhecimento sobre a evolução dos vírus.

Para alguns potenciais vírus, não encontramos sequências de RdRP. Os exemplos mais notáveis são os Bunyavírus AejapBV2 e AevexBV1. Bunyavírus possuem genomas tri segmentados de -ssRNA nos quais o segmento L (*large*) codifica a RdRP, o segmento M (*medium*) codifica a glicoproteína e o segmento S (small) o capsídeo. Com base nas detecções bem-sucedidas de Bunyavirus segmentados em estudos anteriores aplicado nossa abordagem de pequenos RNAs (AGUIAR et al., 2015; OLMO et al., 2023), é improvável que os segmentos L desses vírus estejam presentes nas bibliotecas de *Ae, japonicos* ou *Ae. vexans* e não foram detectados. Além disso, conseguimos montar os segmentos M e S completos para AejapBV2 com sucesso. Nas bibliotecas de *Ae. japonicos,* foi também detectamos outro Bunyavirus, AejapBV1, para o qual os três segmentos L, M e S foram detectados. Rearranjos de segmentos entre Bunyavírus são eventos frequentemente, há evidências de que a maioria, se não todos, os Bunyavírus evoluíram a partir de eventos de rearranjos entre segmentos genômicos (BRIESE; CALISHER; HIGGS, 2013; KAPUSCINSKI et al., 2021) Dada a capacidade dos Bunyavírus de rearranjar segmentos genômicos, AejapBV2 poderia possivelmente usar a RdRp de AejapBV1 para replicação. No entanto, AejapBV1 e AejapBV2 pertencem a duas famílias de Bunyavírus diferentes, *Phenuiviridae* e *Phasmaviridae*, respectivamente, o que pode tornar a complementação ou rearranjo menos provável. Além disso, essa hipótese implica que a replicação de AejapBV2 depende da presença do segmento L de AejapBV1, mas não conseguimos detectar AejapBV1 em nossa biblioteca de pequenos RNAs de *Ae. japonicus* NL_02 enquanto AejapBV2 estava presente. Um padrão interessante é o de que ambos Bunyavirus

sem RdRP detectadas coocorrem com os dsRNAs vírus AejapTV1 e AevexCV1, um Totivirus e um Crhysovirus, em seus hospedeiros. A estratégia de agrupamentos hierárquicos com siRNAs os coloca nos mesmos agrupamentos. Apesar da coincidência, os mecanismos de replicação de vírus de dsRNAs são completamente diferentes dos de -ssRNA segmentados (TAO; YE, 2010), o que torna pouco provável a hipótese de que os Bunyavirus sem RdRP, poderiam, de alguma forma, usar a RdRP desses vírus. O fato de que esses dsRNAs e -ssRNAs fazem parte dos mesmos agrupamentos em bibliotecas de mosquitos distintos quando aplicada nossa técnica de agrupamentos utilizando as quantificações de siRNAs pode indicar alguma relação biológica desses vírus, ou, um exemplo de falha do nosso método que, presumidamente, agrupa contigs do mesmo vírus. Tal exemplo deixa claro que a interpretação das análises de agrupamentos hierárquico para inferir vírus únicos utilizando a quantificação de siRNAs proposta nessa tese exige inspeção visual para lidar com possíveis agrupamentos heterogêneos. Estudos futuros serão necessários para elucidar a estratégia de replicação de AejapBV2 e AevexBV1 em seus mosquitos hospedeiros.

A produção de piRNAs virais se trata de um fenômeno pouco entendido e a relação desses piRNAs com mecanismos antivirais ainda não foi completamente estabelecida experimentalmente em insetos. O fato de que a produção de piRNAs não é observada para todos os vírus infectando mosquitos também é um fenômeno incompreendido. Nessa tese, geramos o maior catálogo de perfil de pequenos RNAs virais de mosquitos de que temos conhecimento. Ao compararmos esses perfis, obtivemos um padrão claro de que a produção de piRNAs virais é um atributo dos vírus de genoma -ssRNA (Classe V). Uma provável explicação para a geração de piRNAs dos Bunyavirus seria o tropismo desses vírus para tecidos reprodutivos do mosquito (AGUIAR et al., 2015), onde as proteínas PIWI estariam mais ativas e gerariam piRNAs virais apenas por uma questão de altas concentrações de RNAs virais e proteínas PIWI facilitando a ocorrência da geração de piRNAs de forma contingente. Porém faltam resultados experimentais para o estabelecimento dessa generalização e sabemos que as proteínas PIWI não são restritas a tecidos reprodutivos em *A. aegypti* (MIESEN; JOOSTEN; RIJ, 2016). Uma outra explicação, seria algum tipo de preferência de substrato das proteínas das vias de piRNAs pelos RNAs dos vírus de -ssRNA.

Ao observarmos os padrões de pequenos RNAs das respostas de RNAi de *A. aegypti*, *A. albopictus*, *A. japonicus*, *A. vittattus* podemos notar uma conservação nos perfis de piRNAs gerados por esses mosquitos para os Bunyavirus que os infectam, com perfis similares até mesmo entre os mesmos segmentos de diferentes vírus infectando diferentes mosquitos. Um mesmo padrão de conservação no perfil de piRNAs gerados pode ser notada para os Anphevírus da família *Xinmoviridae* infectando *A. aegypti*, *A, japonicus* e *A. vittattus*. Tal conservação nas repostas de RNAi na infecção desses vírus pode ser uma evidência de que esses ISVs vêm coevoluindo com os ancestrais desses mosquitos já que a transferência horizontal de ISVs entre espécies é um fenômeno pouco provável.

Interessante notar que o único vírus não -ssRNA que parece produzir piRNAs é o CFAV de genoma +ssRNA, que sabidamente possui uma EVE correspondente em *A. aegypti* com produção ativa de piRNAs (SUZUKI et al., 2020). A maioria dos vírus que deram origem as EVEs encontradas em nosso estudo são de -ssRNA. O viés de EVEs de -ssRNA é algo que parece ocorrer em artrópodes como um todo (HOLMES, 2011; TER HORST et al., 2019). Uma provável explicação é a de que esses vírus produzem mRNAs comumente menores que os demais vírus, facilitando os processos não autônomos de transcrição reversa e endogenização (HOLMES, 2011). O padrão que detectamos aqui ajuda a fundamentar futuros estudos para entendermos a relação entre eventos de endogenização e produção de piRNAs de vírus -ssRNAs e EVEs. Baseada nas observações dos perfis de pequenos RNAs do nosso viroma, se os piRNAs virais realmente possuem papel antiviral, essa proteção parece ser específica contra vírus de -ssRNA em mosquitos.

Outro aspecto relevante dos vírus de -ssRNA que impactou o desenvolvimento desse trabalho é o fato de que muitos contigs desses vírus possuem maiores quantidades de piRNAs do que siRNAs, sendo os contigs desse vírus os maiores confudidores do modelo classificador de vírus e EVEs. Na **Figura 37,** podemos notar algumas misturas de pontos na separação das classes Viral e EVE. Uma inspeção manual desses pontos nos mostrou que sua grande maioria são contigs de vírus -ssRNA. Em geral a capacidade distintiva dos atributos de pequenos RNAs para as duas classes é eficaz, levando a geração de modelos com alto desempenho de classificação para distinguir essas duas classes automaticamente (**Figura 38**). Porém a ressalva de que os vírus de -ssRNAs, ainda com taxas de falsos negativos aceitáveis, podem ser classificados como EVEs devido a sua intrínseca característica

de produção de piRNAs, deve ser feita para quem pretende usar a ferramenta. Um dos passos seguintes de melhora da ferramenta *small RNA Metavir* será a expansão dos dados de treino e teste com o processamento de mais bibliotecas públicas. O atual volume de dados que utilizamos para treinar os modelos influenciou a nossa escolha de embarcar na ferramenta como classificador automático um modelo *Random Forest*, que, por se tratar de um modelo *ensemble* do tipo *bagging*, tem bom desempenho com conjuntos de dados reduzidos em relação as possibilidades de sobre ajuste (*overfitting*).

A análise do EVEroma dos mosquitos vetores revelou um universo de sequências virais abundantes e diversas que superam o viroma em potenciais espécies virais representadas. O EVEroma pode ser fonte de dados para inferir infecções virais ancestrais. A baixa similaridade de sequência dos contigs do viroma com o EVEroma nesse estudo evidencia que as sequências virais endogenizadas podem se tratar de eventos de endogenizações ancestrais dos quais os atuais *loci* genômicos remanescentes já acumularam muitas mutações. Além disso, o EVEroma pode também ser usado para inferir espécies virais que os mosquitos possam vir a se infectar, uma vez que EVEs relacionadas sejam encontradas no genoma. Porém, é preciso cuidado ao inferir espécies virais no EVEroma. Por ser composto, em grande maioria, por sequências pequenas, a determinação de quais espécies virais estão contidas no EVEroma irá depender de regulação de parâmetros na execução e cortes de resultados ao usar alinhadores locais. Por uma questão de conveniência, nesse estudo os flancos não virais dos contigs de EVEs foram removidos, isso facilitou a análise de alinhamentos e remoção de redundâncias para focar nas porções virais dos contigs. Porém, essas porções podem conter informações relevantes sobre a associação das EVEs com elementos transponíveis e regiões genômicas repetitivas (DEZORDI et al., 2020), o que é extremamente importante para se entender o fenômeno de endogenização de EVEs não retrovirais. Apesar das limitações, as EVEs identificadas nesse estudo têm suporte do perfil de piRNAs que conseguimos associar aos contigs e da alta taxa de alinhamentos consistentes nos genomas de A. *aegypti* e *A. albopictus* dos contigs classificados como EVEs montados para essas duas espécies. Se considerarmos que cada um dos alinhamentos virais no genoma de *A. aegypti* pode se tratar de uma EVE, as mais de cinco mil inserções virais que obtivemos é um número maior que o já reportado de EVEs em *A. aegypti* (AGUIAR et al., 2020; PALATINI et al., 2017; WHITFIELD et al., 2017). Porém, reportamos esse

resultado com o cuidado de informar que mais análises devem ser feitas para confirmar se todos esses alinhamentos correspondem a loci de EVEs. Um próximo passo será a curadoria das EVEs de DNA e retrovirais para garantir que não se trata de elementos transponíveis (AGUIAR et al., 2020). Como próximo passo para validar nossa estratégia de detecção de EVEs, já planejamos reações de PCR para confirmar ao menos um grupo de EVEs descobertas nesse estudo.

Nossa avaliação da relação da carga de piRNAs com a carga viral dos 34 pares cognatos EVE/vírus circulantes identificados não evidencia que os piRNAs sendo gerados por essas EVEs poderiam inibir os vírus cognatos. Pelo contrário, os resultados dos altos valores de correlação sugerem um aumento mútuo da carga de piRNA de EVEs com siRNA virais nos pares cognatos. Nem mesmo para a EVE de CFAV, que está entre os 34 pares identificados e que é um exemplo e inibição da replicação viral pelos piRNAs produzidos (SUZUKI et al., 2020), não obtivemos evidências desse fenômeno com a análise de cargas de pequenos RNAs realizada. No trabalho de AGUIAR et al., 2020 nós evidenciamos que os pequenos RNAs gerados por EVEs podem ser apenas, apenas subprodutos da associação desses elementos com elementos transponíveis e mão necessariamente tem uma funçãoantiviral, ou qualquer outra. Nos ´próximos passos do trabalho com o EVEroma, a contextualização genômica das EVEs reportadas será realizada com cuidado para verificar sua associação com elementos transponíveis.

Em relação as bactérias Wolbachias, não há dúvidas de que esse endossimbionte pode modular a resposta antiviral de mosquitos. Atualmente, estima-se que cerca de 66 espécies de mosquitos sejam infectadas por *Wolbachia* (DA SILVA et al., 2021). Os autores da revisão argumentam que esse número é uma provável subestimativa enviesada pelos métodos aplicados e pelo foco da detecção desse endossimbionte apenas em algumas espécies. Até a data da revisão de DA SILVA et al., 2021, não havia estudos de detecção de *Wolbachia* em mosquitos nativos transmissores de mosquitos na África como *Ae. furcifer*, *Ae. taylori* e *Ae. luteocephalus* e em mosquitos do gênero *Sabethes*, nativo da América do Sul. Recentes abordagens de metatrascriptômica tem se mostrado eficientes para detecção e caracterização de infecções por *Wolbachia* em bibliotecas de drosofilideos (ORTIZ-BAEZ et al., 2021) e mosquitos (LI et al., 2023). Esse tipo de detecção em larga escala permite o reaproveitamento de bibliotecas inicialmente planejadas para outros fins, *e.g.* detecção de sequências virais, para avaliação de fenômenos biológicos que

relacionados a amostra levando ao estabelecimento de novas hipóteses sobre o efeito de coinfecções nos hospedeiros.

A abordagem de detecção de pequenos RNAs de *Wolbachia* proposta nesse estudo mostrou robustez e boa precisão na detecção das linhagens *wAlb* e *wMel* em três conjuntos de dados independentes gerados em condições de infecção completamente distintas: infecção artificial por *wAlb* em *A. aegypti* de laboratório, infecção natural por *wAlb* em *A. albopictus* de campo e infecção artificial por *wMel* em *A. aegypti* de campo. O resultado da comparação entre as bibliotecas oxidadas e controles de amostras RNA de *A. albopictus* de Estrasburgo evidencia que os pequenos RNAs de *Wolbachia* não possuem grupamentos metil 2'O, característico dos pequenos RNAs carregados nas proteínas argonautas nas vias de RNAi (AGUIAR; OLMO; MARQUES, 2016; YANG et al., 2007), ou qualquer outro grupamento químico que proteja essas moléculas da degradação decorrente das modificações induzidas pelo processo de oxidação. A grande abundância de pequenos RNAs de tRNAs e rRNAs de *wAlb* infectando *A. aegypti* (**Figura 42**) sugere que boa parte dos pequenos RNAs de *Wolbachia* sendo detectados sejam oriundos da degradação de ncRNAs altamente estruturados e abundantes na célula microbiana (DEUTSCHER, 2003). Além da degradação, tRNAs e rRNAs procarióticos passam por processos de maturação nos quais a clivagem por RNAses podem levar a produção de pequenos RNAs (BROGLIA; LE RHUN; CHARPENTIER, 2023). Para *Staphylococcus aureus,* foi mostrado que a RNAse III dessa bactéria gera pequenos RNAs de ~20nt a partir de substratos de dsRNA formados por sítios onde ocorrem eventos de transcrições sobrepostas (LASA et al., 2011). Wolbachias também codificam pequenos RNAs conservados (~30nt) com potencial papel na interação com *A. aegypti* (MAYORAL et al., 2014). Mais estudos serão necessários para a compreensão dos mecanismos de origem e potenciais funções dos pequenos RNAs que detectamos na interação do endossimbionte com os mosquitos hospedeiros.

A detecção de pequenos RNAs de *Wolbachia*, mais especificamente *wAlb*, apenas em bibliotecas de *A. albopictus* entre as espécies analisadas nesse trabalho é um resultado coerente com a alta prevalência de infecções por *wAlb* já reportada em estudos de campo para essa espécie (ROSS et al., 2020). A não detecção de pequenos RNAs de *Wolbachia* em bibliotecas de *A. aegypti* de campo no projeto ZikaAlliance, mesmo sendo essa a espécie mais bem amostrada em número de bibliotecas, também é um resultado coerente com o que se presume ser um fenômeno

robusto que *A. aegypti* não é infectado naturalmente por *Wolbachia*. As detecções reportadas para *A. aegypti* provavelmente se tratam de contaminações, erros metodológicos ou infecções isoladas recentes sendo estabelecidas em campo, mas que não são replicadas em testes de laboratório (DA SILVA et al., 2021; ROSS et al., 2020). Apesar da consistência das detecções de pequenos RNAs em nossos testes, não é possível afirmar a ausência completa de infecção por *Wolbachia* nas outras espécies de mosquitos avaliadas. Espécies de mosquitos como as do gênero *Sabethes sp*., para a qual não há estudos de detecção do endossimbionte podem ser infectados por linhagens divergentes o bastante para escaparem da nossa estratégia de detecção na qual utilizamos apenas genomas completos de *Wolbachia* disponíveis no Refseq.

O projeto de controle biológico da transmissão de dengue, chikungunya e Zika por *Ae. aegypti* utilizando mosquitos artificialmente infectados por wMel soltos em campo tem obtido sucesso na cidade de Niterói-RJ (PINTO et al., 2021). Em março de 2020, a avaliação das Zonas da cidade em que foram soltos os mosquitos infectados entre os anos de 2017-2019 mostra que entre 33% a 99% de mosquitos das populações nas áreas de soltura se encontram infectados por *wMel* e, mais importante, PINTO et al., 2020 mostram uma redução de 69% nos casos de dengue notificados pelas autoridades de saúde quando comparados dados de pacientes das zonas de soltura de mosquitos com a zona controle. Recentemente, nosso grupo mostrou que ISVs, mais especificamente PCLV e HTV, estão associados a uma maior prevalência de ZIKV e DENV em mosquitos de campo e aumentam a transmissão desses vírus por *A. aegypti* em experimentos de laboratório (OLMO et al., 2023). Os resultados de nossas análises temporais da Zona 3 (**Figura 48**) e correlações das cargas virais desses dois ISVs com a carga de *Wolbachia* nas bibliotecas de mosquitos de Niterói (**Figura 49**) evidenciam que o efeito de redução da carga viral de arbovírus induzidos pela infecção artificial por *wMel* em *A. aegypti* se estende a esses dois vírus.

Os efeitos das potenciais interações de *Wolbachias* com ISVs ainda são pouco conhecidos, principalmente em mosquitos de campo. Em experimentos realizados com cultura de células de *A. aegypti*, infecções por *wMel parecem* inibir a replicação do ISV CFAV e não afeitam a replicação do PCLV (MCLEAN et al., 2019; SCHNETTLER et al., 2016). Também em cultura de células, mas com um efeito contrário, há evidências de que a infecção por *Wolbachia* aumenta a replicação do

ISV *Aedes anphevirus* (PARRY; ASGARI, 2018). Um estudo de campo em Cairns, Australia, avaliou mosquitos coletados em uma região em que o projeto Wolbachia também foi implementado (AMUZU et al., 2018). Os autores compararam os vírus de *A. aegypti* coletados em região de soltura com regiões controle e, ao contrário dos nossos resultados com PCLV e HTV, mostram que *Ae. aegypti* infectados com *Wolbachia* possuem maiores cargas virais de ISVs. Uma limitação relevante desse estudo é que os autores foram restritos a avaliação de ISVs da família *Flaviviridae*, pois fizeram apenas o sequenciamento de amplicons utilizando primers degenerados para a região NS5 de flavivírus. Os mecanismos pelos quais *Wolbachias* afetam a competência vetorial de mosquitos, ou insetos no geral, ainda são poucos conhecidos. Há evidências que possam estar relacionados a interferências direta da bactéria no sistema imune ou competição por recursos celulares com os vírus (PIMENTEL et al., 2021).

Ao considerar nossos resultados com PCLV e HTV, não conseguimos identificar um efeito claro da infecção por *Wolbachia* em ISVs, tanto em cultura celular quanto em mosquitos A. aegypti, entre os estudos publicados. Considerando que esses dois componentes biológicos impactam a competência vetorial de *A. aegypti* para arbovírus, compreender os potenciais efeitos das coinfecções de ISVs e *Wolbachia* é essencial para aprofundar nosso conhecimento sobre a biologia do vetor e esclarecer interações que possam ampliar ou reduzir a efetividade dos mosquitos com Wolbachia na redução da transmissão de arbovírus a longo prazo.

Além disso, uma vez que PCLV e HTV são vírus residentes mais prevalentes, com cargas virais facilmente detectáveis em comparação com cargas de arbovírus em mosquitos de campo (OLMO et al., 2023) , eles podem servir como indicadores para avaliar o impacto das liberações de mosquitos com *Wolbachia* em populações locais ao longo do tempo. Isso pode ser realizado por meio de ensaios de qPCR, caso os resultados observados nesta tese sejam confirmados por mais experimentos.

7 - Conclusões

Neste estudo, analisamos o viroma de 10 espécies de mosquitos vetores coletados ao redor do globo e identificamos 28 vírus, sendo 17 deles potenciais vírus novos e 1726 sequências virais endógenas oriundas de 115 prováveis vírus distintos que podem representar infecções virais ancestrais. Utilizamos uma abordagem de metagenômica baseada em pequenos RNAs previamente desenvolvida por (AGUIAR et al., 2015) e posteriormente refinada e automatizada durante a execução dessa tese. Nossa estratégia de metagenômica nos permitiu abordar três dos principais desafios da metagenômica viral em eucariotos nesse estudo: i) a diferenciação de vírus exógenos replicantes de EVEs utilizando os perfis de pequenos RNAs; ii) a associação de contigs oriundos de diferentes segmentos genômicos ao mesmo vírus com base nas análises de coocorrência e de perfis de pequenos RNAs; iii) a identificação e classificação do segmento genômico de um Narnavirus que não apresentou similaridade significativa com sequências de referência conhecidas, um exemplo claro da matéria escura da metagenômica viral. A automatização e disponibilização do nosso pipeline de análises com a ferramenta *Small RNA Metavir* permitirá futuros estudos em larga escala com bibliotecas de pequenos RNAs de mosquitos e de outros organismos e que outros grupos de pesquisa possam utilizar nossa estratégia de análise de forma padronizada. Além disso, conseguimos quantificar e analisar a carga de pequenos RNAs de *Wolbachia,* um relevante endossimbionte capaz de alterar a competência vetorial de mosquitos, e correlacioná-la com a carga viral de ISVs, mostrando que a bactéria pode impactar a carga viral dos relevantes vírus PCLV e HTV em *A. aegypti*. Os resultados obtidos e estratégias computacionais desenvolvidas permitiram caracterizar de forma precisa o viroma de mosquitos vetores e analisá-lo a luz de outros componentes biológicos relevantes para a competência vetorial de mosquitos. O conhecimento preciso dos vírus circulantes e da interação desses agentes com mosquitos vetores é essencial para fundamentação de estratégias de controle biológico desses importantes transmissores de patógenos.

8 - Perspectivas

1- Realizar análises biogeográficas e filogenéticas integradas de todos os ISVs descobertos nesse estudo;

2- Realizar análise de contexto genômico das sequências de EVEs alinhadas nos genomas de *A. aegypti* e *A. albopictus*;

3- Realizar validação por PCR de um conjunto de EVEs a ser amostrado para suportar as sequências descobertas com nossa estratégia;

4 – Sistematizar os testes e concluir o desenvolvimento da *ferramenta Small RNA Metavir*;

5 – Utilizar a ferramenta *Small RNA Metavir* para processar todas as bibliotecas de pequenos RNAs de mosquitos disponíveis no SRA e explorar a porção de contigs desconhecidos treinando novos classificadores;

6 – Realizar RT-qPCRs para correlacionar os resultados da carga viral medida para genes repórteres de *Wolbachia* com a carga de pequenos RNAs do endossimbionte detectadas nesse estudo.

9 - Referências

ABBO, S. R. et al. The invasive Asian bush mosquito Aedes japonicus found in the Netherlands can experimentally transmit Zika virus and Usutu virus. **PLOS Neglected Tropical Diseases**, v. 14, n. 4, p. e0008217, 13 abr. 2020.

ABBO, S. R. et al. The virome of the invasive Asian bush mosquito Aedes japonicus in Europe. **Virus Evolution**, v. 9, n. 2, p. vead041, 2023.

ABRAHÃO, J. et al. Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. **Nature Communications**, v. 9, n. 1, p. 749, 27 fev. 2018.

ADRIAENSSENS, E. M. et al. Guidelines for public database submission of uncultivated virus genome sequences for taxonomic classification. **Nature Biotechnology**, v. 41, n. 7, p. 898–902, jul. 2023.

Aedes Meigen, 1818 WRBU. Disponível em: <https://wrbu.si.edu/vectorspecies/genera/aedes>. Acesso em: 3 set. 2023.

AGUADO, L. C.; TENOEVER, B. R. RNase III Nucleases and the Evolution of Antiviral Systems. **BioEssays**, v. 40, n. 2, p. 1700173, fev. 2018.

AGUIAR, E. R. G. R. et al. Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host. **Nucleic Acids Research**, v. 43, n. 13, p. 6191–6206, 27 jul. 2015.

AGUIAR, E. R. G. R. et al. A single unidirectional piRNA cluster similar to the flamenco locus is the major source of EVE-derived transcription and small RNAs in Aedes aegypti mosquitoes. **RNA (New York, N.Y.)**, v. 26, n. 5, p. 581–594, maio 2020.

AGUIAR, E. R. G. R.; OLMO, R. P.; MARQUES, J. T. Virus-derived small RNAs: molecular footprints of host-pathogen interactions. **Wiley interdisciplinary reviews. RNA**, v. 7, n. 6, p. 824–837, nov. 2016.

AKBARI, O. S. et al. The Developmental Transcriptome of the Mosquito Aedes aegypti, an Invasive Species and Major Arbovirus Vector. **G3 Genes|Genomes|Genetics**, v. 3, n. 9, p. 1493–1509, 1 set. 2013.

ALARCÓN-ELBAL, P. M. et al. The First Record of Aedes vittatus (Diptera: Culicidae) in the Dominican Republic: Public Health Implications of a Potential Invasive Mosquito Species in the Americas. **Journal of Medical Entomology**, v. 57, n. 6, p. 2016–2021, 13 nov. 2020.

ALEFELDER, S.; PATEL, B. K.; ECKSTEIN, F. Incorporation of terminal phosphorothioates into oligonucleotides. **Nucleic Acids Research**, v. 26, n. 21, p. 4983–4988, 1 nov. 1998.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403–410, 5 out. 1990.

AMGARTEN, D. et al. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. **Frontiers in Genetics**, v. 9, 2018.

AMUZU, H. E. et al. Wolbachia enhances insect-specific flavivirus infection in Aedes aegypti mosquitoes. **Ecology and Evolution**, v. 8, n. 11, p. 5441–5454, 8 maio 2018.

ARAKAKI, N.; MIYOSHI, T.; NODA, H. Wolbachia-mediated parthenogenesis in the predatory thrips Franklinothrips vespiformis (Thysanoptera: Insecta). **Proceedings. Biological Sciences**, v. 268, n. 1471, p. 1011–1016, 22 maio 2001.

ARAUJO, L. M.; FERREIRA, M. L. B.; NASCIMENTO, O. J. Guillain-Barré syndrome associated with the Zika virus outbreak in Brazil. **Arquivos de Neuro-Psiquiatria**, v. 74, n. 3, p. 253–255, mar. 2016.

ARBUCKLE, J. H. et al. The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes in vivo and in vitro. **Proceedings of the National Academy of Sciences**, v. 107, n. 12, p. 5563–5568, 23 mar. 2010.

ASWAD, A.; KATZOURAKIS, A. Paleovirology and virally derived immunity. **Trends in Ecology & Evolution**, v. 27, n. 11, p. 627–636, nov. 2012.

AUSLANDER, N.; GUSSOW, A. B.; KOONIN, E. V. Incorporating Machine Learning into Established Bioinformatics Frameworks. **International Journal of Molecular Sciences**, v. 22, n. 6, p. 2903, 12 mar. 2021.

BAIDALIUK, A. et al. Cell-Fusing Agent Virus Reduces Arbovirus Dissemination in Aedes aegypti Mosquitoes In Vivo. **Journal of Virology**, v. 93, n. 18, p. 10.1128/jvi.00705-19, 28 ago. 2019.

BALTIMORE, D. Expression of Animal Virus Genomes. **BACTERIOL. REV.**, v. 35, 1971.

BANKEVICH, A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. **Journal of Computational Biology**, v. 19, n. 5, p. 455–477, maio 2012.

BARRANGOU, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. **Science (New York, N.Y.)**, v. 315, n. 5819, p. 1709–1712, 23 mar. 2007.

BATSON, J. et al. Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay. **eLife**, v. 10, p. e68353, 27 abr. 2021.

BECHT, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. **Nature biotechnology**, 1 dez. 2018.

BECKMANN, J. F.; RONAU, J. A.; HOCHSTRASSER, M. A Wolbachia deubiquitylating enzyme induces cytoplasmic incompatibility. **Nature Microbiology**, v. 2, p. 17007, 1 mar. 2017.

BEKAERT, M. et al. Towards a computational model for −1 eukaryotic frameshifting sites. **Bioinformatics**, v. 19, n. 3, p. 327–335, 12 fev. 2003.

BENNER, S. A. Defining Life. **Astrobiology**, v. 10, n. 10, p. 1021–1030, dez. 2010.

BHATT, S. et al. The global distribution and burden of dengue. **Nature**, v. 496, n. 7446, p. 504–507, abr. 2013.

BISHOP, C. et al. Analysis of Aedes aegypti microRNAs in response to Wolbachia wAlbB infection and their potential role in mosquito longevity. **Scientific Reports**, v. 12, n. 1, p. 15245, 9 set. 2022.

BOLLING, B. G. et al. Transmission dynamics of an insect-specific flavivirus in a naturally infected Culex pipiens laboratory colony and effects of co-infection on vector competence for West Nile virus. **Virology**, v. 427, n. 2, p. 90–97, 5 jun. 2012.

BRENNER, S. E.; CHOTHIA, C.; HUBBARD, T. J. P. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. **Proceedings of the National Academy of Sciences of the United States of America**, v. 95, n. 11, p. 6073–6078, 26 maio 1998.

BRIESE, T.; CALISHER, C. H.; HIGGS, S. Viruses of the family Bunyaviridae: are all available isolates reassortants? **Virology**, v. 446, n. 1–2, p. 207–216, nov. 2013.

BROGLIA, L.; LE RHUN, A.; CHARPENTIER, E. Methodologies for bacterial ribonuclease characterization using RNA-seq. **FEMS Microbiology Reviews**, v. 47, n. 5, p. fuad049, 1 set. 2023.

BUCHAN, D. W. A. et al. Scalable web services for the PSIPRED Protein Analysis Workbench. **Nucleic Acids Research**, v. 41, n. Web Server issue, p. W349-357, jul. 2013.

BUCHAN, D. W. A.; JONES, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. **Nucleic Acids Research**, v. 47, n. W1, p. W402–W407, 2 jul. 2019.

BUCHFINK, B.; REUTER, K.; DROST, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. **Nature Methods**, v. 18, n. 4, p. 366–368, abr. 2021.

BURRELL, C. J.; HOWARD, C. R.; MURPHY, F. A. Virion Structure and Composition. **Fenner and White's Medical Virology**, p. 27–37, 2017.

BZHALAVA, Z. et al. Machine Learning for detection of viral sequences in human metagenomic datasets. **BMC Bioinformatics**, v. 19, n. 1, p. 336, 24 set. 2018.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC Bioinformatics**, v. 10, n. 1, p. 421, 15 dez. 2009.

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

CARVALHO, C. et al. DIPTERA, Linnaeus, 1758. Em: **Insetos do Brasil: DIvesidade e Taxonomia**. [s.l: s.n.]. p. 702–743.

CENIK, E. S. et al. Phosphate and R2D2 restrict the substrate specificity of Dicer-2, an ATP-driven ribonuclease. **Molecular Cell**, v. 42, n. 2, p. 172–184, 22 abr. 2011.

CERUTTI, H.; CASAS-MOLLANO, J. A. On the origin and functions of RNA-mediated silencing: from protists to man. **Current Genetics**, v. 50, n. 2, p. 81–99, ago. 2006.

CHANG, W. et al. **shiny: Web Application Framework for R**. , 12 ago. 2023. Disponível em: <https://cran.r-project.org/web/packages/shiny/index.html>. Acesso em: 31 ago. 2023

CHAO, J. A. et al. Dual modes of RNA-silencing suppression by Flock House virus protein B2. **Nature Structural & Molecular Biology**, v. 12, n. 11, p. 952–957, nov. 2005.

CHARON, J. et al. Novel RNA viruses associated with Plasmodium vivax in human malaria and Leucocytozoon parasites in avian disease. **PLoS pathogens**, v. 15, n. 12, p. e1008216, dez. 2019.

COCHET, A. et al. Autochthonous dengue in mainland France, 2022: geographical extension and incidence increase. **Eurosurveillance**, v. 27, n. 44, 3 nov. 2022.

COFFIN, J. M. 50th anniversary of the discovery of reverse transcriptase. **Molecular Biology of the Cell**, v. 32, n. 2, p. 91–97, 15 jan. 2021.

COMPEAU, P. E. C.; PEVZNER, P. A.; TESLER, G. How to apply de Bruijn graphs to genome assembly. **Nature Biotechnology**, v. 29, n. 11, p. 987–991, nov. 2011.

COOK, S. et al. Mitochondrial Markers for Molecular Identification of Aedes Mosquitoes (Diptera: Culicidae) Involved in Transmission of Arboviral Disease in West Africa. **Journal of Medical Entomology**, v. 42, n. 1, p. 19–28, 1 jan. 2005.

CORNELIS, G. et al. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 7, p. E432–E441, 14 fev. 2012.

CRABTREE, M. B. et al. Genetic and phenotypic characterization of the newly described insect flavivirus, Kamiti River virus. **Archives of Virology**, v. 148, n. 6, p. 1095–1118, jun. 2003.

CZECH, B.; HANNON, G. J. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. **Trends in Biochemical Sciences**, v. 41, n. 4, p. 324–337, abr. 2016.

DA SILVA, L. M. et al. Systematic Review of Wolbachia Symbiont Detection in Mosquitoes: An Entangled Topic about Methodological Power and True Symbiosis. **Pathogens**, v. 10, n. 1, p. 39, jan. 2021.

DAINTY, K. R. et al. wMel Wolbachia genome remains stable after 7 years in Australian Aedes aegypti field populations. **Microbial Genomics**, v. 7, n. 9, p. 000641, 1 set. 2021.

DARTY, K.; DENISE, A.; PONTY, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. **Bioinformatics**, v. 25, n. 15, p. 1974–1975, 1 ago. 2009.

DAWKINS, R. Chapter 2: The Replicators. Em: **The Selfish Gene**. 2. ed. [s.l.] Oxford University Press, 1990.

DE ALMEIDA, J. P. et al. The virome of vector mosquitoes. **Current Opinion in Virology**, v. 49, p. 7–12, ago. 2021.

DE FARIA, I. J. S. et al. Invading viral DNA triggers dsRNA synthesis by RNA polymerase II to activate antiviral RNA interference in Drosophila. **Cell Reports**, v. 39, n. 12, p. 110976, 21 jun. 2022.

DE FARIAS, S. T.; JOSE, M. V.; PROSDOCIMI, F. Is it possible that cells have had more than one origin? **Bio Systems**, v. 202, p. 104371, abr. 2021.

DERISI, J. L. et al. An exploration of ambigrammatic sequences in narnaviruses. **Scientific Reports**, v. 9, n. 1, p. 17982, 29 nov. 2019.

DEUTSCHER, M. P. Degradation of Stable RNA in Bacteria *. **Journal of Biological Chemistry**, v. 278, n. 46, p. 45041–45044, 14 nov. 2003.

DEZORDI, F. Z. et al. In and Outs of Chuviridae Endogenous Viral Elements: Origin of a Potentially New Retrovirus and Signature of Ancient and Ongoing Arms Race in Mosquito Genomes. **Frontiers in Genetics**, v. 11, 2020.

DI GIACOMO, M. et al. Multiple epigenetic mechanisms and the piRNA pathway enforce LINE1 silencing during adult spermatogenesis. **Molecular Cell**, v. 50, n. 4, p. 601–608, 23 maio 2013.

DIALLO, D. et al. Dengue vectors in Africa: A review. **Heliyon**, v. 8, n. 5, p. e09459, 17 maio 2022.

DINAN, A. M. et al. A case for a negative-strand coding sequence in a group of positive-sense RNA viruses. **Virus Evolution**, v. 6, n. 1, p. veaa007, 1 jan. 2020.

DOMS, R. W. Basic Concepts. **Viral Pathogenesis**, p. 29–40, 2016.

DONALD, C. L. et al. Antiviral RNA Interference Activity in Cells of the Predatory Mosquito, Toxorhynchites amboinensis. **Viruses**, v. 10, n. 12, p. 694, 6 dez. 2018.

DRINNENBERG, I. A. et al. RNAi in budding yeast. **Science (New York, N.Y.)**, v. 326, n. 5952, p. 544–550, 23 out. 2009.

DUDAS, G. et al. Polymorphism of genetic ambigrams. **Virus Evolution**, v. 7, n. 1, p. veab038, jan. 2021.

DYER, K. A.; JAENIKE, J. Evolutionarily Stable Infection by a Male-Killing Endosymbiont in Drosophila innubila. **Genetics**, v. 168, n. 3, p. 1443–1455, nov. 2004.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, 2004.

EDGAR, R. C. et al. Petabase-scale sequence alignment catalyses viral discovery. **Nature**, v. 602, n. 7895, p. 142–147, fev. 2022.

FANCELLO, L.; RAOULT, D.; DESNUES, C. Computational tools for viral metagenomics and their application in clinical research. **Virology**, v. 434, n. 2, p. 162–174, 20 dez. 2012.

FENG, Y. et al. A time-series meta-transcriptomic analysis reveals the seasonal, host, and gender structure of mosquito viromes. **Virus Evolution**, v. 8, n. 1, p. veac006, 2022.

FERREIRA, F. V. et al. The small non-coding RNA response to virus infection in the Leishmania vector Lutzomyia longipalpis. **PLOS Neglected Tropical Diseases**, v. 12, n. 6, p. e0006569, 4 jun. 2018.

FONSECA, P. L. C. et al. Virome analyses of Hevea brasiliensis using small RNA deep sequencing and PCR techniques reveal the presence of a potential new virus. **Virology Journal**, v. 15, n. 1, p. 184, 26 nov. 2018.

FORNI, D. et al. Disease-causing human viruses: novelty and legacy. **Trends in Microbiology**, v. 30, n. 12, p. 1232–1242, 1 dez. 2022.

FORTERRE, P. To be or not to be alive: How recent discoveries challenge the traditional definitions of viruses and life. **Studies in History and Philosophy of Biological and Biomedical Sciences**, v. 59, p. 100–108, out. 2016.

FORTERRE, P.; KRUPOVIC, M.; PRANGISHVILI, D. Cellular domains and viral lineages. **Trends in Microbiology**, v. 22, n. 10, p. 554–558, out. 2014.

FU, L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data. **Bioinformatics (Oxford, England)**, v. 28, n. 23, p. 3150–3152, 1 dez. 2012.

GAGLIA, M. M.; GLAUNSINGER, B. A. Viruses and the cellular RNA decay machinery. **Wiley Interdisciplinary Reviews. RNA**, v. 1, n. 1, p. 47–59, 2010.

GAINETDINOV, I. et al. A Single Mechanism of Biogenesis, Initiated and Directed by PIWI Proteins, Explains piRNA Production in Most Animals. **Molecular Cell**, v. 71, n. 5, p. 775- 790.e5, 6 set. 2018.

GHILDIYAL, M. et al. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. **Science (New York, N.Y.)**, v. 320, n. 5879, p. 1077–1081, 23 maio 2008.

GLORIA-SORIA, A. et al. Global Genetic Diversity of Aedes aegypti. **Molecular ecology**, v. 25, n. 21, p. 5377–5395, nov. 2016.

GÖERTZ, G. P. et al. Subgenomic flavivirus RNA binds the mosquito DEAD/H-box helicase ME31B and determines Zika virus transmission by Aedes aegypti. **Proceedings of the National Academy of Sciences**, v. 116, n. 38, p. 19136–19144, 17 set. 2019.

GRYBCHUK, D. et al. Viral discovery and diversity in trypanosomatid protozoa with a focus on relatives of the human parasite Leishmania. **Proceedings of the National Academy of Sciences of the United States of America**, v. 115, n. 3, p. E506–E515, 16 jan. 2018.

GU, Z.; EILS, R.; SCHLESNER, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. **Bioinformatics**, v. 32, n. 18, p. 2847–2849, 15 set. 2016.

GU, Z.; HÜBSCHMANN, D. Make Interactive Complex Heatmaps in R. **Bioinformatics**, v. 38, n. 5, p. 1460–1462, 1 mar. 2022.

HAHSLER, M.; HORNIK, K.; BUCHTA, C. Getting Things in Order: An Introduction to the R Package seriation. **Journal of Statistical Software**, v. 25, p. 1–34, 18 mar. 2008.

HAN, B. W. et al. piRNA-Guided Transposon Cleavage Initiljates Zucchini-Dependent, Phased piRNA Production. **Science (New York, N.Y.)**, v. 348, n. 6236, p. 817–821, 15 maio 2015a.

HAN, B. W. et al. Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. **Science (New York, N.Y.)**, v. 348, n. 6236, p. 817–821, 15 maio 2015b.

HAN, Y.-H. et al. RNA-based immunity terminates viral infection in adult Drosophila in the absence of viral suppression of RNA interference: characterization of viral small interfering RNA populations in wild-type and mutant flies. **Journal of Virology**, v. 85, n. 24, p. 13153–13163, dez. 2011.

HANDELSMAN, J. et al. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. **Chemistry & Biology**, v. 5, n. 10, p. R245–R249, 1 out. 1998.

HANLEY, K. A.; WEAVER, S. C. Arbovirus Evolution. Em: **Origin and Evolution of Viruses**. [s.l.] Elsevier, 2008. p. 351–391.

HARVEY, E.; HOLMES, E. C. Diversity and evolution of the animal virome. **Nature Reviews Microbiology**, v. 20, n. 6, p. 321–334, jun. 2022.

HILLMAN; ESTEBAN. **Family: Narnaviridae. ICTV Ninth Report; 2009 Taxonomy Release**. Disponível em: <https://ictv.global/report_9th/RNApos/Narnaviridae>. Acesso em: 20 ago. 2023.

HINO, A.; MARUYAMA, H.; KIKUCHI, T. A novel method to assess the biodiversity of parasites using 18S rDNA Illumina sequencing; parasitome analysis method. **Parasitology International**, Current Manual for Parasitological Research. v. 65, n. 5, Part B, p. 572–575, 1 out. 2016.

HOLMES, E. C. The Evolution of Endogenous Viral Elements. **Cell Host & Microbe**, v. 10, n. 4, p. 368–377, 20 out. 2011.

HORIE, M. et al. Endogenous non-retroviral RNA virus elements in mammalian genomes. **Nature**, v. 463, n. 7277, p. 84–87, jan. 2010.

HUANG, X.; MADAN, A. CAP3: A DNA sequence assembly program. **Genome Research**, v. 9, n. 9, p. 868–877, set. 1999.

HUANG, Y. et al. A global dataset of sequence, diversity and biosafety recommendation of arbovirus and arthropod-specific virus. **Scientific Data**, v. 10, n. 1, p. 305, 19 maio 2023.

HUG, L. A. et al. A new view of the tree of life. **Nature Microbiology**, v. 1, n. 5, p. 1–6, 11 abr. 2016.

JONES, D. T.; COZZETTO, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. **Bioinformatics**, v. 31, n. 6, p. 857–863, 15 mar. 2015.

JOOSTEN, J. et al. Endogenous piRNA-guided slicing triggers responder and trailer piRNA production from viral RNA in Aedes aegypti mosquitoes. **Nucleic Acids Research**, v. 49, n. 15, p. 8886–8899, 7 set. 2021.

JUMPER, J. et al. Highly accurate protein structure prediction with AlphaFold. **Nature**, v. 596, n. 7873, p. 583–589, ago. 2021.

KAPUSCINSKI, M. L. et al. Genomic characterization of 99 viruses from the bunyavirus families Nairoviridae, Peribunyaviridae, and Phenuiviridae, including 35 previously unsequenced viruses. **PLoS pathogens**, v. 17, n. 3, p. e1009315, mar. 2021.

KATOH, K. et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic Acids Research**, v. 30, n. 14, p. 3059–3066, 15 jul. 2002.

KATZOURAKIS, A.; GIFFORD, R. J. Endogenous Viral Elements in Animal Genomes. **PLOS Genetics**, v. 6, n. 11, p. e1001191, 18 nov. 2010.

KHURANA, J. S. et al. Adaptation to P element transposon invasion in Drosophila melanogaster. **Cell**, v. 147, n. 7, p. 1551–1563, 23 dez. 2011.

KIM, V. N. MicroRNA biogenesis: coordinated cropping and dicing. **Nature Reviews Molecular Cell Biology**, v. 6, n. 5, p. 376–385, maio 2005.

KOONIN, E. V. et al. Viruses Defined by the Position of the Virosphere within the Replicator Space. **Microbiology and Molecular Biology Reviews : MMBR**, v. 85, n. 4, p. e00193-20, 2021.

KOONIN, E. V.; KRUPOVIC, M.; AGOL, V. I. The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? **Microbiology and Molecular Biology Reviews**, v. 85, n. 3, p. e00053-21, 18 ago. 2021.

KOONIN, E. V.; SENKEVICH, T. G.; DOLJA, V. V. The ancient Virus World and evolution of cells. **Biology Direct**, v. 1, n. 1, p. 29, 19 set. 2006.

KRAEMER, M. U. G. et al. Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. **Nature Microbiology**, v. 4, n. 5, p. 854–863, 4 mar. 2019.

KREUZE, J. F. et al. Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. **Virology**, v. 388, n. 1, p. 1–7, 25 maio 2009.

KRISHNAMURTHY, S. R.; WANG, D. Origins and challenges of viral dark matter. **Virus Research**, v. 239, p. 136–142, 15 jul. 2017.

KRUEGER, F. **Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS**. Altos Labs, , 2015. Disponível em: <https://github.com/FelixKrueger/TrimGalore>

KRUPOVIC, M.; DOLJA, V. V.; KOONIN, E. V. Origin of viruses: primordial replicators recruiting capsids from hosts. **Nature Reviews Microbiology**, v. 17, n. 7, p. 449–458, jul. 2019.

KRUPOVIC, M.; DOLJA, V. V.; KOONIN, E. V. The virome of the last eukaryotic common ancestor and eukaryogenesis. **Nature Microbiology**, v. 8, n. 6, p. 1008–1017, jun. 2023.

KUMAR, S. et al. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. **Molecular Biology and Evolution**, v. 35, n. 6, p. 1547–1549, 1 jun. 2018.

KURUCZ, K. et al. Aedes koreicus, a vector on the rise: Pan-European genetic patterns, mitochondrial and draft genome sequencing. **PloS One**, v. 17, n. 8, p. e0269880, 2022.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 15 fev. 2001.

LANGILLE, M. G. I. et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. **Nature Biotechnology**, v. 31, n. 9, p. 814–821, set. 2013.

LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biology**, v. 10, n. 3, p. R25, 4 mar. 2009.

LASA, I. et al. Genome-wide antisense transcription drives mRNA processing in bacteria. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 50, p. 20172–20177, 13 dez. 2011.

LAZZARINI, L. et al. First autochthonous dengue outbreak in Italy, August 2020. **Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin**, v. 25, n. 36, p. 2001606, set. 2020.

LEINO, K. R. M.; MÜLLER, P. **A Basis for Verifying Multi-threaded Programs**. (G. Castagna, Ed.)Programming Languages and Systems. **Anais**...: Lecture Notes in Computer Science.Berlin, Heidelberg: Springer, 2009.

LETUNIC, I.; BORK, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. **Nucleic Acids Research**, v. 49, n. W1, p. W293–W296, 2 jul. 2021.

LI, C. et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. **Cell**, v. 137, n. 3, p. 509–521, 1 maio 2009a.

LI, C. et al. Metatranscriptomic Sequencing Reveals Host Species as an Important Factor Shaping the Mosquito Virome. **Microbiology Spectrum**, v. 11, n. 2, p. e04655-22, 2023.

LI, C.-X. et al. Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. **eLife**, v. 4, p. e05378, 29 jan. 2015.

LI, H. et al. The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 15 ago. 2009b.

LI, X.-L.; BLACKFORD, J. A.; HASSEL, B. A. RNase L Mediates the Antiviral Effect of Interferon through a Selective Reduction in Viral RNA during Encephalomyocarditis Virus Infection. **Journal of Virology**, v. 72, n. 4, p. 2752–2759, abr. 1998.

LI, Y. et al. RNA Interference Functions as an Antiviral Immunity Mechanism in Mammals. **Science (New York, N.Y.)**, v. 342, n. 6155, p. 10.1126/science.1241911, 11 out. 2013.

LI, Y. et al. SARS-CoV-2 induces double-stranded RNA-mediated innate immune responses in respiratory epithelial-derived cells and cardiomyocytes. **Proceedings of the National Academy of Sciences**, v. 118, n. 16, p. e2022643118, 20 abr. 2021.

LI, Y. et al. Endogenous Viral Elements in Shrew Genomes Provide Insights into Pestivirus Ancient History. **Molecular Biology and Evolution**, v. 39, n. 10, p. msac190, 7 out. 2022.

LIANG, G.; BUSHMAN, F. D. The human virome: assembly, composition and host interactions. **Nature Reviews Microbiology**, v. 19, n. 8, p. 514–527, ago. 2021.

LIGSAY, A.; TELLE, O.; PAUL, R. Challenges to Mitigating the Urban Health Burden of Mosquito-Borne Diseases in the Face of Climate Change. **International Journal of Environmental Research and Public Health**, v. 18, n. 9, p. 5035, 10 maio 2021.

LIN, Q. et al. Sanitizing agents for virus inactivation and disinfection. **VIEW**, v. 1, n. 2, p. e16, 2020.

LIU, B. et al. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. **Nucleic Acids Research**, v. 43, n. W1, p. W65-71, 1 jul. 2015.

LOURENÇO-DE-OLIVEIRA, R. et al. Culex quinquefasciatus mosquitoes do not support replication of Zika virus. **The Journal of General Virology**, v. 99, n. 2, p. 258–264, fev. 2018.

LOUTEN, J. Chapter 4 - Virus Replication. Em: LOUTEN, J. (Ed.). **Essential Human Virology**. Boston: Academic Press, 2016. p. 49–70.

LU, S. et al. CDD/SPARCLE: the conserved domain database in 2020. **Nucleic Acids Research**, v. 48, n. D1, p. D265–D268, 8 jan. 2020.

LYE, L.-F. et al. A Narnavirus-Like Element from the Trypanosomatid Protozoan Parasite Leptomonas seymouri. **Genome Announcements**, v. 4, n. 4, p. e00713-16, 4 ago. 2016.

MA, M. et al. Discovery of DNA viruses in wild-caught mosquitoes using small RNA high throughput sequencing. **PloS One**, v. 6, n. 9, p. e24758, 2011.

MAATEN, L. VAN DER; HINTON, G. Visualizing Data using t-SNE. **Journal of Machine Learning Research**, v. 1, p. 1–48, 2008.

MAHMOUDABADI, G.; PHILLIPS, R. A comprehensive and quantitative exploration of thousands of viral genomes. **eLife**, v. 7, p. e31955, 6 abr. 2018.

MAILLARD, P. V. et al. Antiviral RNA Interference in Mammalian Cells. **Science (New York, N.Y.)**, v. 342, n. 6155, p. 10.1126/science.1241930, 11 out. 2013.

MAILLARD, P. V. et al. Inactivation of the type I interferon pathway reveals long double-stranded RNA-mediated RNA interference in mammalian cells. **The EMBO journal**, v. 35, n. 23, p. 2505–2518, 1 dez. 2016.

MALATHI, K. et al. Small self-RNA generated by RNase L amplifies antiviral innate immunity. **Nature**, v. 448, n. 7155, p. 816–819, ago. 2007.

MALLA, R. K. et al. Numerical analysis of predatory potentiality of Toxorhynchites splendens against larval Aedes albopictus in laboratory and semi-field conditions. **Scientific Reports**, v. 13, n. 1, p. 7403, 6 maio 2023.

MARQUES, J. T. et al. Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila. **Nature Structural & Molecular Biology**, v. 17, n. 1, p. 24–30, jan. 2010.

MATTHEWS, B. J. et al. Improved reference genome of Aedes aegypti informs arbovirus vector control. **Nature**, v. 563, n. 7732, p. 501–507, nov. 2018.

MAVALE, M. S. et al. Vertical and venereal transmission of Chandipura virus (Rhabdoviridae) by Aedes aegypti (Diptera: Culicidae). **Journal of Medical Entomology**, v. 42, n. 5, p. 909–911, set. 2005.

MAYORAL, J. G. et al. Wolbachia small noncoding RNAs and their role in cross-kingdom communications. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 52, p. 18721–18726, 30 dez. 2014.

MCLEAN, B. J. et al. Differential suppression of persistent insect specific viruses in trans-infected wMel and wMelPop-CLA Aedes-derived mosquito lines. **Virology**, v. 527, p. 141–145, 15 jan. 2019.

MCMENIMAN, C. J. et al. Stable introduction of a life-shortening Wolbachia infection into the mosquito Aedes aegypti. **Science (New York, N.Y.)**, v. 323, n. 5910, p. 141–144, 2 jan. 2009.

MEDLOCK, J. M. et al. An entomological review of invasive mosquitoes in Europe. **Bulletin of Entomological Research**, v. 105, n. 6, p. 637–663, dez. 2015.

MENÉNDEZ-ARIAS, L. Mutation rates and intrinsic fidelity of retroviral reverse transcriptases. **Viruses**, v. 1, n. 3, p. 1137–1165, dez. 2009.

Microbiology by numbers. **Nature Reviews Microbiology**, v. 9, n. 9, p. 628–628, set. 2011.

MIESEN, P.; JOOSTEN, J.; RIJ, R. P. VAN. PIWIs Go Viral: Arbovirus-Derived piRNAs in Vector Mosquitoes. **PLOS Pathogens**, v. 12, n. 12, p. e1006017, 29 dez. 2016.

MISOF, B. et al. Phylogenomics resolves the timing and pattern of insect evolution. **Science**, v. 346, n. 6210, p. 763–767, 7 nov. 2014.

MISTRY, J. et al. Pfam: The protein families database in 2021. **Nucleic Acids Research**, v. 49, n. D1, p. D412–D419, 8 jan. 2021.

MLAKAR, J. et al. Zika Virus Associated with Microcephaly. **New England Journal of Medicine**, v. 374, n. 10, p. 951–958, 10 mar. 2016.

MOHN, F.; HANDLER, D.; BRENNECKE, J. piRNA-guided slicing specifies transcripts for Zucchini dependent, phased piRNA biogenesis. **Science (New York, N.Y.)**, v. 348, n. 6236, p. 812–817, 15 maio 2015.

MOREIRA, D.; LÓPEZ-GARCÍA, P. Ten reasons to exclude viruses from the tree of life. **Nature Reviews Microbiology**, v. 7, n. 4, p. 306–311, abr. 2009.

MOREIRA, L. A. et al. A Wolbachia Symbiont in Aedes aegypti Limits Infection with Dengue, Chikungunya, and Plasmodium. **Cell**, v. 139, n. 7, p. 1268–1278, 24 dez. 2009.

MORTELMANS, K.; WANG-JOHANNING, F.; JOHANNING, G. L. The role of human endogenous retroviruses in brain development and function. **APMIS: acta pathologica, microbiologica, et immunologica Scandinavica**, v. 124, n. 1–2, p. 105–115, 2016.

NACCACHE, S. N. et al. The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. **Journal of Virology**, v. 87, n. 22, p. 11966–11977, 15 nov. 2013.

NARITA, S. et al. Unexpected Mechanism of Symbiont-Induced Reversal of Insect Sex: Feminizing Wolbachia Continuously Acts on the Butterfly Eurema hecabe during Larval Development. **Applied and Environmental Microbiology**, v. 73, n. 13, p. 4332–4341, jul. 2007.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION. **Astrobiology at NASA: Life in the universe**. Disponível em: <https://astrobiology.nasa.gov/research/life-detection/about/>. Acesso em: 16 jul. 2023.

NAYAK, A. et al. RNA interference-mediated intrinsic antiviral immunity in invertebrates. **Current Topics in Microbiology and Immunology**, v. 371, p. 183–200, 2013.

NISHIMASU, H. et al. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. **Nature**, v. 491, n. 7423, p. 284–287, nov. 2012.

NUGENT, T.; JONES, D. T. Transmembrane protein topology prediction using support vector machines. **BMC Bioinformatics**, v. 10, n. 1, p. 159, 26 maio 2009.

OLMO, R. P. et al. Mosquito vector competence for dengue is modulated by insect-specific viruses. **Nature Microbiology**, v. 8, n. 1, p. 135–149, jan. 2023.

O'NEILL, S. L. et al. 16S rRNA phylogenetic analysis of the bacterial endosymbionts associated with cytoplasmic incompatibility in insects. **Proceedings of the National Academy of Sciences of the United States of America**, v. 89, n. 7, p. 2699–2702, 1 abr. 1992.

ORTIZ-BAEZ, A. S. et al. RNA virome diversity and Wolbachia infection in individual Drosophila simulans flies. **The Journal of General Virology**, v. 102, n. 10, p. 001639, 27 out. 2021.

OUTAMMASSINE, A.; ZOUHAIR, S.; LOQMAN, S. Global potential distribution of three underappreciated arboviruses vectors (*Aedes japonicus* , *Aedes vexans* and *Aedes vittatus*) under current and future climate conditions. **Transboundary and Emerging Diseases**, v. 69, n. 4, jul. 2022.

PALATINI, U. et al. Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus. **BMC Genomics**, v. 18, p. 512, 5 jul. 2017.

PALATINI, U. et al. Improved reference genome of the arboviral vector Aedes albopictus. **Genome Biology**, v. 21, n. 1, p. 215, 26 ago. 2020.

PARRY, R.; ASGARI, S. Aedes Anphevirus: an Insect-Specific Virus Distributed Worldwide in Aedes aegypti Mosquitoes That Has Complex Interplays with Wolbachia and Dengue Virus Infection in Cells. **Journal of Virology**, v. 92, n. 17, p. e00224-18, 1 set. 2018.

PETIT, M. et al. piRNA pathway is not required for antiviral defense in Drosophila melanogaster. **Proceedings of the National Academy of Sciences**, v. 113, n. 29, p. E4218–E4227, 19 jul. 2016.

PIMENTEL, A. C. et al. The Antiviral Effects of the Symbiont Bacteria Wolbachia in Insects. **Frontiers in Immunology**, v. 11, 2021.

PINTO, S. B. et al. Effectiveness of Wolbachia-infected mosquito deployments in reducing the incidence of dengue and other Aedes-borne diseases in Niterói, Brazil: A quasi-experimental study. **PLOS Neglected Tropical Diseases**, v. 15, n. 7, p. e0009556, 12 jul. 2021.

POTTER, S. C. et al. HMMER web server: 2018 update. **Nucleic Acids Research**, v. 46, n. W1, p. W200–W204, 2 jul. 2018.

POWELL, J. R.; TABACHNICK, W. J. History of domestication and spread of Aedes aegypti - A Review. **Memórias do Instituto Oswaldo Cruz**, v. 108, n. suppl 1, p. 11–17, 2013.

PROSDOCIMI, F. et al. Decoding viruses: An alternative perspective on their history, origins and role in nature. **Biosystems**, v. 231, p. 104960, set. 2023.

PUTRI, G. H. et al. Analysing high-throughput sequencing data in Python with HTSeq 2.0. **Bioinformatics**, v. 38, n. 10, p. 2943–2945, 13 maio 2022.

RAND, T. A. et al. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. **Cell**, v. 123, n. 4, p. 621–629, 18 nov. 2005.

RAOULT, D.; FORTERRE, P. Redefining viruses: lessons from Mimivirus. **Nature Reviews. Microbiology**, v. 6, n. 4, p. 315–319, abr. 2008.

RETALLACK, H. et al. Persistence of Ambigrammatic Narnaviruses Requires Translation of the Reverse Open Reading Frame. **Journal of Virology**, v. 95, n. 13, p. e0010921, 10 jun. 2021.

REUTER, J. S.; MATHEWS, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. **BMC Bioinformatics**, v. 11, n. 1, p. 129, 15 mar. 2010.

RIAZ, A. et al. Ovine herpesvirus-2 encoded microRNAs target virus genes involved in virus latency. **The Journal of general virology**, v. 95, n. Pt 2, p. 472–480, fev. 2014.

ROBERT, X.; GOUET, P. Deciphering key features in protein structures with the new ENDscript server. **Nucleic Acids Research**, v. 42, n. Web Server issue, p. W320-324, jul. 2014.

ROOSSINCK, M. J. The good viruses: viral mutualistic symbioses. **Nature Reviews Microbiology**, v. 9, n. 2, p. 99–108, fev. 2011.

ROOVERS, E. F. et al. Piwi proteins and piRNAs in mammalian oocytes and early embryos. **Cell Reports**, v. 10, n. 12, p. 2069–2082, 31 mar. 2015.

ROSENBERG, R. et al. Search strategy has influenced the discovery rate of human viruses. **Proceedings of the National Academy of Sciences**, v. 110, n. 34, p. 13961–13964, 20 ago. 2013.

ROSS, P. A. et al. An elusive endosymbiont: Does Wolbachia occur naturally in Aedes aegypti? **Ecology and Evolution**, v. 10, n. 3, p. 1581–1591, 16 jan. 2020.

ROUNDY, C. M. et al. Insect-Specific Viruses: A Historical Overview and Recent Developments. **Advances in Virus Research**, v. 98, p. 119–146, 2017.

ROUX, S.; MATTHIJNSSENS, J.; DUTILH, B. E. Metagenomics in Virology. **Encyclopedia of Virology**, p. 133–140, 2021.

SADRAEIAN, M. et al. Viral inactivation by light. **eLight**, v. 2, n. 1, p. 18, 26 set. 2022.

SALYER, S. J. et al. Prioritizing Zoonoses for Global Health Capacity Building—Themes from One Health Zoonotic Disease Workshops in 7 Countries, 2014–2016. **Emerging Infectious Diseases**, v. 23, n. 13, dez. 2017.

SAMUEL, G. H. et al. Yellow fever virus capsid protein is a potent suppressor of RNA silencing that binds double-stranded RNA. **Proceedings of the National Academy of Sciences of the United States of America**, v. 113, n. 48, p. 13863–13868, 29 nov. 2016.

SANJUÁN, R. From Molecular Genetics to Phylodynamics: Evolutionary Relevance of Mutation Rates Across Viruses. **PLOS Pathogens**, v. 8, n. 5, p. e1002685, 3 maio 2012.

SANJUÁN, R.; DOMINGO-CALAP, P. Mechanisms of viral mutation. **Cellular and Molecular Life Sciences**, v. 73, n. 23, p. 4433–4448, 2016.

SCHNETTLER, E. et al. Wolbachia restricts insect-specific flavivirus infection in Aedes aegypti cells. **The Journal of General Virology**, v. 97, n. 11, p. 3024–3029, 10 nov. 2016.

SCHRODINGER, E. **What is Life? The physical aspect of the living cell**. [s.l.] Cambridge University Press, 1951.

SHABALINA, S. A.; KOONIN, E. V. Origins and evolution of eukaryotic RNA interference. **Trends in ecology & evolution**, v. 23, n. 10, p. 578–587, out. 2008.

SHI, M. et al. Redefining the invertebrate RNA virosphere. **Nature**, v. 540, n. 7634, p. 539–543, dez. 2016.

SICARD, A. et al. The Strange Lifestyle of Multipartite Viruses. **PLOS Pathogens**, v. 12, n. 11, p. e1005819, 3 nov. 2016.

SIDDELL, S. G. et al. Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV). **Journal of General Virology**, v. 104, n. 5, p. 001840, 2023.

SIEVERS, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. **Molecular Systems Biology**, v. 7, p. 539, 11 out. 2011.

SIMMONDS, P. et al. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla –selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. **BMC Genomics**, v. 14, n. 1, p. 610, 10 set. 2013.

SIMMONDS, P. et al. ICTV Virus Taxonomy Profile: Flaviviridae. **Journal of General Virology**, v. 98, n. 1, p. 2–3, 2017.

SIMMONDS, P. et al. Four principles to establish a universal virus taxonomy. **PLOS Biology**, v. 21, n. 2, p. e3001922, 13 fev. 2023.

SINHA, A. et al. Complete Genome Sequence of the Wolbachia wAlbB Endosymbiont of Aedes albopictus. **Genome Biology and Evolution**, v. 11, n. 3, p. 706–720, 1 mar. 2019.

SLONCHAK, A.; KHROMYKH, A. A. Subgenomic flaviviral RNAs: What do we know after the first decade of research. **Antiviral Research**, v. 159, p. 13–25, nov. 2018.

SNYDER, T. The Mosquito: A Human History of Our Deadliest Predator. **Emerging Infectious Diseases**, v. 26, n. 10, p. 2536, out. 2020.

SOFUKU, K. et al. Influence of Endogenous Viral Sequences on Gene Expression. Em: **Gene Expression and Regulation in Mammalian Cells - Transcription From General Aspects**. [s.l.] IntechOpen, 2018.

SPERSCHNEIDER, J.; DATTA, A. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. **Nucleic Acids Research**, v. 38, n. 7, p. e103, 1 abr. 2010.

STEINHAUER, D. A.; DOMINGO, E.; HOLLAND, J. J. Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. **Gene**, v. 122, n. 2, p. 281–288, 15 dez. 1992.

STOLLAR, V.; THOMAS, V. L. An agent in the Aedes aegypti cell line (Peleg) which causes fusion of Aedes albopictus cells. **Virology**, v. 64, n. 2, p. 367–377, abr. 1975.

SUZUKI, R.; SHIMODAIRA, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. **Bioinformatics**, v. 22, n. 12, p. 1540–1542, 15 jun. 2006.

SUZUKI, Y. et al. Non-retroviral Endogenous Viral Element Limits Cognate Virus Replication in Aedes aegypti Ovaries. **Current biology: CB**, v. 30, n. 18, p. 3495- 3506.e6, 21 set. 2020.

TAO, Y. J.; YE, Q. RNA Virus Replication Complexes. **PLOS Pathogens**, v. 6, n. 7, p. e1000943, 22 jul. 2010.

TEICH, V.; ARINELLI, R.; FAHHAM, L. Aedes aegypti e sociedade: o impacto econômico das arboviroses no Brasil. **Jornal Brasileiro de Economia da Saúde**, v. 9, n. 3, p. 267–276, dez. 2017.

TEIXEIRA, L.; FERREIRA, Á.; ASHBURNER, M. The Bacterial Symbiont Wolbachia Induces Resistance to RNA Viral Infections in Drosophila melanogaster. **PLOS Biology**, v. 6, n. 12, p. e1000002, 23 dez. 2008.

TER HORST, A. M. et al. Endogenous Viral Elements Are Widespread in Arthropod Genomes and Commonly Give Rise to PIWI-Interacting RNAs. **Journal of Virology**, v. 93, n. 6, p. e02124-18, 15 mar. 2019.

THAKUR, A.; KUMAR, M. AntiVIRmiR: A repository of host antiviral miRNAs and their expression along with experimentally validated viral miRNAs and their targets. **Frontiers in Genetics**, v. 13, p. 971852, 8 set. 2022.

THOMPSON. **The Diptera Site: The biosystematic database of world Diptera. Nomenclature status statistics.** Database. Disponível em: <https://diptera.myspecies.info/>. Acesso em: 11 jul. 2023.

TYCOWSKI, K. T. et al. Viral noncoding RNAs: more surprises. **Genes & Development**, v. 29, n. 6, p. 567–584, 15 mar. 2015.

UMBACH, J. L. et al. MicroRNAs expressed by herpes simplex virus 1 during latent infection regulate viral mRNAs. **Nature**, v. 454, n. 7205, p. 780–783, 7 ago. 2008.

VAINIO, E. J. et al. Diagnosis and discovery of fungal viruses using deep sequencing of small RNAs. **Journal of General Virology**, v. 96, n. 3, p. 714–725, 2015.

VALENTINE, M. J.; MURDOCK, C. C.; KELLY, P. J. Sylvatic cycles of arboviruses in non-human primates. **Parasites & Vectors**, v. 12, n. 1, p. 463, 2 out. 2019.

VARJAK, M.; LEGGEWIE, M.; SCHNETTLER, E. The antiviral piRNA response in mosquitoes? **Journal of General Virology**, v. 99, n. 12, p. 1551–1562, 2018.

VINGA, S. Editorial: Alignment-free methods in computational biology. **Briefings in Bioinformatics**, v. 15, n. 3, p. 341–342, 1 maio 2014.

VIRALZONE. **Human viruses and associated pathologies**. Disponível em: <https://viralzone.expasy.org/678>. Acesso em: 19 jul. 2023.

WAGIH, O. ggseqlogo: a versatile R package for drawing sequence logos. **Bioinformatics**, v. 33, n. 22, p. 3645–3647, 15 nov. 2017.

WANG, W. et al. The initial uridine of primary piRNAs does not create the tenth adenine that Is the hallmark of secondary piRNAs. **Molecular Cell**, v. 56, n. 5, p. 708–716, 4 dez. 2014.

WEAVER, S. C.; REISEN, W. K. Present and future arboviral threats. **Antiviral Research**, v. 85, n. 2, p. 328–345, fev. 2010.

WHITFIELD, Z. J. et al. The diversity, structure and function of heritable adaptive immunity sequences in the Aedes aegypti genome. **Current biology : CB**, v. 27, n. 22, p. 3511- 3519.e7, 20 nov. 2017.

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

WICKHAM, H. ggplot2. **WIREs Computational Statistics**, v. 3, n. 2, p. 180–185, 2011.

WIGINGTON, C. H. et al. Re-examination of the relationship between marine virus and microbial cell abundances. **Nature Microbiology**, v. 1, p. 15024, 25 jan. 2016.

WILDER-SMITH, A. et al. Epidemic arboviral diseases: priorities for research and public health. **The Lancet Infectious Diseases**, v. 17, n. 3, p. e101–e106, mar. 2017.

WILKERSON, R. C. et al. Making Mosquito Taxonomy Useful: A Stable Classification of Tribe Aedini that Balances Utility with Current Knowledge of Evolutionary Relationships. **PLOS ONE**, v. 10, n. 7, p. e0133602, 30 jul. 2015.

WILKINS, M. R. et al. Protein identification and analysis tools in the ExPASy server. **Methods in Molecular Biology (Clifton, N.J.)**, v. 112, p. 531–552, 1999.

WOESE, C. R.; FOX, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 11, p. 5088–5090, nov. 1977.

WORLD HEALTH ORGANIZATION. **Global vector control response 2017-2030**. Geneva: World Health Organization, 2017.

XIAO, C. et al. Cryo-electron Microscopy of the Giant Mimivirus. **Journal of Molecular Biology**, v. 353, n. 3, p. 493–496, out. 2005.

YANG, Z. et al. Approaches for Studying MicroRNA and Small Interfering RNA Methylation In Vitro and In Vivo. **Methods in enzymology**, v. 427, p. 139–154, 2007.

ZADRA, N.; RIZZOLI, A.; ROTA-STABELLI, O. Chronological Incongruences between Mitochondrial and Nuclear Phylogenies of Aedes Mosquitoes. **Life**, v. 11, n. 3, p. 181, mar. 2021.

ZELLER, H.; VAN BORTEL, W.; SUDRE, B. Chikungunya: Its History in Africa and Asia and Its Spread to New Regions in 2013–2014. **Journal of Infectious Diseases**, v. 214, n. suppl 5, p. S436–S440, 15 dez. 2016.

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. **Genome Research**, v. 18, n. 5, p. 821–829, maio 2008.

ZHANG, G. et al. Cell fusing agent virus and dengue virus mutually interact in Aedes aegypti cell lines. **Scientific Reports**, v. 7, n. 1, p. 6935, 31 jul. 2017.

ZHOU, A.; HASSEL, B. A.; SILVERMAN, R. H. Expression cloning of 2-5A-dependent RNAase: a uniquely regulated mediator of interferon action. **Cell**, v. 72, n. 5, p. 753–765, 12 mar. 1993.

ZUG, R.; HAMMERSTEIN, P. Still a host of hosts for Wolbachia: analysis of recent data suggests that 40% of terrestrial arthropod species are infected. **PloS One**, v. 7, n. 6, p. e38544, 2012.

10 - Produção acadêmica durante o doutorado

Trabalhos publicados relacionados ao tema da tese

1- Abbo, Sandra R., ***João PP de Almeida**, Roenick P. Olmo, Carlijn Balvers, Jet S. Griep, Charlotte Linthout, Constantianus JM Koenraadt et al. **The virome of the invasive Asian bush mosquito Aedes japonicus in Europe**. *Virus Evolution* 9, no. 2 (2023): vead041.
Primeira autoria compartilhada

2**- de Almeida, João PP**, Eric RGR Aguiar, Juliana N. Armache, Roenick P. Olmo, and João T. Marques. **The virome of vector mosquitoes**. *Current Opinion in Virology* 49 (2021): 7-12.

3- Eric Roberto Guimarães Rocha, **João Paulo Pereira de Almeida**, Lucio Rezende Queiroz, Liliane Santana Oliveira, Roenick Proveti Olmo, Isaque João da Silva de Faria, Jean-Luc Imler, Arthur Gruber, Benjamin J. Matthews, and João Trindade Marques. **A single unidirectional piRNA cluster similar to the flamenco locus is the major source of EVE-derived transcription and small RNAs in *Aedes aegypti* mosquitoes**. *RNA* 26, no. 5 (2020): 581-594.

4- Olmo, Roenick P., Yaovi MH Todjro, Eric RGR Aguiar, **João Paulo P. de Almeida**, Flávia V. Ferreira, Juliana N. Armache, Isaque JS de Faria et al. **Mosquito vector competence for dengue is modulated by insect-specific viruses**. *Nature Microbiology* 8, no. 1 (2023): 135-149.

5- I. Sardi, Silvia, Rejane H. Carvalho, Luis G. C. Pacheco, **João P. P. d. Almeida**, Emilia M. M. d. A. Belitardo, Carina S. Pinheiro, Gúbio S. Campos, and Eric RGR Aguiar. **High-quality resolution of the outbreak-related Zika virus genome and discovery of new viruses using ion torrent-based metatranscriptomics**. *Viruses* 12, no. 7 (2020): 782.

Trabalhos publicados de colorações durante o enfrentamento a pandemia do SARS-CoV-2

1- Dos Santos, Letícia Adrielle, Pedro Germano de Góis Filho, Ana Maria Fantini Silva, João Victor Gomes Santos, Douglas Siqueira Santos, Marília Marques Aquino, Rafaela Mota de Jesus et al. **Recurrent COVID-19 including evidence of reinfection and enhanced severity in thirty Brazilian healthcare workers**. *Journal of Infection* 82, no. 3 (2021): 399-406.

2- Carlos, Renata Santiago Alberto, Ana Paula Melo Mariano, Bianca Mendes Maciel, Sandra Rocha Gadelha, Mylene de Melo Silva, Emilia Maria Medeirosde Andrade Belitardo, Danilo Jobim Passos Gil Rocha et al. **First genome sequencing of SARS-CoV-2 recovered from an infected cat and its owner in Latin America**. *Transboundary and emerging diseases* 68, no. 6 (2021): 3070-3074.

3- Campos, Gubio S., Silvia I. Sardi, Melissa B. Falcao, Emilia MMA Belitardo, Danilo JPG Rocha, Carolina A. Rolo, Aline D. Menezes et al. **Ion torrent-based nasopharyngeal swab metatranscriptomics in COVID-19**. *Journal of Virological Methods* 282 (2020): 113888.

Trabalhos publicados de outras colaborações

1- Fonseca, Paula LC, Joel AM Porto, Juliana N. Armache, **João Paulo P. de Almeida**, Felipe F. da Silva, Roenick P. Olmo, Isaque J. da S. Faria et al. **Genome-wide identification of miRNAs and target regulatory network in the invasive ectoparasitic mite Varroa destructor**. *Genomics* 113, no. 4 (2021): 2290-2303.

2- Gabriela B. Caldas-Garcia, Vinícius Castro Santos, Paula Luize Camargos Fonseca, **João Paulo Pereira Almeida**, Marco Antonio Costa, Eric Roberto Guimarães Rocha Aguiar. **The Viromes of Six Ecosystem Service Provider Parasitoid Wasps.** *Aceito na Revista Viruses* (2023)

3- Silva, Bruno M., Lucianna H. Santos, **João Paulo P. de Almeida**, and Mariana TQ de Magalhães. **Rad5 HIRAN domain: Structural insights into its interaction with ssDNA through molecular modeling approaches**. *Journal of Biomolecular Structure and Dynamics* 41, no. 7 (2023): 3062-3075.

Manuscritos em fase de submissão

1- **João Paulo Pereira de Almeida**; Juliana N. Armache; Roenick Proveti Olmo; Paula Luize Camargos Fonseca; Isaque João da Silva de Faria; Weider Cristiano Santana; Bergmann M. Ribeiro; João Trindade Marques; Marco Antônio Costa; Eric Roberto Guimarães Rocha Aguiar.
Identification of Lake Sinai virus and its induced RNAi immune response in Brazilian honeybees (*Apis mellifera*). *A ser submetido para revista Viruses em 2023***.**

2- Amanda de Freitas, Fernanda Rezende, Silvana de Mendonça, Lívia Baldon, Emanuel Silva, Flávia Ferreira, **João Paulo de Almeida**, Siad Amadou, Bruno Marçal, Sara Comini, Marcele Rocha, Hegger Fritsch, Marta Giovanetti, Luiz Alcântara, Luciano Moreira, Alvaro Ferreira.
High Transmission Efficiency of Chikungunya Virus by Brazilian *Aedes aegypti* Mosquitoes. *Artigo submetido para revista Emerging Infectious Diseases – CDC 2023*

11 – Apêndices

Tabelas Suplementares

Tabela Suplementar 1 – Dados de coleta dos mosquitos, protocolos de preparo e sequenciamento das 122 bibliotecas de pequenos RNAs sequenciadas de amostras de 10 espécies de mosquitos.

Tabela Suplementar 2 – Classificação dos contigs montados com (filtrado) ou sem (não-filtrado) remoção dos reads alinhados aos genomas de mosquitos nas 122 bibliotecas do projeto ZikaAlliance.

Tabela Suplementar 3 – Teste de execução da ferramenta *Small RNA Metavir* em bibliotecas públicas de pequenos RNAs

As três tabelas Suplementares citadas na tese se encontram no link:

https://github.com/JPbio/Tabelas_SUplementares_Tese