UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE MEDICINA
DEPARTAMENTO DE MEDICINA CLÍNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS MÉDICAS

PEDRO MIGUEL CARNEIRO JERONIMO

MONITORAMENTO DA EVOLUÇÃO VIRAL E APRIMORAMENTO DE ESTRATÉGIAS DE SEQUENCIAMENTO PARA O VIRUS DA DENGUE NO BRASIL

FORTALEZA
2025

PEDRO MIGUEL CARNEIRO JERONIMO

MONITORAMENTO DA EVOLUÇÃO VIRAL E APRIMORAMENTO DE ESTRATÉGIAS DE SEQUENCIAMENTO PARA O VIRUS DA DENGUE NO BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Médicas da Faculdade de Medicina da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Ciências Médicas.

Orientador: Prof. Dr. Fábio Miyajima

FORTALEZA

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

J54m Jeronimo, Pedro Miguel Carneiro.
Monitoramento da evolução viral e aprimoramento de estratégias de sequenciamento para o vírus da dengue no Brasil / Pedro Miguel Carneiro Jeronimo. – 2025.
89 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Faculdade de Medicina, Programa de Pós-Graduação em Ciências Médicas, Fortaleza, 2025.
Orientação: Prof. Dr. Fábio Miyajima.

1. Variação Genética. 2. Painéis de sequenciamento. 3. Vigilância epidemiológica. 4. Síndrome febril. I. Título.

CDD 610

PEDRO MIGUEL CARNEIRO JERONIMO

MONITORAMENTO DA EVOLUÇÃO VIRAL E APRIMORAMENTO DE ESTRATÉGIAS DE SEQUENCIAMENTO PARA O VIRUS DA DENGUE NO BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Médicas da Faculdade de Medicina da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Ciências Médicas.

Orientador: Prof. Dr. Fábio Miyajima

Aprovado em: __/__/____

BANCA EXAMINADORA

_____

Dr. Fabio Miyajima (Orientador)
Universidade Federal do Ceará (UFC)

_____

Dr. Roberto Dias Lins Neto
Instituto Aggeu Magalhães (IAM)

_____

Dr. Leonardo Soares Basto
Instituto Oswaldo Cruz (Fiocruz)

_____

Dr. Felipe Gomes Naveca (Suplente)
Instituto Oswaldo Cruz (Fiocruz)

RESUMO

A diversidade genética do vírus da dengue (DENV) no Brasil apresenta desafios para a vigilância epidemiológica e o controle da doença. Este estudo investigou a variabilidade genética dos sorotipos I e II do DENV, com o objetivo de desenvolver painéis de sequenciamento genômico otimizados para o cenário brasileiro e avaliar o impacto dessa diversidade em estratégias de vigilância. Os resultados mostraram uma significativa variabilidade genética regional, afetando a compatibilidade dos *primers* de sequenciamento. Essa análise genética também evidenciou mutações específicas em diferentes regiões e períodos, ressaltando a importância de considerar o contexto local para a interpretação dos resultados. Painéis desenvolvidos especificamente para o Brasil, levando em conta essa diversidade, demonstraram, *in sílico*, desempenho superior em comparação aos painéis internacionais. Além disso, a escolha do tamanho do *amplicon* (400, 600 ou 800 pb) revelou-se um fator crucial, necessitando equilíbrio entre a qualidade das amostras e a durabilidade dos painéis, especialmente em amostras degradadas. Os painéis apresentaram coberturas genômicas médias de 97,7%, 97,3% e 96,6%, respectivamente, com falhas completas observadas em duas amostras e baixa cobertura em outra, independentemente do protocolo utilizado. A análise também identificou um número reduzido de *primers* defeituosos, sem um padrão claro entre os diferentes tamanhos de *amplicon*. A introdução do genótipo cosmopolita (DENV2-II) pode ter contribuído para o aumento de casos, embora a possibilidade de reinfecção pelo mesmo sorotipo ainda precise ser explorada em estudos futuros. Como parte do estudo, foi desenvolvido o software Hydra, voltado para o monitoramento da qualidade do sequenciamento. O Hydra se mostrou uma ferramenta fundamental ao analisar a cobertura, detectar deleções e identificar variações nas regiões de *primer*, sendo essencial para a vigilância genômica contínua do DENV. Por fim, a integração de dados epidemiológicos, sociodemográficos, ambientais e econômicos provou-se vital para o controle eficaz da dengue. Embora a modelagem preditiva de surtos ainda apresente desafios, o estudo reforça a importância da otimização dos painéis de sequenciamento e do monitoramento constante da qualidade dos dados, essenciais para aprimorar a resposta à dengue no Brasil.

Palavras-chave: Variação Genética, painéis de sequenciamento, vigilância epidemiológica, síndrome febril.

ABSTRACT

The genetic diversity of the dengue virus (DENV) in Brazil poses challenges for epidemiological surveillance and disease control. This study investigated the genetic variability of DENV serotypes I and II, aiming to develop genomic sequencing panels optimized for the Brazilian context and to assess the impact of this diversity on surveillance strategies. The results showed significant regional genetic variability, affecting the compatibility of sequencing *primers*. This genetic analysis further revealed specific mutations across different regions and time periods, highlighting the importance of considering the local context when interpreting results. Panels specifically developed for Brazil, considering this diversity, demonstrated superior in silico performance compared to international panels. Additionally, the choice of *amplicon* size (400, 600, or 800 bp) proved to be a crucial factor, requiring a balance between sample quality and panel durability, especially for degraded samples. The panels achieved average genomic coverages of 97.7%, 97.3%, and 96.6%, respectively, with complete failures observed in two samples and low coverage in another, regardless of the protocol used. The analysis also identified a small number of defective *primers*, with no clear pattern among the different *amplicon* sizes. The introduction of the cosmopolitan genotype (DENV2-II) may have contributed to the increase in cases, although the possibility of reinfection by the same serotype still needs to be explored in future studies. As part of this study, the Hydra software was developed to monitor sequencing quality. Hydra proved to be a fundamental tool for analyzing coverage, detecting deletions, and identifying variations in *primer* regions, making it essential for continuous genomic surveillance of DENV. Finally, the integration of epidemiological, sociodemographic, environmental, and economic data proved vital for the effective control of dengue. Although predictive outbreak modeling still presents challenges, this study reinforces the importance of optimizing sequencing panels and continuously monitoring data quality, which are essential for improving dengue response efforts in Brazil.

Keywords: genetic diversity, sequencing panels, epidemiological surveillance, febrile syndrome.

LISTA DE ABREVIATURAS E SIGLAS

AA	Aminoácidos
AR	Modelo Autorregressivo
ARMA	Modelo Autoregressivo de Média Móvel
ARIMA	Modelo Autoregressivo Integrado de Média Móvel
BAM	Formato Binário de Alinhamento/Mapa
BED	*Browser Extensible Data*
BGZF	*Blocked* GNU Zip *Format*
C	Proteína Capsídeo
CIDACS	Centro de Integração de Dados e Conhecimentos para Saúde
CNES	Cadastro Nacional de Estabelecimentos de Saúde
CNES ST	Estabelecimentos de Saúde
CSV	*Comma-Separated Values*
DENV	Vírus da Dengue
DENV1	Vírus da Dengue sorotipo 1
DENV1-V	Vírus da Dengue sorotipo 1, genótipo V
DENV2	Vírus da Dengue sorotipo 2
DENV2-II	Vírus da Dengue sorotipo 2, genótipo II (cosmopolita)
DENV2-III	Vírus da Dengue sorotipo 2, genótipo III (*Asian-American*)
DDoS	Ataque de Negação de Serviço Distribuído (do inglês *Distributed Denial of Service*)
DNA	Ácido Desoxirribonucleico
E	Proteína Envelope
EDI	Domínio central da proteína E
EDII	Domínio de dimerização da proteína E
EDIII	Domínio terminal C da proteína E, similar a imunoglobulina
EDTA	Laboratório de Ecologia de Doenças Transmissíveis da Amazônia
ELISA	Ensaio Imunoenzimático (do inglês *Enzyme-Linked Immunosorbent Assay*)
FASTA	Formato de arquivo para sequências biológicas
FastQ	Formato de arquivo para sequências de nucleotídeos e seus *scores* de qualidade
Fiocruz	Fundação Oswaldo Cruz
G	Comprimento do genoma do vírus
GISAID	Iniciativa Global para Compartilhamento de Todos os Dados de Influenza (do inglês *Global Initiative on Sharing All Influenza Data*)
GISAID EpiArbo	Plataforma do GISAID para dados de arbovírus
Hydra	*High-throughput Yield Data Read Analyzer, software* para visualização de arquivos BAM
HyINDEL	*Software* para detecção de inserções e deleções em dados de sequenciamento
IBGE	Instituto Brasileiro de Geografia e Estatística
IBGE POP	Dados de população do IBGE
IBGE PIB	Dados de Produto Interno Bruto (PIB) do IBGE
IgG	Imunoglobulina G

IgM	Imunoglobulina M
iiq	Intervalo interquartil
IGV	*Integrative Genomics Viewer, software* para visualização de dados genômicos
INMET	Instituto Nacional de Meteorologia
JSON	JavaScript *Object Notation*
Kb	Quilobases
Lacen	Laboratório Central de Saúde Pública
LARBOH	Laboratório de Arbovírus e Vírus Hemorrágicos
M	Número de mutações em região de *primer*
MA	Modelo de Média Móvel
Mafft	*Software* para alinhamento múltiplo de sequências
MapBiomas	Plataforma de mapeamento de cobertura e uso da terra
mcv	Mediana da cobertura vertical
minia	*Software* para montagem de genomas
minimap2	*Software* para alinhamento de sequências
Miseq	Sequenciador de DNA da Illumina
MMA	Município, Mês, Ano
MongoDB	Sistema de banco de dados NoSQL orientado a documentos
mosdepth	*Software* para análise de profundidade de sequenciamento
Mosqlimate	Plataforma de modelagem climática e saúde
MPOS	Posição mais à esquerda baseada em 1 do parceiro
MRNM	Nome do parceiro de referência (‘=’ se o mesmo que RNAME)
MS	Ministério da Saúde
MySQL	Sistema de gerenciamento de banco de dados relacional de código aberto
N	Bases ambíguas em sequências de DNA
NGS	Sequenciamento de Nova Geração (do inglês *Next-Generation Sequencing*)
NS1	Proteína não estrutural 1
NS2A	Proteína não estrutural 2A
NS2B	Proteína não estrutural 2B
NS3	Proteína não estrutural 3
NS4A	Proteína não estrutural 4A
NS4B	Proteína não estrutural 4B
NS5	Proteína não estrutural 5
NUC	Nucleotídeos
OOT	*Out-of-Time*, período de tempo posterior ao usado em treino e teste de modelos de aprendizado de máquina
PAHO	Organização Pan-Americana da Saúde (do inglês *Pan American Health Organization*)
PCR	Reação em Cadeia da Polimerase (do inglês *Polymerase Chain Reaction*)
Phred	Escala de qualidade de sequenciamento de DNA
Plotly	Biblioteca para geração de gráficos interativos
POS	Posição mais à esquerda baseada em 1 do alinhamento truncado
PostgreSQL	Sistema de gerenciamento de banco de dados relacional de código aberto

prM	Proteína precursora da Membrana
Primal Scheme	Plataforma *online* para desenho de *primers*
Python	Linguagem de programação
QNAME	Nome da consulta do *read* ou do par de *reads*
QUAL	Qualidade da consulta (ASCII-33=Qualidade da base Phred)
RAM	Memória de Acesso Aleatório (do inglês *Random Access Memory*)
Redis	Sistema de gerenciamento de banco de dados não relacional para armazenamento de dados em memória
RNA	Ácido Ribonucleico
RT-PCR	Reação em Cadeia da Polimerase com Transcrição Reversa (do inglês *Reverse Transcription Polymerase Chain Reaction*)
SAM	*Sequence Alignment/Map*
samtools	*Software* para manipulação de arquivos SAM/BAM
SARIMA	Modelo Autorregressivo Integrado de Média Móvel Sazonal
SARS-CoV-2	Síndrome Respiratória Aguda Grave Coronavírus 2
SIM	Sistema de Informação sobre Mortalidade
SIM DO	Sistema de Informação sobre Mortalidade por Doenças de Notificação Compulsória
SINAN	Sistema de Informação de Agravos de Notificação
SINAN Dengue	Sistema de Informação de Agravos de Notificação para Dengue
SQL	Linguagem de Consulta Estruturada (do inglês *Structured Query Language*)
SQL Injection	Técnica de ataque que explora vulnerabilidades em bancos de dados SQL
Streamlit	*Framework* para criação de aplicativos web
TE	Tris-EDTA, solução tampão
Tm	Temperatura de *melting*
WAF	*Firewall* de Aplicação Web (do inglês *Web Application Firewall*)
WHO	Organização Mundial da Saúde (do inglês *World Health Organization*)
zlib	Biblioteca de compressão de dados
2K	Região do genoma do DENV

SUMÁRIO

1. INTRODUÇÃO

A dengue, uma doença viral transmitida por mosquitos, representa um sério desafio para a saúde pública no Brasil, caracterizada por surtos recorrentes e elevadas taxas de incidência. A América do Sul, incluindo o Brasil, é uma das regiões mais afetadas, com um número expressivo de casos anualmente (Bhatt *et al.*, 2013). Ademais, com o aumento da temperatura do planeta devido ao aquecimento global, há crescente espaço geográfico suscetível para o completo ciclo de vida do vetor e, consequentemente, expansão do número possível de casos (Delrieu *et al.*, 2023).

A adaptação viral a diferentes pressões seletivas, no entanto, afeta a amplificação do genoma viral por métodos de enriquecimento por *amplicon* para o sequenciamento genético. Exemplo é, devido à ausência de atividade revisora em sua RNA polimerase (Guzman *et al.*, 2010; Leitmeyer *et al.*, 1999), e graças a baixa estabilidade do RNA quando comparado ao DNA, a ocorrência de mutações que, eventualmente, se fixam. Logicamente, quanto mais casos, mais oportunidades o vírus dispõe para acumular mutações e se diferenciar geneticamente no espaço e no tempo. Por outro lado, a dispersão dessas variantes, por dependerem de um vetor, possui menor alcance, por exemplo, quando comparada a outros patógenos com transmissão pelo ar. Essas características somadas têm o potencial de gerar genótipos região-específicos.

A diversidade viral subexplorada, principalmente em países com pouca ou ausente histórico de sequenciamento genômico e de vigilância sindrômica, é hipótese passível de explicar a baixa sensibilidade que é apresentada por kits diagnósticos sorológicos (24 – 86%) (Alidjinou *et al.*, 2022; Aryati *et al.*, 2013; Guzman *et al.*, 2010; Hermann *et al.*, 2014; Lima *et al.*, 2010).

A detecção do vírus por meios confiáveis favorece o manejo adequado dos pacientes, bem como a vigilância epidemiológica, a qual monitora a emergência e o ressurgimento de sorotipos ou de genótipos específicos (Ayukepi *et al.*, 2017; Muller; Depelsenaire; Young, 2017). A complexidade da doença é ampliada pela diversidade genética dos sorotipos do vírus da dengue (DENV), especialmente dos sorotipos I e II, o que torna essencial a identificação precisa das variantes em circulação para a formulação de estratégias de controle eficazes (Phadungsombat; Nakayama; Shioda, 2024). Nesse contexto, a introdução de painéis de sequenciamento genômico padronizados surge como uma solução promissora para melhorar a detecção e monitoramento das cepas virais.

Este estudo investiga a implementação de painéis de sequenciamento genômico dos sorotipos I e II do DENV no Brasil, visando a desenvolver métodos para acompanhar o

desempenho dos *primers* ao longo do tempo. A proposta se baseia na integração de informações epidemiológicas, sociodemográficas, ambientais e econômicas, que são cruciais para a previsão de picos de casos e para a antecipação da necessidade de atualização dos painéis de sequenciamento.

2. REVISÃO BIBLIOGRÁFICA

2.1.Transmissão

Os arbovírus, como os responsáveis por doenças como dengue, zika e chikungunya, dependem de vetores artrópodes para completar seu ciclo de vida e garantir sua transmissão. Entre esses vetores, o *Aedes aegypti* se destaca como o principal transmissor em áreas urbanas devido à sua ampla distribuição e alta capacidade de adaptação às condições humanas (Bhatt *et al.*, 2013).

Ainda, estudos sugerem um comportamento complexo entre os vetores e os vírus carregados: a infecção viral pode alterar o comportamento do vetor, como o aumento da frequência de picadas em humanos, o que amplifica a eficiência da transmissão do vírus (Xiang *et al.*, 2021). Essa dinâmica reforça a cadeia humano-mosquito-humano, permitindo que a transmissão ocorra mesmo em indivíduos assintomáticos, um fator crítico na disseminação de arbovírus (Duong *et al.*, 2015).

Contudo, a dependência de vetores específicos impõe limitações geográficas e ambientais à propagação dos arbovírus. A densidade populacional do mosquito vetor é um fator chave, influenciando diretamente o risco de surtos. Isso significa que a transmissão está fortemente associada a ambientes e períodos que favoreçam a reprodução e sobrevivência dos mosquitos, como climas quentes e úmidos que ampliam seu nicho ecológico (Simmons *et al.*, 2012).

Ao contrário de vírus de transmissão direta entre humanos, como o SARS-CoV-2, os arbovírus dependem do espaço geográfico e da interação com condições ecológicas específicas para sua disseminação. Essa dependência ecológica favorece a circulação de genótipos adaptados a regiões específicas, tornando o controle ainda mais desafiador, especialmente em áreas onde a presença do vetor é endêmica e sua proliferação é difícil de controlar.

2.2. Dispersão geográfica e genética dos sorotipos e genótipos de DENV

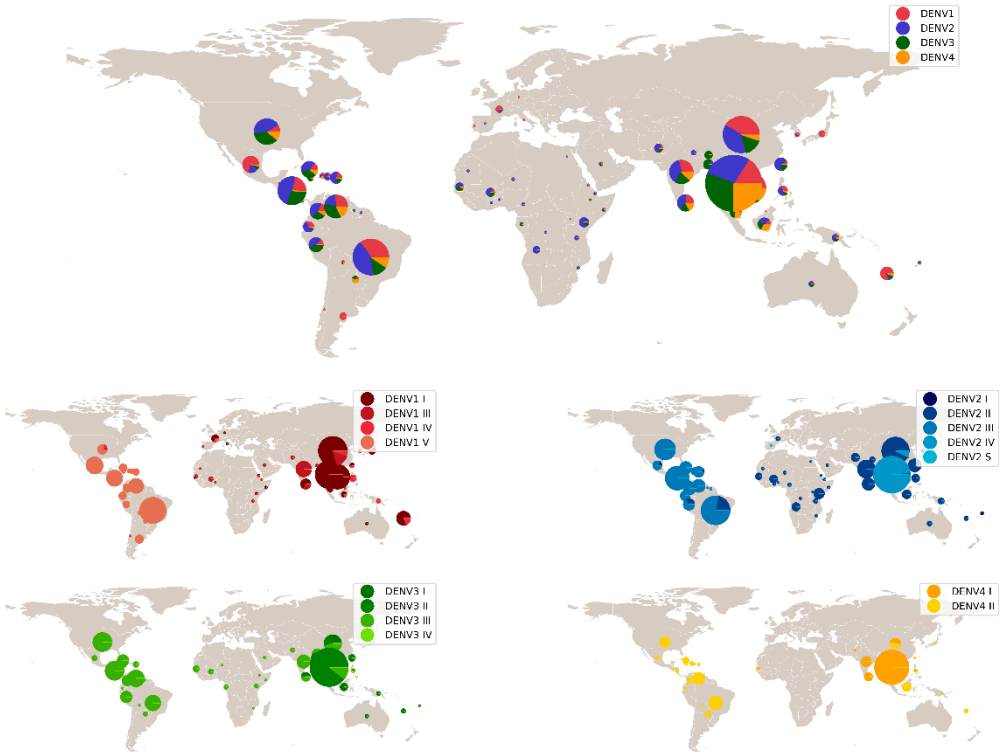
A dispersão geográfica do vírus da Dengue tem sido amplamente influenciada por uma série de fatores que têm expandido os seus limites de endemicidade. Estes incluem, mas não se limitam, a mudanças no ambiente ecológico que favorecem o aumento da população de mosquitos vetores, tais como a urbanização desenfreada, os efeitos do aquecimento global que permitem a adaptação dos mosquitos a novas áreas geográficas e o aumento das viagens

internacionais que facilitam a dispersão do vírus para regiões previamente não afetadas. Esses fenômenos convergentes têm contribuído significativamente para a expansão geográfica do DENV, criando desafios para o controle e prevenção da doença (Hopp; Foley, 2001; Lambrechts; Scott; Gubler, 2010; Mousson *et al.*, 2005).

A sua expansão pode ser evidenciada, por exemplo, pela emergência do genótipo II do sorotipo DENV 2, conhecido como "cosmopolita", o qual exibe uma notável dispersão geográfica (Twiddy *et al.*, 2002). Este genótipo, caracterizado por sua adaptabilidade e prevalência em diversos ambientes, foi identificado mais recentemente no Brasil (Giovanetti *et al.*, 2022a), embora tenha suas origens na Ásia (Yenamandra *et al.*, 2021).

No entanto, é fundamental reconhecer que o comportamento singular desse genótipo específico não representa a norma. A capacidade de atravessar fronteiras geográficas entre países é uma característica excepcional entre os genótipos do vírus da Dengue. De maneira geral, observa-se uma notável diferenciação na prevalência desses genótipos em diversas áreas geográficas, evidenciando a complexidade das interações entre fatores ambientais, comportamentais e genéticos que moldam a disseminação do vírus. (Allicock *et al.*, 2012; Holmes; Twiddy, 2003).

Figura 01 - Dispersão geográfica do Vírus Dengue (DENV) e seus distintos sorotipos. Cada sorotipo é subdividido posteriormente em genótipos, baseando-se na sequência genética da proteína Envelope.

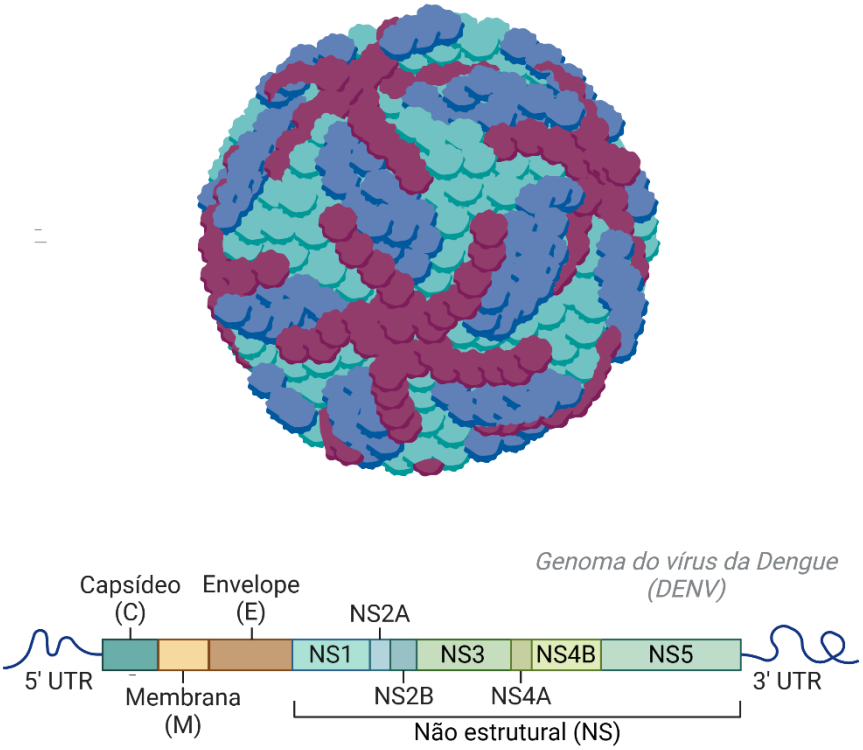


Fonte: Elaborado pelo autor com dados disponíveis em https://nextstrain.org/dengue/all/genome.

2.3. Estrutura molecular viral

É um vírus de RNA de fita única e sentido positivo de aproximadamente 10 mil bases e possui uma única poliproteína que codifica 3 proteínas estruturais Capsídeo (C), Membrana (M) e Envelope (E), e 7 não estruturais (NS1, NS2a, NS2B, NS3, NS4A, NS4B, NS5) (Guzman *et al.*, 2010; Leitmeyer *et al.*, 1999).

Figura 02 - Genoma e estrutura superficial do vírus Dengue. Seu genoma é dividido em genes codificantes para proteínas estruturais (formadores da partícula viral) e não estruturais (proteínas que interagem com o sistema imune e participam do ciclo de vida do vírus)



Fonte: Adaptado pelo autor a partir de um modelo pré-disponível em https://www.biorender.com/.

2.3.1. Proteína não estrutural 1 (NS1)

A proteína NS1 (46 kDa) pode ser encontrada na forma de dímero, quando associada à membrana celular e outros componentes celulares (Brown *et al.*, 2016) e como hexâmero associado a lipídeos (Gutsche *et al.*, 2011), sendo esta última a forma encontrada secretada e utilizada para o diagnóstico por testes diretos de detecção de antígeno (Alcon *et al.*, 2002).

2.3.2. Envelope (E)

A proteína E, envelope, é uma glicoproteína extensa, contendo 495 resíduos de aminoácidos, com uma abundância de cisteína, onde 12 resíduos altamente conservados formam seis pontes dissulfeto, e exibindo múltiplas funções. (Fahimi *et al.*, 2018). Pesquisas utilizando a técnica de cristalografia revelaram que a proteína E do vírus da dengue é constituída por três domínios claramente diferenciados em sua estrutura: um domínio central (EDI), um

domínio de dimerização (EDII) e um domínio terminal C que se assemelha a uma imunoglobulina (Ig) (EDIII) (Modis *et al.*, 2004, 2005).

2.3.3. Domínios imunogênicos das proteínas NS1 e Envelope

Os domínios imunogênicos das proteínas NS1 e envelope são de particular interesse devido ao seu potencial como alvos terapêuticos e de diagnóstico. Para a NS1, estudos destacaram os epítopos localizados nas regiões 01-24, 25-38, 21-40, 36-45, 50-65, 57-77, 74-88, 80-89, 103-112, 108-129, 101-130, 121-130, 127-143, 156-175, 166-178, 187-196, 193-209, 206-215, 233-245, 231-255, 264-281, 275-290, 295-304, 315-324, 296-325, 311-330 e 341-351 (Freire *et al.*, 2017; Hertz *et al.*, 2017; Jiang *et al.*, 2010; Nagar; Savargaonkar; Anvikar, 2020; Rocha *et al.*, 2017). Por sua vez, na proteína de envelope, os epítopos relevantes foram identificados nas regiões 02-25, 67-80, 98-109, 201-226, 310-330, 390-409 e 411-428 (Bergamaschi *et al.*, 2019; Nagar; Savargaonkar; Anvikar, 2020). Esses achados ressaltam a complexidade das interações entre o vírus e o sistema imunológico, além de fornecerem subsídios cruciais para o desenvolvimento de estratégias de prevenção e tratamento eficazes contra a dengue e o zika.

2.4. Sequenciamento genômico

2.4.1. Sanger

A técnica de sequenciamento de DNA concebida por Frederick Sanger em 1977, que incorporava o uso de dideoxinucleotídeos, desempenhou um papel crucial ao longo de vários anos na obtenção do primeiro esboço do genoma humano (Hutchison, 2007). Essa abordagem envolve a realização de análises de PCR utilizando tanto deoxinucleotídeos comuns quanto dideoxinucleotídeos, estes últimos atuando como bloqueadores da reação. Esse método gera fragmentos de DNA com diferentes comprimentos, passíveis de serem separados por eletroforese e posteriormente identificados (Metzker, 2005).

No cenário atual, a tecnologia empregada para a execução do sequenciamento Sanger evoluiu consideravelmente em relação à sua implementação original. Técnicas modernas e avançadas, como a eletroforese capilar e a detecção de fluoróforos, são agora empregadas (ThermoFisher Scientific, 2016).

2.4.2. NGS - Next-Generation Sequencing

Um dos progressos mais significativos na tecnologia de *Next-Generation Sequencing* (NGS), em comparação com o sequenciamento Sanger, reside na capacidade de gerar grandes volumes de dados a um custo relativamente baixo, viabilizando o sequenciamento de diversos organismos e facilitando análises evolutivas (Metzker, 2010). Em determinadas situações, devido à escassez de amostras disponíveis para sequenciamento, algumas análises demandam etapas de amplificação da amostra inicial (Chang; Li, 2013). Embora essa amplificação seja frequentemente necessária, sua suscetibilidade a erros e vieses, resultantes do uso de iniciadores específicos, destaca a importância do controle de qualidade antes, durante e após a montagem dos genomas (Chen *et al.*, 2018).

De maneira geral, no processo de construção de bibliotecas genômicas, a amostra é fragmentada, seja por enzimas ou métodos físicos, gerando fragmentos menores (Bronner; Quail, 2019). Esses fragmentos de DNA são então ligados a adaptadores, que facilitam a hibridização de *primers*, permitindo a amplificação da região desejada. Uma prática comum em várias tecnologias de NGS é a fixação da biblioteca genômica em uma

matriz sólida, permitindo o sequenciamento simultâneo de múltiplas amostras (Metzker, 2010).

Na tecnologia *Illumina*, que adota a estratégia de terminação cíclica reversível, a reação é interrompida após a adição de um nucleotídeo, possibilitando a clivagem do terminador e, em seguida, a continuação da reação em cadeia (Metzker, 2010). Esse método permite ao sequenciador identificar, por fluorescência, a base que foi adicionada, sendo a identificação de um nucleotídeo específico derivada do consenso da informação transmitida pelo cluster em um determinado ciclo de amplificação (Bronner; Quail, 2019).

Independentemente da tecnologia utilizada, o resultado do sequenciamento são os arquivos FastQ, que representam as sequências identificadas durante o processo, juntamente com parâmetros de qualidade essenciais para algoritmos durante a montagem do genoma (Cock *et al.*, 2009). A montagem do genoma, por sua vez, pode ser conduzida tanto por referência quanto de novo. A abordagem por referência envolve o mapeamento das leituras sequenciadas do arquivo FastQ contra uma sequência padrão (Behjati; Tarpey, 2013). Embora apresente limitações, essa metodologia é valiosa quando o objetivo da pesquisa é a comparação com um genoma de referência e a anotação de pequenas variações (Metzker, 2010).

2.4.3. Enriquecimento de material genético

2.4.3.1. Hibridização

A técnica de enriquecimento por hibridização utiliza um conjunto de sondas biotiniladas ligadas a *beads* magnéticas, pequenas esferas paramagnéticas que permitem a captura e separação do material genético alvo. Essa abordagem, devido ao maior tamanho da sonda, é capaz de sustentar a hibridização com material genético mesmo frente a um dado grau polimorfismos. Por outro lado, o aumento da sensibilidade acompanha uma redução na especificidade, onde sondas específicas para SARS-CoV-2 podem, por exemplo, capturam consideravelmente material genético do hospedeiro não especificamente e comprometendo a quantidade de *reads* disponíveis para os reais alvos (Von Sydow *et al.*, 2023).

2.4.3.2. Amplicon

O enriquecimento por *amplicon*, por outro lado, é útil quando não se é esperado muita variabilidade dentro do próprio organismo, uma vez que mutações individuais tem o potencial de causar quedas de cobertura na respectiva região flanqueada pelos *primers* (Arana *et al.*, 2022). Esse conhecimento a priori reflete a necessidade de se conhecer bem o organismo estudado e sua variabilidade genética. Esta técnica permite o enriquecimento ao amplificar toda a extensão do genoma ou região de estudo, sendo altamente específica e sensível (Chiara *et al.*, 2021) e, quando comparada com hibridização, revela simplicidade e custo-efetividade (Von Sydow *et al.*, 2023). O desenho pode ser feito facilmente por softwares e sites, facilitando o processo (Quick *et al.*, 2017a).

Ao mesmo tempo, devido a persistência dos *amplicons* no ambiente de trabalho, um cuidado extensivo com contaminação deve ser tomado (Quick *et al.*, 2017a). Mutações que ocorrem em regiões de hibridização dos *primers*, mas que não são suficientes para o impedimento do alongamento da cadeia de nucleotídeos durante a replicação, podem ocasionar falsas mutações de reversão nos genomas estudados (Sanderson; Barrett, 2021).

O tamanho do *amplicon* é algo que deve ser considerado, pois um tamanho de *amplicon* maior significa num menor número de *primers* necessários para cobrir a totalidade da região alvo. *Amplicons* grandes refletem menos problemas com *mismatches* em relação à sequência de referência, mas encontra problemas ao trabalhar com materiais degradados ou com baixa concentração de material genético (Quick *et al.*, 2017a; Su *et al.*, 2022).

2.5.Formatos de arquivo de bioinformática

2.5.1. Fasta

O formato de arquivo FASTA foi introduzido concomitantemente com o software homônimo (Pearson; Lipmant, 1988). Reconhecido como um padrão na bioinformática, sua adoção ampla se justifica pela sua simplicidade, que simplifica a manipulação e análise de sequências utilizando ferramentas de processamento de texto. Uma entrada em formato FASTA inicia-se com uma linha de descrição, seguida por linhas contendo os

dados da sequência. A linha de descrição é diferenciada dos dados da sequência pelo símbolo de maior-que (">") na primeira coluna.

2.5.2. BED

O formato de arquivo BED (*Browser Extensible Data*) é uma estrutura de texto utilizada para armazenar informações sobre regiões genômicas, como coordenadas e suas respectivas anotações (Kent *et al.*, 2002). Este formato foi concebido durante o desenvolvimento do Projeto Genoma Humano (Kent *et al.*, 2002).

Um arquivo BED é composto, no mínimo, por três colunas, podendo opcionalmente incluir até nove colunas adicionais, totalizando doze colunas. As primeiras três colunas contêm os identificadores dos cromossomos, segmentos de *scaffolds* ou, em casos de vírus não segmentados, a referência do genoma, além do início e do fim das coordenadas das sequências em consideração. As nove colunas adicionais subsequentes podem conter anotações relacionadas a essas sequências. A separação entre as colunas deve ser feita por espaços ou tabulações, sendo esta última preferível por motivos de compatibilidade entre diferentes programas de análise.

2.5.3. SAM/BAM

O formato SAM é crucial na bioinformática, armazenando dados de alinhamento genômico para análises como identificação de variantes e expressão gênica.

O formato SAM compreende uma seção de cabeçalho e outra de alinhamento. As linhas no cabeçalho iniciam com o caractere '@', enquanto as linhas na seção de alinhamento não apresentam tal característica (Li *et al.*, 2009).

Cada linha de alinhamento no formato SAM possui 11 campos obrigatórios e um número variável de campos opcionais. Esses campos obrigatórios são brevemente descritos na tabela 1:

Tabela 01 - Descrição dos campos presentes no arquivo SAM: Esta tabela apresenta os campos que compõem um arquivo SAM (*Sequence Alignment/Map*), um formato amplamente utilizado para armazenar dados de alinhamento de sequências em bioinformática.

Nome	**Descrição**
QNAME	Nome da consulta do *read* ou do par de *reads*
FLAG	FLAG em bits (emparelhamento, fita, fita do par, etc.)
RNAME	Nome da sequência de referência
POS	Posição mais à esquerda baseada em 1 do alinhamento truncado
MAPQ	Qualidade de mapeamento (escalada em Phred)
CIGAR	*String* CIGAR estendida (operações: MIDNSHP)
MRNM	Nome do parceiro de referência (‘=’ se o mesmo que RNAME)
MPOS	Posição mais à esquerda baseada em 1 do parceiro
ISIZE	Tamanho do *insert* inferido
SEQ	Sequência da consulta na mesma fita que a referência
QUAL	Qualidade da consulta (ASCII-33=Qualidade da base Phred)

Fonte: Elaborado pelo autor

Adicionalmente, visando melhorar a eficiência, foi concebido um formato binário complementar denominado Formato Binário de Alinhamento/Mapa (BAM). O BAM representa em formato binário as mesmas informações contidas no SAM, mantendo sua integridade. A compressão do BAM é realizada pela biblioteca BGZF, uma ferramenta genérica desenvolvida para permitir acesso aleatório rápido em arquivos comprimidos compatíveis com zlib (Li *et al.*, 2009).

2.6. Bancos de dados

Os bancos de dados desempenham um papel crucial no armazenamento e gerenciamento de grandes volumes de informações em diversas áreas, como saúde e bioinformática. Esses sistemas permitem a organização e recuperação eficiente de dados, sendo utilizados amplamente em análises científicas. Existem diferentes tipos de bancos de dados, incluindo relacionais e não relacionais, cada um adequado a diferentes aplicações e necessidades (Tracz; Plechawska-Wójcik, 2024). Bancos de dados relacionais, como MySQL e PostgreSQL, utilizam tabelas e são baseados em SQL, enquanto os não relacionais, como MongoDB, lidam com dados de maneira mais flexível, sendo ideais para dados não estruturados (Sliusarenko; Filatov, 2023). A escolha do tipo de banco depende da natureza e do volume de dados, bem como das necessidades específicas de consulta e análise (Tracz; Plechawska-Wójcik, 2024).

2.7. Aprendizado de máquina

O aprendizado de máquina consiste no uso de algoritmos que reconhecem padrões em dados de forma automatizada, possibilitando prever resultados ou apoiar decisões em contextos incertos. Este conjunto de técnicas é utilizado em grandes conjuntos de dados, como registros eletrônicos de saúde, para detectar padrões e conexões que poderiam passar despercebidos em análises tradicionais. Essa abordagem é especialmente valiosa na epidemiologia da saúde, ajudando a compreender a disseminação de doenças e seus desfechos (Leung *et al.*, 2023; Sylvestre *et al.*, 2022; Wiens; Shenoy, 2018). Ele engloba diferentes métodos, como o aprendizado supervisionado, aplicado em tarefas de previsão e categorização, e o aprendizado não supervisionado, voltado para a identificação de padrões ocultos (Kreatsoulas; Subramanian, 2018). A regressão em *machine learning* é uma técnica utilizada para modelar relações entre variáveis, onde o objetivo principal é prever ou estimar uma variável dependente de natureza numérica (Loh, 2011). Já, a classificação, é uma técnica voltada para prever categorias ou rótulos de uma variável dependente de natureza qualitativa (Sen; Hajra; Ghosh, 2020).

Dentro desse campo, algumas métricas básicas são amplamente utilizadas para avaliar a performance de modelos preditivos de classificação. A *precision* mede a proporção de predições positivas corretas em relação ao total de predições positivas realizadas pelo modelo, refletindo sua capacidade de evitar falsos positivos (Powers; Ailab, 2011). Já o *recall* avalia a proporção de verdadeiros positivos identificados em relação ao total de exemplos positivos reais, destacando a habilidade do modelo de evitar falsos negativos (Powers; Ailab, 2011).

O *f1-score* é uma métrica que combina *precision* e *recall* em uma única medida harmônica, sendo útil para cenários com classes desbalanceadas, ao equilibrar a relação entre essas duas métricas (Powers; Ailab, 2011). Essa métrica é particularmente relevante quando há necessidade de balancear a importância de evitar falsos positivos e falsos negativos em um contexto específico, como na detecção de doenças.

Durante o processo de modelagem, é fundamental dividir o conjunto de dados em três partes: treino, teste e *out-of-time* (OOT). O conjunto de treino é utilizado para ajustar os parâmetros do modelo e ensiná-lo a identificar padrões nos dados (Medar; Rajpurohit, 2017). O conjunto de teste, por sua vez, é reservado para avaliar o desempenho do modelo em dados não vistos durante o treino, fornecendo uma estimativa mais realista de sua generalização (Medar; Rajpurohit, 2017). A técnica OOT vai além e consiste em separar

uma porção de dados referente a um período posterior ao usado no treino e teste, testando a capacidade do modelo de prever eventos futuros com base no conhecimento do passado (Maldonado; López; Iturriaga, 2022). Esta prática é especialmente importante em contextos em que o componente temporal é relevante, como em predições epidemiológicas.

Além disso, durante o processo de modelagem, é importante estar atento ao fenômeno de *overfitting*. Ocorre *overfitting* quando o modelo se ajusta excessivamente aos dados de treinamento, perdendo a capacidade de generalizar para novos dados, o que pode levar a um desempenho inferior em conjuntos de teste. Por outro lado, o *underfitting* acontece quando o modelo é excessivamente simples e não consegue capturar a complexidade dos dados, resultando em baixa performance tanto nos dados de treinamento quanto nos de teste (Kingsmore *et al.*, 2021).

A escolha de um modelo adequado, a correta separação dos dados e a avaliação de métricas são essenciais para garantir que o modelo seja eficiente, evitando problemas como *overfitting* e *underfitting*, que podem comprometer a sua aplicabilidade em cenários reais (Leung *et al.*, 2023; Naeem *et al.*, 2021).

3. OBJETIVO GERAL

Introduzir painéis de sequenciamento genômico dos sorotipos I e II do DENV no Brasil e desenvolver métodos para acompanhar o desempenho dos painéis de *primers* ao longo do tempo, integrando com informações epidemiológicas, sociodemográficas, ambientais e econômicas, para possibilitar a previsão de picos de casos e antecipar a necessidade de atualização dos painéis de sequenciamento.

3.1. Objetivos específicos

1. Analisar o impacto da diversidade genética do vírus da dengue no Brasil, levando em consideração fatores regionais e temporais, para contextualizar sua influência nas estratégias de vigilância e controle;
2. Desenvolver e validar um painel de *primers* padronizado para o sequenciamento genômico dos sorotipos I e II de DENV, levando em consideração a variabilidade genética observada no Brasil;

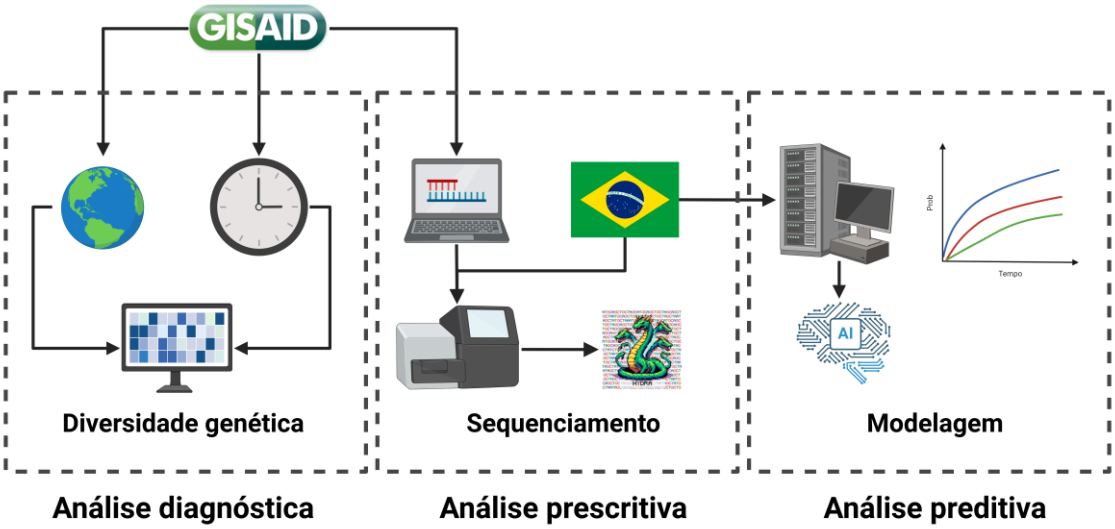
3. Desenvolver um software de monitoramento para avaliar o desempenho dos *primers*, sugerindo ajustes conforme surgem mutações nas regiões-alvo dos sorotipos;
4. Construir uma base de dados modular com dados públicos do Brasil que subsidie investigações epidemiológicas e integre informações sociais, econômicas e ambientais;
5. Desenvolver um modelo preditivo para os picos de casos de dengue, relacionando o aumento no número de casos ao surgimento de mutações e à consequente necessidade de atualização dos painéis de sequenciamento, utilizando variáveis epidemiológicas, demográficas e ambientais.

4. METODOLOGIA

O estudo, cujo fluxo é ilustrado pela figura 3, iniciou com a avaliação genética do vírus da dengue ao longo do tempo, com foco em dados disponibilizados pelo GISAID. A partir dessa análise, foram desenvolvidos *primers* específicos, e elaborou-se um modelo matemático para prever o tempo de vida útil desses *primers*, permitindo acompanhar sua eficácia ao longo do tempo. Além disso, foi desenvolvido um software para avaliação de sequências genéticas e identificação de *primers* defeituosos, otimizando sua utilização. Em etapa subsequente, foi criada uma base de dados integrada, utilizando informações públicas de saúde, ambientais e socioeconômicas. Por fim, foram aplicados algoritmos de *machine learning* e o Teorema de Bayes para a predição de casos de dengue, com técnicas de ajuste sazonal, geração de *Lags* e probabilidades atualizadas para classes menos frequentes.

Figura 03 - Fluxo geral da metodologia desenvolvida na dissertação, abrangendo desde a obtenção e o tratamento de dados de fontes públicas até a modelagem preditiva de casos de dengue. O diagrama ilustra as principais etapas: coleta de dados brutos, integração e padronização em um banco de dados PostgreSQL, análise e classificação de *primers*, geração de variáveis derivadas, modelagem preditiva com algoritmos avançados e disponibilização dos resultados por meio de uma aplicação web. O fluxo destaca a interligação das bases de dados, o uso de técnicas de

otimização computacional e os mecanismos de segurança implementados para o armazenamento e acesso aos dados.



Fonte: Elaborado pelo autor.

4.1.Avaliação contextualizada da variabilidade genética de DENV

4.1.1. Análise espacial

4.1.1.1. Obtenção das sequências

Para esta análise, foram recuperadas todas as sequências de DENV1-V (N = 4249), DENV2-II (cosmopolita, N = 3088) e DENV2-III (*asian-american*, N = 1854) com mais de 9000 bp disponíveis no GISAID EpiArbo até o dia 25/05/2024. Sequências com regiões classificadas como "*unknown*" foram removidas. Utilizou-se o Nextclade v.3.3.1 customizado para DENV, disponível em https://github.com/nextstrain/dengue/tree/main/nextclade/datasets, e *heatmaps* foram gerados para as proteínas NS1 e E.

4.1.1.2. Significância estatística

Para comparar a frequência entre mutações nas amostras provenientes do estado do Ceará, os testes chi-quadrado e teste exato de Fisher foram empregados, quando

apropriados, comparando-se a frequência de mutação entre as amostras do Ceará e a frequência das amostras das outras regiões investigadas. A correção de Bonferroni foi empregada para correção do valor de p para comparações múltiplas. Mutações com diferença significativa a nível significância (α) de 5% são assumidas, neste trabalho, como mutações de interesse.

4.1.2. Análise temporal

4.1.2.1. Obtenção das sequências para avaliar a diferença genética

No dia 20/06/2024, foram baixados todos os genomas de DENV1 e DENV2 disponíveis desde 1990, utilizando os filtros "*low cov excluded*: excluir entradas com mais de 25% de Ns" e "possuindo ao menos as regiões E e NS1". O número total de genomas baixados foi de 5021 para DENV1 e 5376 para DENV2, sendo, para DENV2, 3327 pertenciam ao genótipo II e 2049 ao genótipo III.

4.1.2.2. Alinhamento das sequências

O alinhamento das sequências foi realizado utilizando o Nextclade v3.3.1. O *dataset* necessário para a anotação e para o alinhamento dos dados foi obtido no repositório oficial do grupo Nextstrain: https://github.com/nextstrain/dengue/tree/main/nextclade/datasets.

Com o *dataset* específico para o sorotipo 2 do dengue, alterou-se o valor base do argumento "*min-seed-cover*" para permitir o alinhamento, juntamente do sorotipo 2, do sorotipo 1. O comando utilizado foi:

nextclade run --in-order --input-ref=./data/reference.fasta --input-annotation=./data/genome_annotation.gff3 --input-tree=./data/tree.json --input-pathogen-json=./data/virus_properties.json --output-fasta=output/nextclade.aligned.fasta --output-tsv=output/nextclade.tsv --min-seed-cover=0.05 [input].fasta.

4.1.2.3. Identificação e processamento de sequências-alvo

Rico-Hesse' (1990), em seu trabalho, utilizou uma única região específica que compreende as regiões codificantes das proteínas E e NS1 de DENV para a classificação

entre genótipos e sorotipos, indicando uma distância genética de 30% para separar sorotipos de DENV e de 12% para separar genótipos dentro de um sorotipo. No trabalho, a região de interesse foi identificada utilizando os *primers* D2/2452 (CCACATTTCAGTTCTTT) e D2/2578 (TTACTGAGCGGATTCCACAGATGC). Essa mesma região foi recortada nas sequências obtidas nos passos anteriores e uma matriz de distância entre sorotipos e genótipos foi construída para se avaliar a validade destes pressupostos atualmente.

As sequências que não possuíam a região de interesse completa ou que apresentavam Ns na região foram removidas. Em seguida, a matriz de distância de nucleotídeos entre as regiões dos *primers* foi calculada utilizando o software Geneious Prime 2021.2.2 (https://www.geneious.com).

4.1.2.4. Investigação de outliers para a distância genética de nucleotídeos

Para as amostras cuja diferença entre os genótipos II e III do sorotipo 2 de DENV foi próxima à distância entre os sorotipos 1 e 2, etapas adicionais foram realizadas. Foi calculada as divergências de cada gene dos *outliers* quando comparado com as referências adotadas pelo banco de dados EpiArbo do *Gisaid*: NC_001477.1 (DENV1) e NC_001474.2 (DENV2). A matriz de distâncias foi, então, utilizada para se realizar uma correlação linear de Spearman pareando as amostras por ano e as separando por genótipo, comparando as combinações dos genótipos DENV1-V, DENV2-II e DENV2-III.

4.2.Desenho de *primers* para sequenciamento genômico

4.2.1. Obtenção das sequências genômicas

No dia 13/10/2023 a plataforma Gisaid EpiArbo foi acessada e buscou-se individualmente por dengue 1, dengue 2. Todos os vírus foram filtrados para conter apenas amostras do Brasil, com data de coleta completa e superior a 01/01/2018 e para conter mais do que 9000 nucleotídeos. Em seguida, as amostras de DENV1 e de DENV2 foram analisadas na plataforma DengueSurver disponíveis no Gisaid.

4.2.2. Controle de qualidade e amostragem

Como controle de qualidade, foram removidas as sequências que possuíam mais que 10% de bases ambíguas (N) ou ausência da informação de estado de coleta. Posteriormente, devido ao elevado número de amostras encontradas, uma etapa de amostragem se fez necessária. Esta etapa se deu ao se escolher aleatoriamente 3 amostras por mês, estado de coleta, sorotipo e genótipo. Em casos em que havia menos do que 3 amostras, todas foram escolhidas.

4.2.3. Alinhamento múltiplo de sequências

Inicialmente, as sequencias foram separadas em grupos (DENV1-V N = 250, DENV2-II-III N = 67 + 132 = 198). Após, o alinhamento das sequências foi executado, para cada grupo/genótipo, usando o seguinte comando com o software Mafft v.7.515 (Katoh; Standley, 2013):

mafft --thread -1 --globalpair --maxiterate 1000 [input].fasta > aln.G-INS-i.[input].fasta

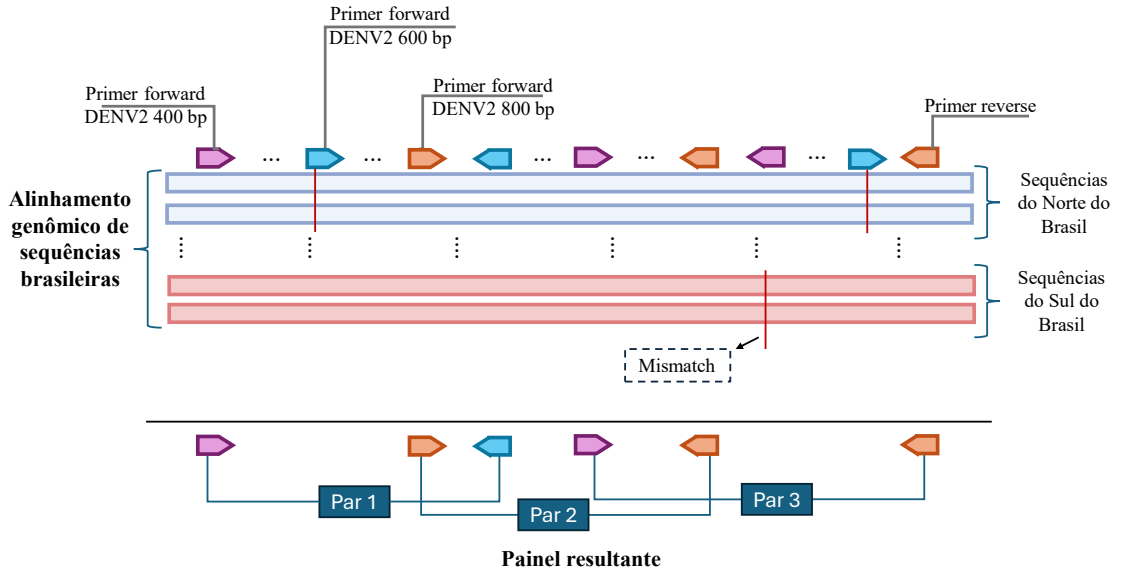
4.2.4. Construção dos painéis de sequenciamento genômico

Os alinhamentos foram visualizados com o *software* Geneious 2021.2.2 (https://www.geneious.com) para a geração de um genoma consenso com uma similaridade de 75%, resultando em 2 genomas consenso. Essa similaridade foi definida com base na soma das frequências relativas dos dois nucleotídeos mais abundantes em cada posição, que deve exceder X% para justificar a anotação de um nucleotídeo degenerado. O critério foi escolhido com o objetivo de minimizar a ocorrência de bases nitrogenadas degeneradas no genoma consenso. Os painéis de sequenciamento completo por *amplicon* foram feitos pela plataforma *online* Primal Scheme (Quick *et al.*, 2017b) para os tamanhos aproximado de *amplicon* de 400, 600 e 800 bases.

Para a construção do painel “Pan-DENV2”, abrangendo os genótipos II e III, utilizou-se o software Geneious para integrar *primers* derivados de diferentes painéis de *amplicons* (400, 600 e 800 pares de bases) gerados pelo *Primal Scheme*. Durante a curadoria manual, *primers* contendo qualquer *mismatch* foram descartados. Novos pares foram formados ao combinar *primers* existentes, priorizando a ausência de *mismatches*.

Por exemplo, caso o *primer* reverse original apresentasse incompatibilidade, ele era substituído pelo próximo *primer* reverse disponível sem *mismatch*, garantindo a integridade do painel O processo de construção do painel é ilustrado na figura 04.

Figura 04 - Processo de construção do painel "Pan-DENV2". Os *primers* gerados para painéis de 400, 600 e 800 pares de bases pelo *Primal Scheme* foram combinados no *software* Geneious. Durante a curadoria manual, *primers* com qualquer *mismatch* foram descartados, e novos pares foram formados ao substituir *primers* incompatíveis, assegurando a cobertura das sequências dos genótipos II e III do DENV2.



Fonte: Elaborado pelo autor.

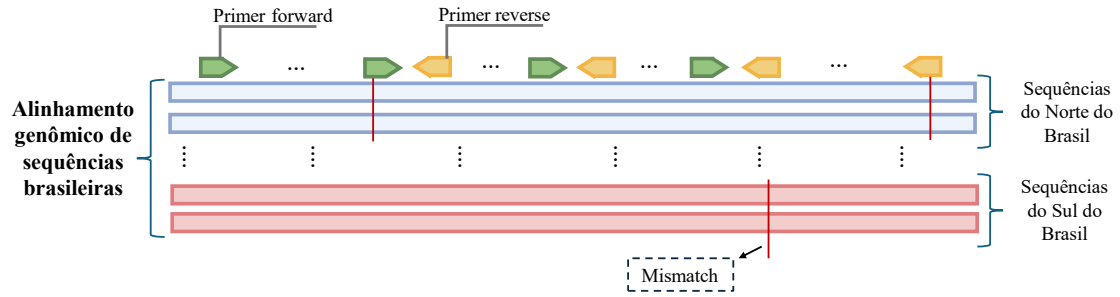
4.2.5. Validação in silico dos painéis de sequenciamento

Os *primers* foram submetidos a uma rotina customizada, usando a biblioteca *primer*3py da linguagem de programação *Python*, para avaliação da formação de estruturas secundárias de DNA: grampos, homodímeros e heterodímeros.

Além disso, os *primers* selecionados (DENV1 400bp, DENV1 600bp, DENV1 800bp, PanDENV2BR, DENV1 ref. e DENV2 ref. (Vogels *et al.*, 2023)) foram mapeados contra as sequências genômicas brasileiras usadas no desenvolvimento dos painéis, permitindo verificar a quantidade de *mismatches* por sorotipo, conforme descrito pela figura 05.

Figura 05 - Esquematização do mapeamento e contagem de *mismatches* para os diversos painéis de *primers* utilizados neste trabalho (DENV1 400bp, DENV1 600bp, DENV1 800bp,

PanDENV2BR, DENV1 ref. e DENV2 ref.) contra as sequências brasileiras de diferentes regiões do Brasil pelo software Geneious Prime.



Fonte: Elaborado pelo autor.

4.3.Sequenciamento genômico

4.3.1. Obtenção de amostras

As amostras foram obtidas de duas localidades, abrangendo as regiões Norte e Sudeste do Brasil. No Amazonas, as amostras foram coletadas pela Fiocruz Amazonas – Laboratório de Ecologia de Doenças Transmissíveis na Amazônia (EDTA). No estado do Rio de Janeiro, as amostras foram obtidas da Fiocruz – Laboratório de Arbovírus e Vírus Hemorrágicos (LARBOH).

4.3.2. Desenho experimental

4.3.2.1. Normalização e Preparação dos Pools de Primers

A normalização dos *primers* foi realizada utilizando o equipamento *Dragonfly,* ajustando a concentração para 100 µM com TE (Tris-EDTA). Após a normalização, cada *primer* foi diluído em um pool de 200 µL de modo que a concentração final de cada *primer* no pool seja de 1 µM, também com TE como diluente, assegurando uniformidade nas concentrações.

Esse procedimento de normalização e preparação foi repetido para todos os conjuntos de *primers* testados, assegurando que todas as reações subsequentes de PCR utilizassem *primers* com concentrações consistentes, minimizando variações nas reações de amplificação.

4.3.2.2. Validação dos Primers para Sequenciamento Genômico

A validação dos *primers* foi realizada em duas etapas principais: validação experimental por PCR e eletroforese em gel, e o sequenciamento dos produtos amplificados. A primeira etapa consistiu na realização de reações de PCR utilizando amostras de cDNA viral previamente recebidas. Para DENV1, os conjuntos de *primers* com tamanhos de *amplicon* de 400, 600 e 800 bases foram validados com 12 amostras de cada região do Sudeste do Brasil e 3 do Norte do Brasil.

4.3.2.3. Sequenciamento dos Produtos de PCR

Os produtos amplificados por PCR foram sequenciados utilizando o sistema Miseq da Illumina. A etapa de sequenciamento foi realizada com um cartucho V2 300 micro, empregando 45 (12 – DENV1 (RJ) + 3 – DENV1 (AM)) tratamentos.

4.3.2.4. Miniaturização e Pareamento de Amostras

Para a etapa de comparação dos *primers*, foi utilizado um sistema de miniaturização com os equipamentos Mosquito/Dragonfly, que permitiu o pareamento eficiente das amostras com volumes reduzidos. Esta etapa foi fundamental para maximizar o número de amostras processadas simultaneamente e garantir a consistência na comparação dos resultados. Além dos *primers* com *amplicons* de 400, 600 e 800 bases para DENV1 e do painel Pan-DENV2, foram empregados os *primers* DENV1 e DENV2 desenvolvido por Vogels et al. (2023) para verificar a superioridade ou equivalência em relação aos *primers* projetados neste estudo.

Foram utilizadas amostras de DENV de duas regiões do Brasil: Amazonas (AM) e Rio de Janeiro (RJ), totalizando 66 amostras, sendo 12 DENV1 AM, 14 DENV2 AM, 27 DENV1 RJ e 13 DENV2 RJ. Essas amostras foram distribuídas entre os diferentes painéis de *primers* avaliados, totalizando 210 tratamentos, dos quais 156 correspondem a DENV1 e 54 a DENV2. O experimento foi planejado para comparar o desempenho dos painéis desenvolvidos com *primers* de referência internacional, utilizando um cartucho V3 no sequenciador MiSeq (Illumina), visando uma profundidade média de sequenciamento de 2000x.

Durante a validação das bibliotecas de sequenciamento com os robôs, observou-se uma discrepância na quantificação do DNA, além de um resultado abaixo do esperado. A análise no equipamento TapeStation indicou uma concentração de 0,5 ng/µL, enquanto as medições no Qubit, tanto com o kit *Broad Range* (BR) quanto com o *High Sensibility* (HS), resultaram em valores abaixo do limite de detecção. Diante desse resultado, decidiu-se não prosseguir com o sequenciamento para evitar o desperdício de um cartucho, optando por interromper a etapa final do experimento.

4.3.2.5. Equipamento Utilizado para o Sequenciamento

O sequenciamento genômico do vírus será realizado no Miseq Illumina, que possui capacidade adequada para o sequenciamento de produtos de PCR de diferentes tamanhos de *amplicon*. A escolha dos cartuchos V2 micro e V3 foi baseada na quantidade de amostras e na profundidade de sequenciamento necessária para cada etapa.

4.3.2.6. Montagem dos genomas

Os genomas foram montados por meio da ferramenta *ViralFlow* v1.0.0 (Silva *et al.*, 2024), utilizando como referência os genomas NC_001477.1 para DENV1 e NC_001474.2 para DENV2. De forma breve, o pipeline realiza o pré-processamento das leituras brutas aplicando filtros de qualidade com escore *Phred* mínimo de 20, além de remover automaticamente possíveis adaptadores remanescentes empregados durante o sequenciamento, com a ferramenta fastp v0.23.4 (Chen *et al.*, 2018). A remoção das regiões de *primers* é feita de eficientemente com o comando *ampliconclip* do *samtools* v1.11 (Li *et al.*, 2009), utilizando arquivos BED que indicam as posições dos *primers*.

As sequências que passam nos filtros são alinhadas ao genoma de referência usando o BWA v0.7.17 (Li, 2013). Por fim, os genomas consenso (fasta) são gerados com base no alelo mais frequente em cada posição, utilizando o *samtools* v1.11 em conjunto com o iVar v1.4.2 (Grubaugh *et al.*, 2019), estabelecendo uma profundidade mínima de 5 leituras para definição do consenso.

4.4. Aprimoramento da atualização de painéis de *primers*

4.4.1. Estimativa do tempo de vida útil de um painel de primers

Pode-se usar a função de probabilidade da distribuição binomial para encontrar a probabilidade (*p*) de encontrarmos pelo menos uma mutação em região de *primer* (*M = 1*) após *n* tentativas.

$$p = \frac{\sum_{i=1}^{N_{primers}} C_i}{G} \quad (1)$$

Onde *G* é o comprimento do genoma do vírus e C_i é o comprimento do *primer i*.

$$q = 1 - p \quad (2)$$

Onde *q* é a probabilidade de não ocorrer a mutação na região de *primer*.

$$P(M = 1 \mid N = n) = 1 - q^n \quad (3)$$

Pode-se considerar *n* como sendo o número de tentativas para um sucesso, em um ano, devido a unidade de μ, onde μ é a taxa de substituições por sítio por ano.

$$n = G \times \frac{\mu}{12} \times t \quad (4)$$

Substituindo *n* na equação (3), temos

$$P(M = 1 \mid T = t) = 1 - q^{G \times \frac{\mu}{12} \times t} \quad (5)$$

Onde *t* é o número de meses. Substituindo a equação 1 em 5, temos:

$$P(M = 1 \mid T = t) = 1 - \left(1 - \frac{\sum_{i=1}^{N_{primers}} C_i}{G}\right)^{G \times \frac{\mu}{12} \times t} \quad (6)$$

É possível, ainda, generalizar a equação 6 para um número de pelo menos *m* mutações em região de *primer* em um tempo *t* ao se aplicar a seguinte equação, caso *m-1* for menor ou igual a *n*:

$$P(M = m \mid T = t) = 1 - \sum_{k=0}^{m-1} \binom{n}{k} p^k (1-p)^{n-k} \quad (7)$$

Se $m - 1 > n$, então $P(M = m \mid T = t) = 0$ devido ao fato de que não se pode acontecer *m* mutações em regiões de *primers* sem que ocorram ao menos *m* mutações.

4.4.1.1. Cálculo das Variáveis

A porcentagem de cobertura do genoma viral pelas regiões-alvo dos *primers* foi calculada somando-se os comprimentos individuais de cada *primer* desenvolvido neste estudo. O comprimento do genoma viral foi baseado na sequência de referência NC_001474.2, com 10.902 nucleotídeos. A taxa de mutação por sítio por ano utilizada foi de $8{,}94 \times 10^{-4}$ (Wei; Li, 2017).

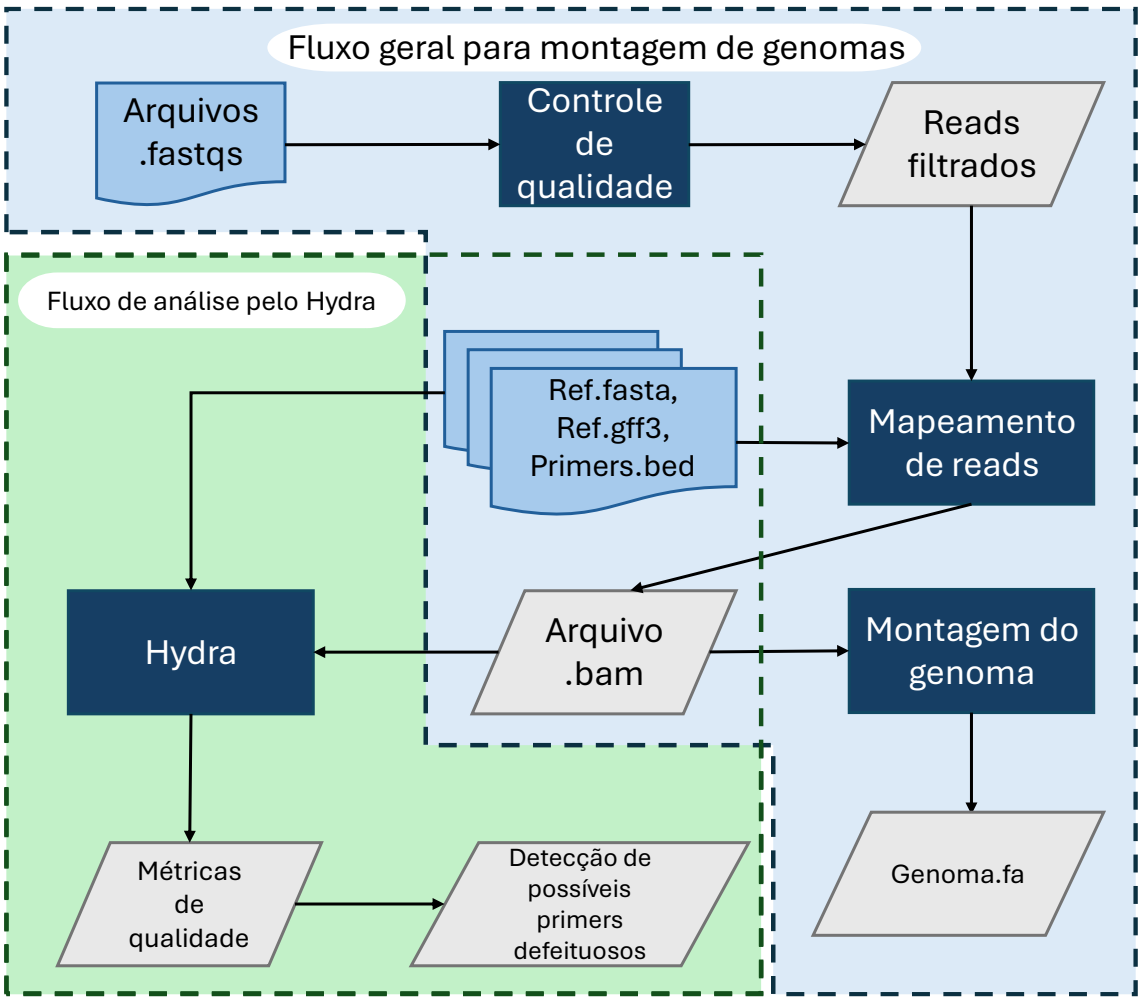
4.4.2. Hydra: software para monitoramento de painéis de amplicon

4.4.2.1. Construção e encapsulamento do software

O software para visualização de arquivos BAM, Hydra (*High-throughput Yield Data Read Analyzer*), foi desenvolvido utilizando uma combinação de ferramentas de bioinformática e bibliotecas em *Python*. Durante o desenvolvimento, foram utilizadas as seguintes ferramentas e versões: *bam-readcount* (versão 1.0.1-unstable-8-c7c76e6, commit c7c76e6), *samtools* (versão 1.19.2), minimap2 (versão 2.24-r1122), minia (versão 3.2.5), *mosdepth* (versão 0.3.7) e HyINDEL (disponível em: https://github.com/alok123t/HyINDEL). Para a interface de usuário e visualização dos dados genômicos, foram empregadas as bibliotecas *Streamlit*, para a criação de interfaces interativas e dinâmicas, e *Plotly*, para a geração de gráficos interativos. Os softwares foram encapsulados em uma solução *Docker* e disponibilizados via *Dockerhub*,

permitindo o compartilhamento de modo reprodutível. A figura 6 ilustra a inserção do *software* Hydra dentro de uma *pipeline* genérica para montagem de genomas virais.

Figura 06: Fluxograma do processo de análise de arquivos BAM no software Hydra. O fluxo principal utiliza dados de sequenciamento para identificar *primers* potencialmente problemáticos, como aqueles com *mismatches*, baseando-se na variabilidade intra-hospedeiro do vírus sequenciado.



Fonte: Elaborado pelo autor.

4.4.2.2. Cálculo de métricas

4.4.2.2.1. Métrica de variabilidade de cobertura vertical

$$Escore\ de\ qualidade = \frac{1}{(iiq) * (mcv)} \quad (8)$$

Onde iiq representa o intervalo interquartil da cobertura vertical, e mcv corresponde à mediana da cobertura vertical da região sequenciada. A escolha desses fatores visa minimizar o impacto de valores extremos (*outliers*). A inversão da multiplicação na fórmula tem como objetivo facilitar a interpretação, associando diretamente a métrica à qualidade do sequenciamento, de forma proporcional.

4.4.2.2.2. Classificação de *primers* afetados por mutações futuras

Para classificar um *primer* como potencialmente afetado por mutações futuras, foi definido um critério arbitrário de frequência mínima de 10%. Esse critério estabelece que um nucleotídeo de menor frequência deve apresentar uma frequência mínima de 10% e ser diferente da respectiva base que compõe o *primer*. Esse valor pode ser ajustado de acordo com a necessidade do usuário.

4.4.2.3. Testes de tempo e memória

Para as análises, utilizaram-se os *benchmarks* disponibilizados no repositório do CDC no GitHub, disponível em: https://github.com/CDCgov/datasets-sars-cov-2. O conjunto de dados (*Dataset 4 - VOI/VOC lineages*) contém 16 amostras com diferentes níveis de qualidade (boa, média e baixa), garantindo a reprodutibilidade dos resultados obtidos. A configuração do sistema computacional utilizado foi: Processador Intel Core i7-10700KF CPU @ 3.80GHz (8 núcleos, 16 *threads*), 32 GB de RAM (2x16 GB DDR4, 2666 MHz, Team Group Inc). Os arquivos BAM foram gerados utilizando a ferramenta ViralFlow 1.0.0 (Silva, Da *et al.*, 2024). A escolha do benchmark foi baseada na diversidade de qualidade das amostras. Para a avaliação dos *primers*, utilizaram-se os do protocolo ARTIC V5.3.2, que eram os mais atualizados no momento deste estudo. As bibliotecas *time* e *psutil* do *Python* foram empregadas para monitorar o tempo de

execução e o uso de recursos computacionais. Todas as análises foram realizadas em triplicata para assegurar a consistência dos resultados.

4.5. Arquitetura e modelagem de dados para vigilância epidemiológica

4.5.1. Escolha das bases de dados

Foram selecionadas bases de dados públicas como o Sistema de Informação de Agravos de Notificação (SINAN) e o Sistema de Informação sobre Mortalidade (SIM), com foco inicial na dengue, devido ao seu impacto recente na saúde pública no Brasil (Amorim *et al.*, 2023; Gräf *et al.*, 2023) e ao escopo do atual estudo. Além dessas, foram incluídas as bases do Cadastro Nacional de Estabelecimentos de Saúde (CNES), Instituto Brasileiro de Geografia e Estatística (IBGE), Instituto Nacional de Meteorologia (INMET) e *MapBiomas*.

A integração dos dados do SINAN e SIM permite uma análise detalhada de surtos e suas correlações com a mortalidade. O CNES fornece informações sobre a infraestrutura de saúde, essenciais para avaliar a capacidade de resposta a surtos. Combinando esses dados com informações populacionais e econômicas do IBGE, é possível investigar a influência de fatores socioeconômicos sobre a saúde pública. Adicionalmente, os dados climáticos do INMET e as informações sobre uso do solo do MapBiomas acrescentam uma dimensão ambiental, possibilitando a análise do impacto de variáveis climáticas e mudanças no uso do solo na disseminação de doenças.

4.5.2. Obtenção dos dados

Os dados das bases SINAN Dengue, SIM DO, CNES ST e IBGE foram obtidos por meio da aplicação de transferência de arquivos do DATASUS/MS (disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/), e sua conversão foi realizada utilizando a biblioteca *read.dbc* (https://github.com/danicat/*read*.dbc/). Para as bases do INMET (https://portal.inmet.gov.br/) e *MapBiomas* (https://brasil.mapbiomas.org/), a coleta foi realizada diretamente de seus respectivos sites. Após a coleta, os dados passaram por um processo sistemático de tratamento para assegurar a qualidade e integridade das informações. O armazenamento inicial ocorreu em uma camada de dados brutos, preservando-os conforme obtidos. Para a construção da camada tratada, *scripts*

desenvolvidos internamente foram utilizados para decodificar variáveis categóricas representadas numericamente. Os códigos desenvolvidos para a obtenção e tratamento dos dados estão disponíveis em: https://github.com/pdrmglc/ETL-ML-Dengue.

4.5.3. Tratamento dos dados

Para as bases do SINAN, SIM, IBGE e CNES foi aplicada apenas uma etapa de decodificação das variáveis categóricas, utilizando o dicionário de dados disponibilizado nas respectivas plataformas. Conflitos entre versões de dicionários de dados foram resolvidos com base em um dicionário unificado, disponível em https://github.com/pdrmglc/ETL-ML-Dengue, que foi desenvolvido para corrigir ambiguidades e atualizar a nomenclatura das diferentes classificações de gravidade da doença (WHO, 2009a). A coleta de dados do INMET dependeu das estações meteorológicas distribuídas pelos municípios, mas nem todos os municípios mantêm operações contínuas dessas estações ao longo do tempo, resultando em lacunas nos dados. Para tratar esses valores ausentes, foi aplicada a técnica de imputação, utilizando a mediana das variáveis correspondentes, calculada por estado, mês e ano. Já no caso dos dados do *MapBiomas*, a camada bruta apresentava os anos como colunas (formato *wide*). Para padronizar o formato com as demais bases, foi necessário transformar a estrutura para o formato *long*. Por fim, todas as bases de dados foram delimitadas para conter o período de 2010 a 2021 como intersecção.

4.5.4. Métodos de Disponibilização de Dados

Após o tratamento e decodificação, os dados da camada tratada foram carregados em um banco de dados PostgreSQL, hospedado em um servidor com Ubuntu Server 22.04 LTS. O servidor possui um processador Intel Core i7-10700KF CPU @ 3.80GHz (8 núcleos, 16 *threads*), placa de vídeo NVIDIA GeForce GTX 1660 Ti e 32 GB de RAM (2x16 GB DDR4 Synchronous 2666 MHz, Team Group Inc). Com o banco de dados devidamente estruturado e codificado, foi implementada uma metodologia para facilitar o acesso aos resultados por meio de uma interface web. Para isso, utilizou-se o framework *Flask* em *Python*. Os scripts desenvolvidos para a criação da aplicação web foram armazenados em um repositório disponível em: https://github.com/pdrmglc/brasildatahub.

4.5.5. Segurança de Acesso aos Dados

Para garantir a segurança e integridade dos dados, a aplicação é hospedada em um servidor da Fiocruz, que conta com soluções corporativas de antivírus e uma equipe dedicada ao monitoramento e à manutenção dos serviços. Além disso, foram implementadas estratégias recomendadas, como mecanismos de limitação de taxa (*rate limiting*), monitoramento contínuo e o uso de firewalls de aplicação web (WAF), para reforçar a proteção contra acessos não autorizados e potenciais ameaças. (Clincy; Shahriar, 2018).

4.5.6. Otimização e velocidade na entrega dos dados

Dada a grande quantidade de dados, foi implementada uma arquitetura de cache otimizada para aprimorar a eficiência na entrega das informações. Utilizando Redis, o servidor armazena em memória RAM os caminhos dos arquivos estáticos, permitindo que a API os entregue de forma mais rápida. Para essa funcionalidade, foram geradas todas as possíveis combinações entre as bases de dados e os anos de consulta, garantindo que os arquivos estivessem localmente disponíveis para envio imediato. Além disso, para otimizar ainda mais a entrega, os arquivos foram compactados, reduzindo a necessidade de processamento e uso de rede. A ferramenta ZIP nativa do Linux foi usada para compactar os arquivos CSV, enquanto o GZIP foi aplicado para a compactação otimizada de arquivos JSON.

O desempenho dessa solução foi avaliado com o *software* Locust, que simulou o comportamento de usuários reais através de agentes simuladores realizando requisições simultâneas. Os experimentos reproduziram cenários de acessos simultâneos à aplicação utilizando *bots*, aproximando-se de situações de uso real. Cada simulação foi realizada em triplicata para lidar com a variabilidade, e o cálculo do tempo de resposta foi baseado na estimativa do tamanho médio dos arquivos, variando o número de usuários simultâneos. Para cada replicata, foi registrada a mediana do tempo necessário para o servidor responder a cada solicitação.

4.5.7. Critérios de exclusão de variáveis

As variáveis que não possuíam uma descrição clara ou adequada nos seus respectivos dicionários de dados e definições originais foram removidas durante o

processo de transformação, a fim de evitar possíveis interpretações equivocadas. Adicionalmente, as variáveis de metadados presentes nos bancos analisados foram excluídas, por não estarem alinhadas com o objetivo deste estudo.

4.5.8. Aspectos éticos

Todas as bases de dados utilizadas neste estudo são de acesso público e irrestrito. Além disso, os dados são previamente anonimizados pela fonte geradora, garantindo a impossibilidade de identificar qualquer indivíduo nas informações utilizadas.

4.5.9. Agregação de dados e engenharia de variáveis

O banco de dados PostgreSQL criado anteriormente foi utilizado para construir uma nova camada de tabelas, agora agrupadas por mês, ano e código de 6 dígitos de cada município brasileiro. Assim, as bases de dados SINAN, CNES, SIM e INMET, as quais estão a nível de indivíduo, foram agrupadas para atingir o nível de granularidade de mês, ano e município semelhante às demais bases. Desta forma, para variáveis quantitativas, foram calculadas as seguintes métricas: média, desvio padrão, mediana, intervalo interquartil, máximo e mínimo. Já para as variáveis categóricas, obteve-se a contagem e a porcentagem de cada categoria frente a variável original. O apêndice A indica todas as variáveis que foram derivadas a partir das bases SINAN, CNES, SIM e INMET para a construção da tabela base analítica ao se conectar todas as bases pela chave composta. As demais bases (IBGE PIP, IBGE POP e *MapBiomas* – nível de município e ano) foram utilizadas diretamente após o tratamento dos dados mencionado anteriormente e as variáveis utilizadas podem ser vistas em https://github.com/pdrmglc/ETL-ML-Dengue/blob/main/dicionarios/dicionarios_banco.zip.

4.5.10. Classificação da Situação Mensal dos Municípios

A classificação da situação mensal de cada município foi realizada com base no "PLANO DE CONTINGÊNCIA PARA RESPOSTA ÀS EMERGÊNCIAS EM SAÚDE PÚBLICA POR DENGUE, CHIKUNGUNYA E ZIKA" (Ministério da Saúde, 2022). As categorias definidas foram "Resposta Inicial", "Alerta", "Emergência" e "Normal", sendo esta última utilizada para municípios cuja situação não se enquadrava nas categorias anteriores. Essa abordagem permite uma melhor compreensão do estado de saúde pública

em relação a essas doenças, facilitando a alocação de recursos e a tomada de decisões estratégicas por parte das autoridades sanitárias. Contudo, dada a análise inicial dos dados, identificou-se um desbalanceamento substancial entre as categorias de alerta para dengue. As classes "Normal" (0) e "Resposta Inicial" (1) abrangem a maioria das observações, enquanto as categorias "Alerta" (2) e "Emergência" (3) possuem frequências muito baixas, com menos de 0,05% cada. Para melhorar a capacidade preditiva do modelo e mitigar vieses decorrentes do desbalanceamento extremo, optou-se por uma reclassificação, consolidando as categorias "Resposta Inicial", "Alerta" e "Emergência" em uma única classe denominada como “Não normal”.

*4.5.11. Mode*lag*em de Predição*

Modelos preditivos foram desenvolvidos para estimar casos futuros de dengue com uma previsão de três meses. O objetivo principal foi criar modelos que permitam uma resposta antecipada, possibilitando a implementação de ações adequadas em tempo hábil.

Ainda, devido ao significativo desbalanceamento entre as classes (frequência das categorias da variável dependente), foi adotado um sistema de pesos para favorecer as classes menos frequentes, garantindo que o modelo possa captar melhor as variáveis associadas a essas categorias. Essa abordagem visa melhorar a acurácia das previsões e, consequentemente, a eficácia das intervenções de saúde pública.

4.5.12. Divisão dos Dados

Para a criação das categorias a serem preditas, foram utilizadas a incidência média e o intervalo de confiança de 95% para os dados de 2010 a 2012 como calibrador para o estabelecimento do canal endêmico e dos limites superior e inferior conforme descrito em Ministério da Saúde, (2022). Com isso, a divisão final dos dados foi realizada da seguinte maneira: o conjunto de treino compreendeu os dados de 2013 a 2020, representando 80% do total; o conjunto de teste incluiu 20% dos dados de 2013 a 2020; a validação foi realizada por meio de uma abordagem "*out-of-time*" (OOT), onde os dados de 2021 foram totalmente segregados como conjunto de validação (n = 13.453). O objetivo é avaliar a capacidade de generalização do modelo e a precisão de suas previsões em dados futuros, além dos conjuntos de treino e teste convencionais.

Considerando a natureza não convencional da abordagem de séries temporais, foi assumido que cada combinação de mês, ano e município representa um indivíduo distinto. Essa suposição permitiu a integração de um grande volume de dados sem a necessidade de criar séries temporais separadas por município, possibilitando responder à seguinte questão: "Dado que um município possui essas características (*features*), qual é o número esperado de casos no mês seguinte?" Essa metodologia visa melhorar a precisão das previsões e fornecer abordagens mais robustas para a gestão de surtos de dengue.

4.5.13. Modelos Utilizados

O algoritmo *Gradient Boosting* foi selecionado para a etapa de seleção de variáveis devido à sua capacidade de lidar com a complexidade dos dados, realizando uma seleção eficiente das *features* mais relevantes. Além disso, o *Gradient Boosting* pode capturar relações complexas e não lineares entre as variáveis preditoras, embora não faça uso explícito de métodos tradicionais de séries temporais. Sua escolha se deve também à capacidade de incorporar os erros cometidos em etapas anteriores, de forma semelhante ao que ocorre em modelos de média móvel (MA), otimizando o desempenho preditivo.

4.5.14. Geração de Lags

Foi implementada uma função para a criação de *Lags* para variáveis relevantes, semelhante à abordagem utilizada em modelos de séries temporais, como o modelo autorregressivo (AR). Essa técnica visa aprimorar a capacidade preditiva do modelo, permitindo a captura de padrões temporais e aumentando a precisão das previsões futuras. A inclusão de *Lags* constitui uma estratégia para mimetizar a componente sazonal observada em modelos de séries temporais.

*4.5.15. Rodadas de Mode***lag***em*

Antes da primeira rodada de modelagem, foi gerada a matriz de correlação entre todas as variáveis. As variáveis com correlação igual ou superior a 0,8, considerada uma correlação muito forte (Papageorgiou, 2022) , foram removidas, mantendo aleatoriamente apenas uma de cada par. A primeira rodada de modelagem utilizou os hiperparâmetros padrão do XGBoost para identificar as *features* mais relevantes. Na segunda rodada,

também com hiperparâmetros padrão, foi incluída a criação de *Lags* de 1 a 13 meses para todas as variáveis que apresentaram relevância para o modelo, seguido de um novo filtro para remover variáveis muito correlacionadas, como na etapa anterior. Por fim, foi realizada uma nova rodada de modelagem, utilizando apenas as variáveis relevantes e suas correspondentes *Lags*. As rodadas de modelagem foram empregadas como uma forma de selecionar *features* com alguma relação com a variável dependente, utilizando uma abordagem gulosa (*greedy*) devido ao elevado número de variáveis independentes. Além disso, após a construção dos *Lags*, buscou-se também, de forma gulosa, encontrar os parâmetros de sazonalidade para cada uma das variáveis relevantes no período anterior.

4.5.16. Atualização de Probabilidades Usando o Teorema de Bayes

Para atualizar a probabilidade de ocorrência da categoria "Não normal" com base nas previsões do modelo, foi utilizado o Teorema de Bayes. Primeiramente, as probabilidades a priori das categorias "Normal" e "Não normal" foram determinadas a partir da distribuição inicial das classes no conjunto de dados original. As frequências foram definidas como $P(Normal)$ e $P(Não\ normal)$.

A atualização da probabilidade baseou-se na aplicação da seguinte fórmula bayesiana para eventos condicionais:

$$P(Não\ normal|Positivo) = \frac{P(Positivo|Não\ normal) \times P(Não\ normal)}{P(Positivo)} \quad (9)$$

onde:

- P(Não normal | Positivo) representa a probabilidade a posteriori de a categoria ser "Não normal" dado que o modelo a classificou como tal;
- P(Positivo | Não normal) é a métrica de *Recall* (sensibilidade) do modelo para a categoria "Não normal," que expressa a taxa de verdadeiros positivos para essa categoria;
- P(Não normal) é a probabilidade a priori da categoria "Não normal";
- P(Positivo), a probabilidade marginal de uma previsão positiva (de pertencer a categoria “Não normal”).

5. RESULTADOS

5.1.Avaliação contextualizada da variabilidade genética de DENV

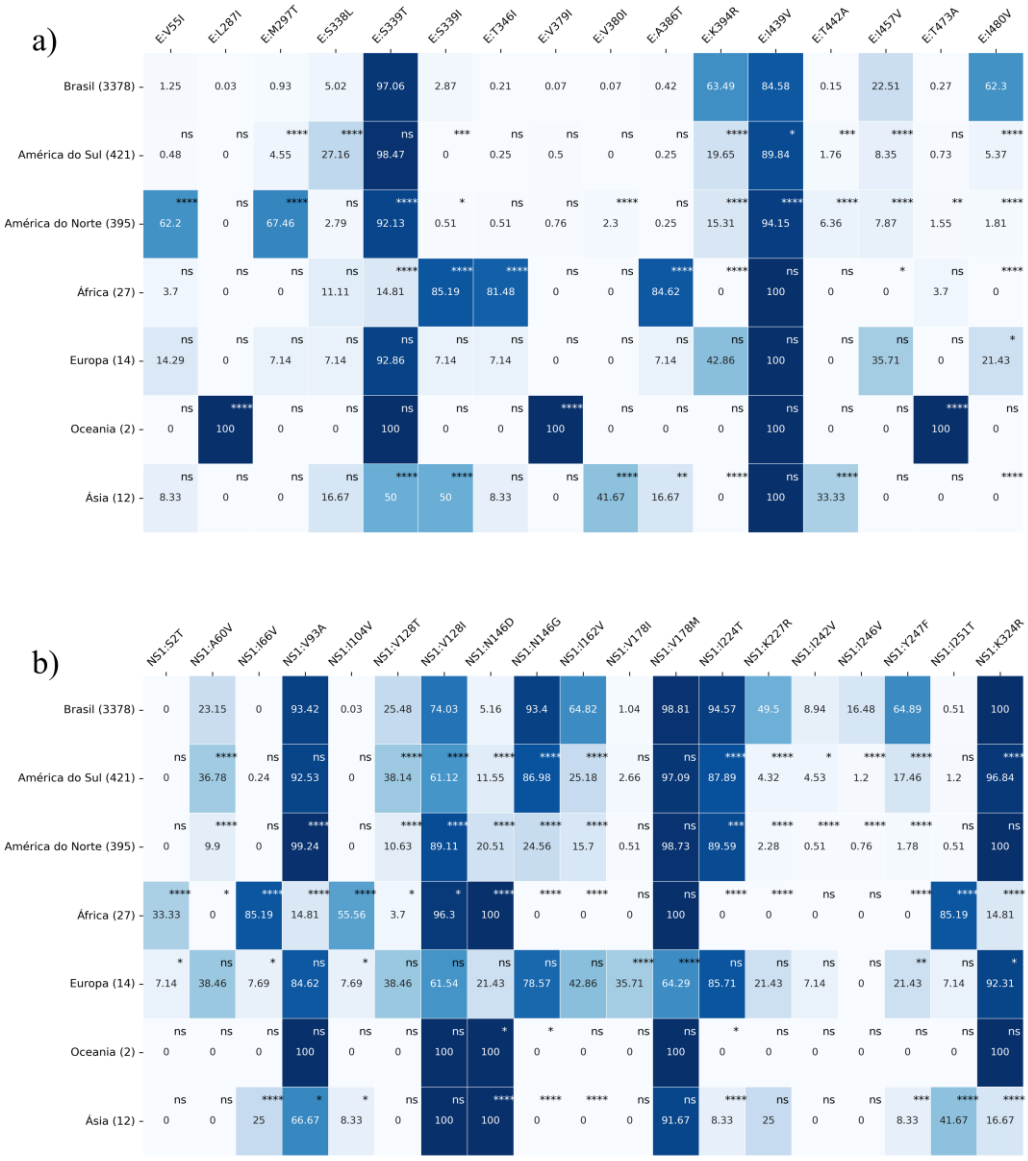
5.1.1. Avaliação geográfica

5.1.1.1. DENV1

5.1.1.1.1. E e NS1

A análise mutacional da proteína E do DENV1 para as diferentes regiões geográficas avaliadas (Brasil, América do Sul, América do Norte, África, Europa, Oceania e Ásia) revelou a mutação E:I480V (62,3%) como única diferença estatística em relação às demais regiões. A proteína NS1, por sua vez, apresentou a mutação NS1:Y247F (64,89%), significativamente diferente em todas as regiões, exceto na Oceania (N = 2).

Figura 07: Distribuição geográfica de mutações nas proteínas E e NS1 de DENV1-V. O *heatmap* apresenta a frequência relativa de cada mutação em diferentes regiões geográficas. As cores mais intensas indicam maior frequência. Diferenças estatísticas significativas na frequência da mutação entre o Brasil e as demais regiões foram avaliadas pelo teste de qui-quadrado ou teste exato de fisher (p-ajustado pelo método de Bonferroni). *p < 0,05; **p < 0,01; ***p < 0,001. ns: não significativo.



Fonte: Elaborado pelo autor com dados disponíveis em gisaid.org

5.1.1.2.DENV2

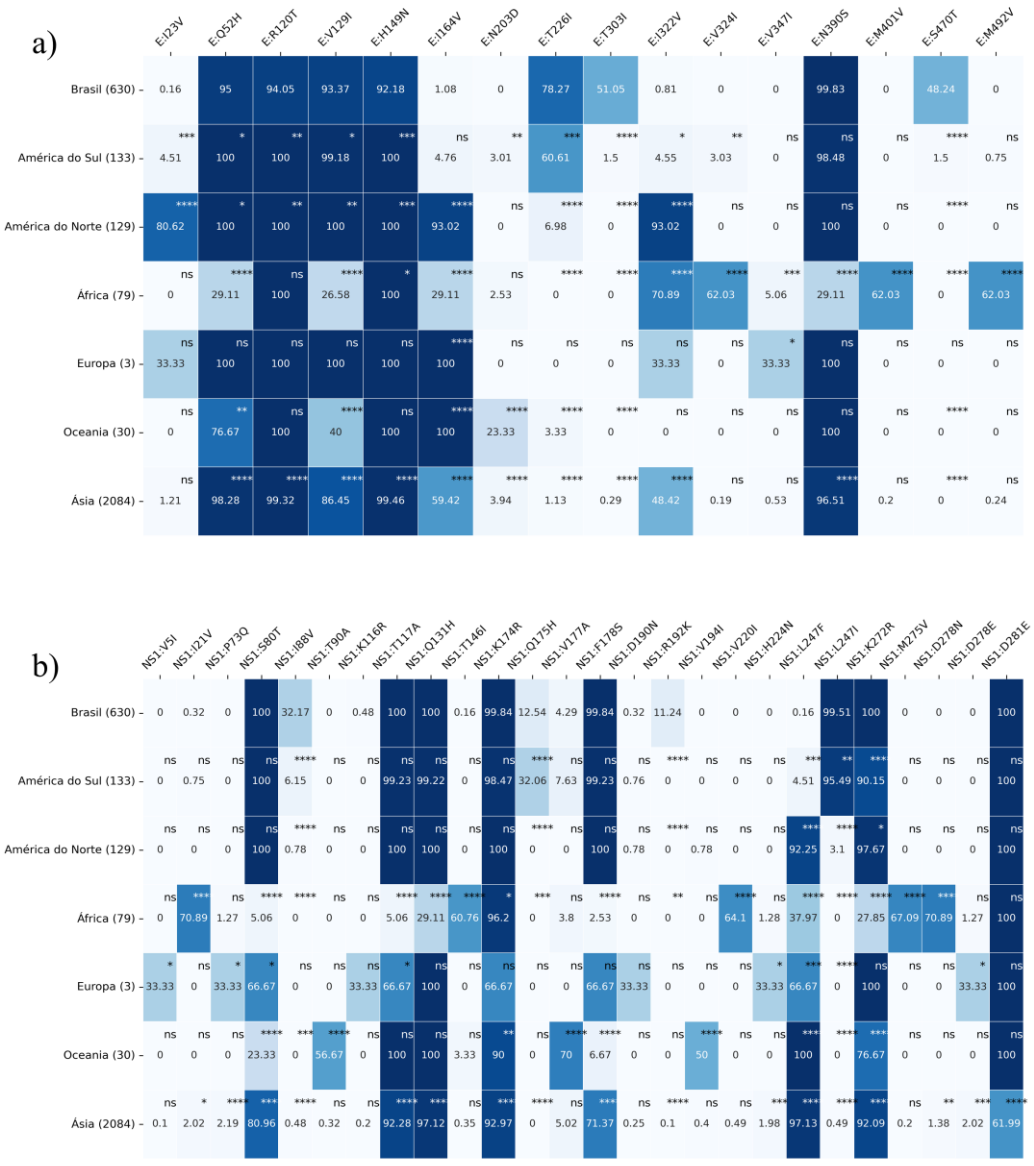
5.1.1.2.1. Genótipo II

5.1.1.2.1.1. E e NS1

A análise da proteína E do DENV2-II revelou as mutações E:T226I (78,27%), E:T303I (51,05%) e E:S470T (48,24%) como significativamente mais frequentes no Brasil em comparação com as demais regiões, exceto a Europa (N=3). A mutação

E:V129I, embora presente em 96,37% dos casos, não apresentou diferenças significativas entre as regiões. A proteína NS1, por sua vez, apresentou as mutações NS1:I88V (32,17%), NS1:Q175R (12,54%) e NS1:R192K (11,24%) como mais frequentes no Brasil, com exceção da Oceania (N=30) e Europa.

Figura 08 - Distribuição geográfica de mutações nas proteínas E e NS1 de DENV2-II. O *heatmap* apresenta a frequência relativa de cada mutação em diferentes regiões geográficas. As cores mais intensas indicam maior frequência. Diferenças estatísticas significativas na frequência da mutação entre o Brasil e as demais regiões foram avaliadas pelo teste de qui-quadrado ou teste exato de fisher (p-ajustado pelo método de Bonferroni). *p < 0,05; **p < 0,01; ***p < 0,001. ns: não significativo.



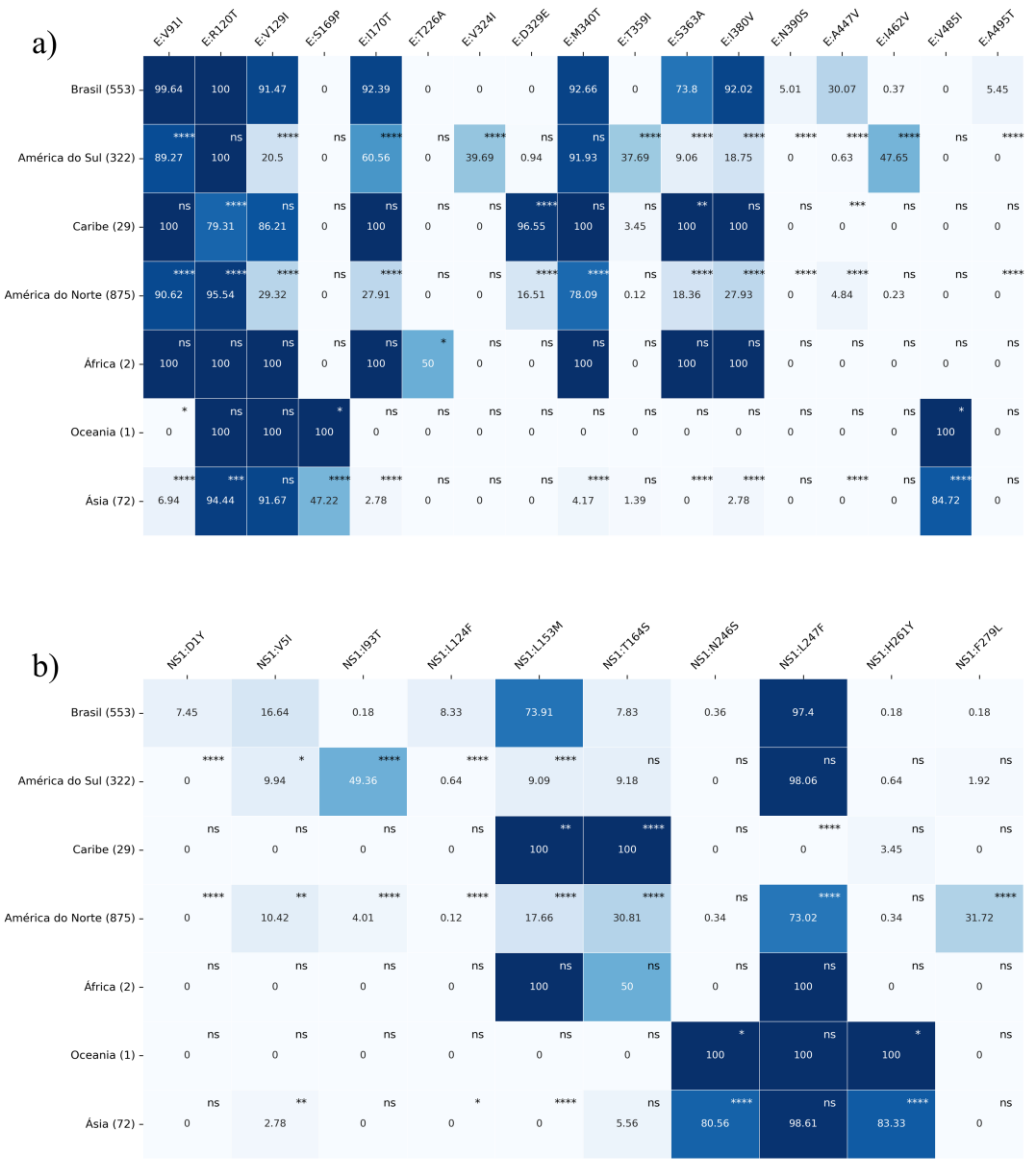
Fonte: Elaborado pelo autor com dados disponíveis em gisaid.org

5.1.1.2.2. Genótipo III

5.1.1.2.2.1. E e NS1

A mutação E:S363A, embora presente no Caribe (100%), mostrou-se significativamente diferente no Brasil (73,8%), sendo majoritariamente encontrada apenas nessas regiões. A mutação E:A447V tem sua frequência significativamente distinta para todas as regiões, exceto África (N=2) e Oceania (N=1). A proteína NS1 apresentou a mutação NS1:L153M, que também está presente no Caribe, porém com uma frequência significativamente diferente.

Figura 09 - Distribuição geográfica de mutações nas proteínas E e NS1 de DENV2-III. O heatmap apresenta a frequência relativa de cada mutação em diferentes regiões geográficas. As cores mais intensas indicam maior frequência. Diferenças estatísticas significativas na frequência da mutação entre o Brasil e as demais regiões foram avaliadas pelo teste de qui-quadrado ou teste exato de fisher (p-ajustado pelo método de Bonferroni). *p < 0,05; **p < 0,01; ***p < 0,001. ns: não significativo.

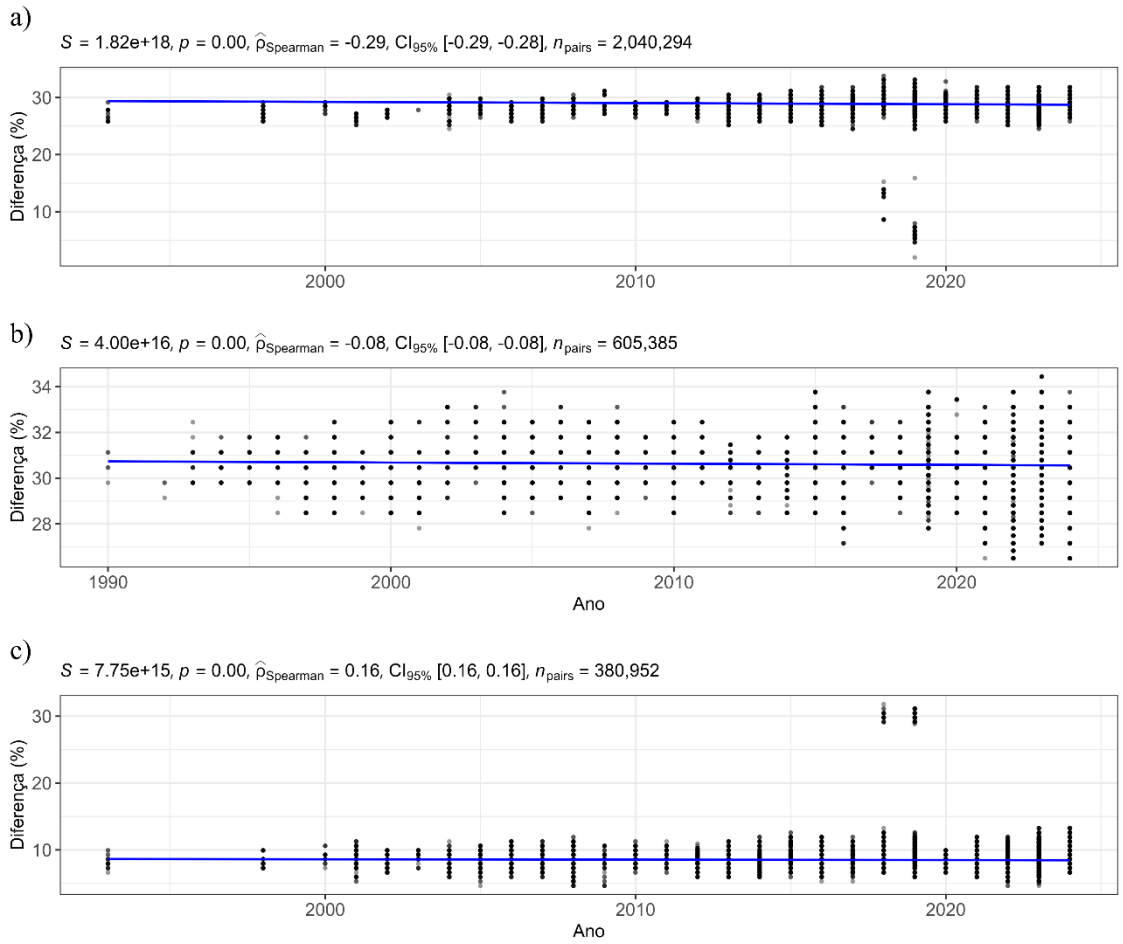


Fonte: Elaborado pelo autor com dados disponíveis em gisaid.org

5.1.2. Avaliação temporal

Todos os testes de comparação entre os grupos DENV1 vs. DENV2-II, DENV1 vs DENV2-III e DENV2-II vs DENV2-III mostraram significância estatística (p-valor < 0,05). A correlação de Spearman foi fraca e negativa para DENV1 vs. DENV2-II (n = 2.040.294), muito fraca e negativa para DENV1 vs DENV2-III (n = 605.385), e fraca e positiva para DENV2-II vs DENV2-III (n = 380.952).

Figura 10 - Gráficos de dispersão representando a análise estatística das diferenças entre os grupos DENV1, DENV2-II e DENV2-III ao longo dos anos. Correlações de Spearman indicam uma relação fraca e negativa entre DENV1 e DENV2-II (a), muito fraca e negativa entre DENV1 e DENV2-III (b), e fraca e positiva entre DENV2-II e DENV2-III (c). Significância estatística foi alcançada em todas as comparações (p-valor < 0,05), com a presença de outliers nos gráficos de DENV1-V vs DENV2-II e DENV2-II vs DENV2-III.



Fonte: Elaborado pelo autor com dados disponíveis em gisaid.org

Houve, ainda, outliers perceptíveis nos gráficos comparando DENV1-V e DENV2-II e comparando DENV2-II e DENV2-III. As sequências em questão pertencem a dois códigos de acesso específicos, EPI_ISL_18579091 e EPI_ISL_18579100, tabela 2. A comparação com os genomas de referência para DENV1 e DENV2 revela uma maior similaridade das proteínas C, prM, E e NS1 com DENV1 para a sequência EPI_ISL_18579091 - NUC: 88,40% ~ 95,00%; AA: 92,10% ~ 97,00% vs. DENV2 - NUC: 68,20% ~ 72,20%; AA: 69,30% ~ 75,80%. Ao passo que, para as proteínas NS2A, NS2B, NS3, NS4A, 2K, NS4B e NS5, a similaridade é maior para DENV2, onde para DENV1 – NUC: 47,60% ~ 72,20%; AA: 39,00% ~ 79,60% vs. DENV2 – NUC: 89,50% ~ 92,50%; AA: 96,90% ~ 100,00%.

Para a sequência EPI_ISL_18579100, a comparação com os genomas de referência para DENV1 e DENV2 revela uma maior similaridade apenas com as proteínas C, prM, E com DENV1 - NUC: 84,80% ~ 92,90%; AA: 87,50% ~ 92,80% vs. DENV2 - NUC: 72,10% ~ 72,90%; AA: 71,90% ~ 75,20%. Ao passo que, para as proteínas NS1, NS2A, NS2B, NS3, NS4A, 2K, NS4B e NS5, a similaridade é maior para DENV2, onde para DENV1 – NUC: 47,70% ~ 78,10%; AA: 39,50% ~ 83,80% vs. DENV2 – NUC: 84,70% ~ 91,50%; AA: 87,80% ~ 100,00%.

Tabela 02 - identidade percentual de nucleotídeos (NUC) e aminoácidos (AA) entre sequências dos genes C, prM, E, NS1, NS2A, NS2B, NS3, NS4A, 2K, NS4B e NS5 dos isolados virais hDenV2/Sri_Lanka/UNSW-S521/2019 e hDenV2/Sri_Lanka/UNSW-S082/2018, comparados com os genomas de referência DENV1 (NC_001477.1) e DENV2 (NC_001474.2). A comparação revela maior similaridade de C, prM, E e NS1 com DENV1 e das proteínas NS2A, NS2B, NS3, NS4A, 2K, NS4B e NS5 com DENV2.

Referência	Gene	Identidade			
		hDenV2/Sri_Lanka/UNSW-S521/2019|EPI_ISL_18579091|2019-07-09		hDenV2/Sri_Lanka/UNSW-S082/2018|EPI_ISL_18579100|2018-04-01	
		Nucleotídeo	Aminoácido	Nucleotídeo	Aminoácido
NC_001477.1 (DENV1)	C	95.00%	95.60%	92.90%	91.20%
	prM	91.00%	97.00%	88.20%	92.80%
	E	88.40%	92.10%	84.80%	87.50%
	NS1	89.90%	93.50%	77.00%	80.40%
	NS2A	47.60%	39.00%	47.70%	39.50%
	NS2B	62.80%	60.00%	62.80%	60.00%
	NS3	72.20%	79.60%	72.60%	79.40%
	NS4A	57.70%	59.80%	57.50%	59.80%
	2K	52.20%	56.50%	52.20%	56.50%
	NS4B	68.90%	77.00%	71.90%	77.40%
	NS5	70.90%	79.30%	78.10%	83.80%
NC_001474.2 (DENV2)	C	70.50%	69.30%	72.90%	71.90%
	prM	69.10%	75.80%	72.10%	75.20%
	E	68.20%	71.50%	72.30%	74.30%
	NS1	72.60%	75.00%	84.70%	87.80%
	NS2A	90.40%	97.20%	90.40%	96.80%
	NS2B	92.10%	96.90%	91.50%	96.90%
	NS3	91.90%	98.40%	91.10%	97.70%
	NS4A	89.50%	97.60%	89.50%	97.60%
	2K	91.30%	100.00%	91.30%	100.00%
	NS4B	92.30%	98.00%	86.60%	93.10%
	NS5	92.50%	97.60%	85.00%	90.10%

Fonte: Elaborado pelo autor com dados disponíveis em gisaid.org

5.2. Desenho de *primers* para sequenciamento genômico

5.2.1. Características físico-químicas dos primers

O conjunto de *primers* gerados pode ser visto na tabela suplementar 1. Nenhum dos *primers* investigados apresentou temperatura de *melting* (Tm) maior do que 50 °C, com apenas a formação de um heterodímero com ΔG menor que -9 kcal/mol entre os *primers* DENV1_800_PS_2_LEFT e DENV1_800_PS_14_LEFT.

5.2.2. Compatibilidade de primers para sequências genômicas brasileiras

O pareamento *in silico* dos *primers*, utilizados neste trabalho, com as sequências que geraram a construção dos painéis específicos para o Brasil, resultou na identificação de *mismatches*. Especificamente, apenas os *primers* obtidos do protocolo internacional apresentaram *mismatches*. A tabela 3 resume o número de variações encontradas para cada painel de sequenciamento.

Tabela 03 - Desempenho dos painéis de *primers* para DENV1 e DENV2. A tabela apresenta o número de *mismatches*, *offtargets* e *primers* hibridizados para diferentes painéis de sequenciamento direcionados aos sorotipos DENV1 e DENV2. "DENV1 (ref.)" e "DENV2 (ref.)" referem-se a painéis internacionais direcionados para DENV1 e DENV2, respectivamente. "DENV1 400bp", "DENV1 600bp" e "DENV1 800bp" correspondem aos painéis de *primers* desenhados para amplificar regiões de diferentes tamanhos do genoma do DENV1. "Pan-DENV2BR" refere-se a um painel de *primers* desenvolvido especificamente para DENV2 no Brasil.

Painel	**Origem**	**Nº *mismatches***	**Nº *offtargets***	**Nº *primers* hibridizados**
DENV1 (ref.)	Vogels *et al.,* 2023	26	1	45
DENV1 400bp	Este estudo	0	0	66
DENV1 600bp	Este estudo	0	0	39
DENV1 800bp	Este estudo	0	0	28
DENV2 (ref.)	Vogels *et al.,* 2023	29	0	61
Pan-DENV2BR	Este estudo	0	0	54

Fonte: Elaborado pelo autor.

5.2.3. Sequenciamento genético

A avaliação dos painéis de *primers*, tabela 4, revelou uma cobertura genômica consistente para a maioria das amostras testadas, independentemente do tamanho do *amplicon* utilizado. As coberturas médias obtidas foram de aproximadamente 97,7% para o painel de 400 bases, 97,3% para o painel de 600 bases e 96,6% para o painel de 800 bases.

Duas amostras (RJ001 e RJ010) apresentaram falha completa no sequenciamento, independentemente do painel utilizado, com cobertura genômica de 0% e profundidade de leitura ausente. Além disso, a amostra RJ005 apresentou desempenho abaixo do esperado, com coberturas de 57,8%, 46,1% e 36,5% para os painéis de 400, 600 e 800 bases, respectivamente, acompanhadas de baixas profundidades de sequenciamento. O controle negativo (NTC) não apresentou amplificação significativa para os painéis de 400 e 600 bases, mas gerou uma cobertura de 29,1% no painel de 800 bases.

O número de *primers* defeituosos foi baixo em todos os protocolos, sendo detectado apenas em três amostras (RJ003, RJ006 e RJ012). Nenhum padrão claro de falha foi identificado entre os diferentes painéis.

Tabela 04 - Desempenho dos painéis de *primers* de diferentes tamanhos de *amplicon* (400, 600 e 800 bases) na validação do protocolo de sequenciamento para DENV-1. Os valores apresentados correspondem à cobertura genômica (%), profundidade média de sequenciamento e o número de *primers* considerados defeituosos para cada amostra testada.

Código Amostra	**Protocolo - Tamanho médio do *amplicon***								
	400 bases			**600 bases**			**800 bases**		
	Cobertura (%)	**Profundidade**	***Bad primers***	**Cobertura (%)**	**Profundidade**	***Bad primers***	**Cobertura (%)**	**Profundidade**	***Bad primers***
RJ001	0	0	0	0	0	0	0	0	0
RJ002	97.7%	1758	0	97.3%	2940	0	96.6%	3806	0
RJ003	97.7%	1384	1	97.1%	2060	1	96.6%	3299	0
RJ004	97.7%	1502	0	97.3%	2114	0	96.6%	3711	0
RJ005	57.8%	7	0	46.1%	3	0	36.5%	2	0
RJ006	97.7%	1623	1	97.3%	2455	0	96.6%	3230	0
RJ007	97.7%	257	0	95.0%	84	0	96.6%	604	0
RJ008	97.7%	591	0	97.3%	445	0	96.6%	1438	0
RJ009	97.7%	884	0	97.3%	1294	0	96.6%	2244	0
RJ010	0.7%	0	0	0.0%	0	0	0.0%	0	0
RJ011	97.7%	392	0	97.3%	561	0	96.6%	1082	0
RJ012	97.7%	333	0	97.3%	297	0	96.5%	785	1
AM001	97.7%	1387	0	97.3%	1606	0	96.6%	2615	0
AM002	97.7%	1815	0	97.3%	2011	0	96.6%	2119	0
AM003	97.7%	1880	0	97.3%	2626	0	96.6%	2806	0
NTC	0	0	-	0	0	-	29.1%	3	-

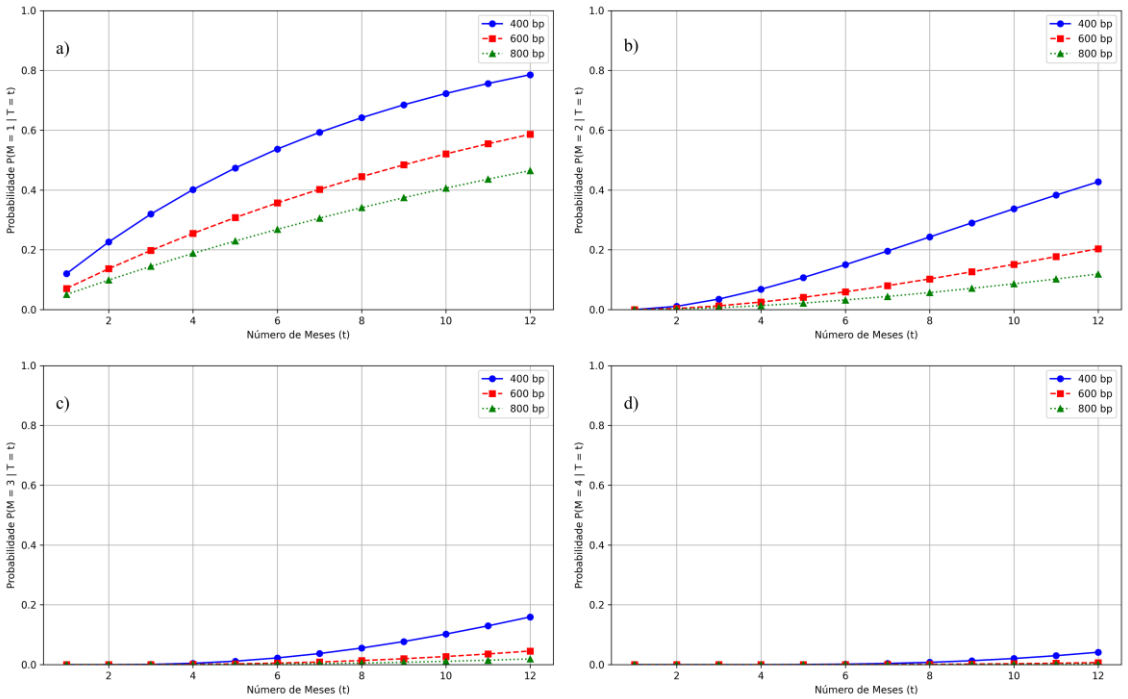
5.3. Aprimoramento da atualização de painéis de *primers*

5.3.1. Tempo de vida útil dos primers por tamanho de amplicon

A figura 11a ilustra a probabilidade de ocorrência de mutações em regiões de *primer* ao longo do tempo, para diferentes tamanhos de *amplicons*, modelada a partir da equação 6. Observa-se que a probabilidade de mutação aumenta com o tempo de circulação contínua do vírus em todos os tamanhos de *amplicons* analisados. Especificamente, para *amplicons* de 400 bp, a probabilidade de ocorrer uma mutação superior a 50%, valor arbitrariamente escolhido para representar uma linha de corte, é atingida após aproximadamente 6 meses de circulação viral. Para *amplicons* de 600 bp, este ponto é alcançado em torno de 9 meses, enquanto para *amplicons* de 800 bp, a probabilidade de 50% é atingida após aproximadamente 12 meses. Este padrão se repete para os casos de ocorrência de pelo menos 2, 3 e 4 mutações em regiões de *primer*, mas com uma redução expressiva no aumento da probabilidade. Após 12 meses de circulação contínua do vírus, a probabilidade de ocorrer pelo menos 2 mutações em regiões de *primer* é maior para *amplicons* menores, devido a uma maior porcentagem do genoma do organismo sendo coberta por *primers*. *Amplicons* de 400 bp apresentaram a maior probabilidade (42.74%), seguidos por *amplicons* de 600 bp (20.38%) e 800 bp (11.91%). No entanto, a probabilidade de ocorrer 3 ou 4 mutações em 12 meses é extremamente baixa, sendo praticamente nula para todos os tamanhos de *amplicon* analisados.

Figura 11 - Gráfico de estimativa de tempo de vida útil de *primers* com diferentes tamanhos de *amplicon*. O eixo x representa o período de circulação contínua do vírus (em meses), enquanto o eixo y indica a probabilidade de ocorrer pelo menos *m* mutações em uma região de *primer* ao longo do tempo. Cada painel (a-d) corresponde a diferentes valores de mmm, onde mmm é o

número mínimo de mutações necessárias para comprometer a integridade da região de *primer*: **a)** m = 1, **b)** m = 2, **c)** m = 3 e **d)** m = 4.



Fonte: Elaborado pelo autor.

5.3.2. Hydra: software para acompanhamento de painéis de amplicon

Os resultados das análises realizadas em triplicata foram os seguintes: na primeira execução, o tempo foi de 174,45 segundos, com uso de 394,09 MB de memória; na segunda execução, o tempo foi de 176,02 segundos, com memória utilizada de 296,96 MB; e na terceira execução, o tempo foi de 175,96 segundos, com uso de 250,56 MB de memória. Os dados estão compilados na tabela 5.

Tabela 05 - Tabela de resultados das execuções em triplicata, com registro do tempo de processamento em segundos e uso de memória em megabytes (MB).

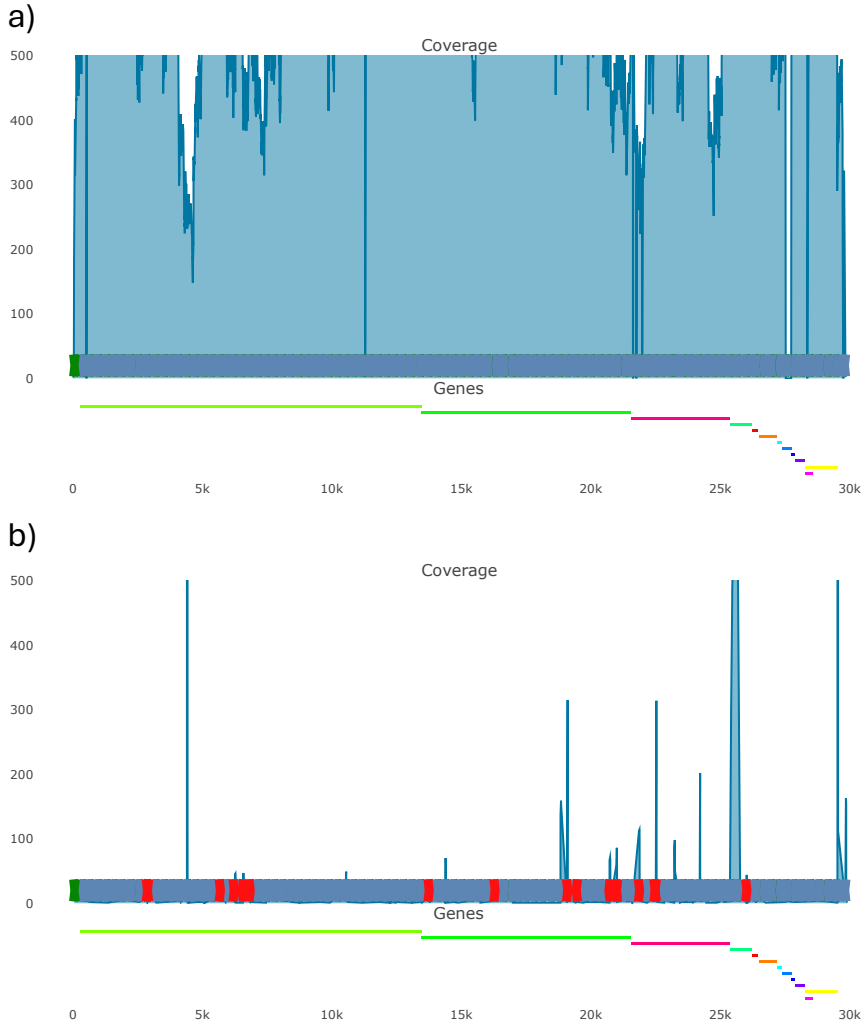
Execução	**Tempo (segundos)**	**Memória (MB)**
1ª	174.45	394.09
2ª	176.02	296.96
3ª	175.96	250.56
Média	175.48	313.87
Desvio padrão	0.73	59.80

Fonte: Elaborado pelo autor.

A figura 12 apresenta, em a), uma amostra com perfil aceitável para sequenciamento, enquanto em b) mostra uma amostra com perfil que deve ser rejeitado. Logo abaixo dos gráficos de cobertura vertical, estão representados os *primers* do arquivo .bed, com *primers forward* em verde, reverse em azul, e *primers* com possível *mismatch* destacados em vermelho. O gráfico inferior exibe a distribuição dos genes ao longo do genoma. A interface gráfica do Hydra também oferece uma funcionalidade interativa, permitindo a visualização da cobertura vertical em nível de nucleotídeo, indicando a profundidade para cada base nitrogenada.

Figura 12 - Gráficos de cobertura vertical para amostras de SARS-CoV-2: a) Amostra com perfil aceitável para sequenciamento; b) Amostra com perfil rejeitado. Abaixo dos gráficos de cobertura vertical, estão representados os *primers* do arquivo .bed: *primers forward* em verde, *reverse* em

azul e *primers* com possível problema de *mismatch* em vermelho. O gráfico inferior ilustra a distribuição dos genes ao longo do comprimento do genoma.



Fonte: Elaborado pelo autor.

5.4. Arquitetura e modelagem de dados para vigilância epidemiológica

5.4.1. Compilação dos dados

A tabela 6 compila as bases de dados utilizadas, incluindo características relevantes de cada uma. Algumas fontes de dados exigiram um tratamento adicional, conforme

descrito na seção de metodologia. O SINAN fornece dados diários sobre notificações de dengue, totalizando 16.328.086 registros e 107 variáveis. O SIM disponibiliza dados diários sobre mortalidade no mesmo período, com 15.779.052 registros e 54 variáveis. O CNES contém dados diários sobre estabelecimentos de saúde, somando 40.192.000 registros e 183 variáveis. O IBGE oferece dados anuais sobre população e PIB, com 66.830 registros e 3 variáveis para a população, e 66.825 registros e 44 variáveis para o PIB. O INMET disponibiliza dados climáticos mensais, totalizando 803.917 registros e 22 variáveis. Por fim, o *MapBiomas* fornece dados anuais sobre uso do solo, com 78.432 registros e 34 variáveis.

Tabela 06 - Resumo das Bases de Dados Utilizadas no Estudo. A tabela apresenta um resumo das bases de dados utilizadas no estudo, detalhando as subdivisões, categorias, granularidade temporal, período da série histórica (2010 a 2021), quantidade de linhas e colunas, e a origem dos dados. As bases incluem notificações de dengue, mortalidade, estabelecimentos de saúde, dados socioeconômicos, PIB, clima e uso do solo, provenientes de fontes como DATASUS, IBGE, INMET e *MapBiomas*.

Base	Subdivisão	Categoria	Granularidade	Número de registros	Número de variáveis	Origem
SINAN	Dengue	Notificação de agravos	dia	16328086	107	https://datasus.saude.gov.br/transferencia-de-arquivos/
SIM	-	Mortalidade	dia	15779052	54	https://datasus.saude.gov.br/transferencia-de-arquivos/
CNES	ST	Estabelecimentos de saúde	dia	40192000	183	https://datasus.saude.gov.br/transferencia-de-arquivos/
IBGE	População	Socioeconômico	ano	66830	3	https://datasus.saude.gov.br/transferencia-de-arquivos/
	PIB		ano	66825	44	https://datasus.saude.gov.br/transferencia-de-arquivos/
INMET	-	Clima	mês	803917	22	https://bdmep.inmet.gov.br/
MapBiomas	-	Uso do solo	ano	78432	34	https://brasil.mapbiomas.org/estatisticas/

Fonte: Elaborado pelo autor.

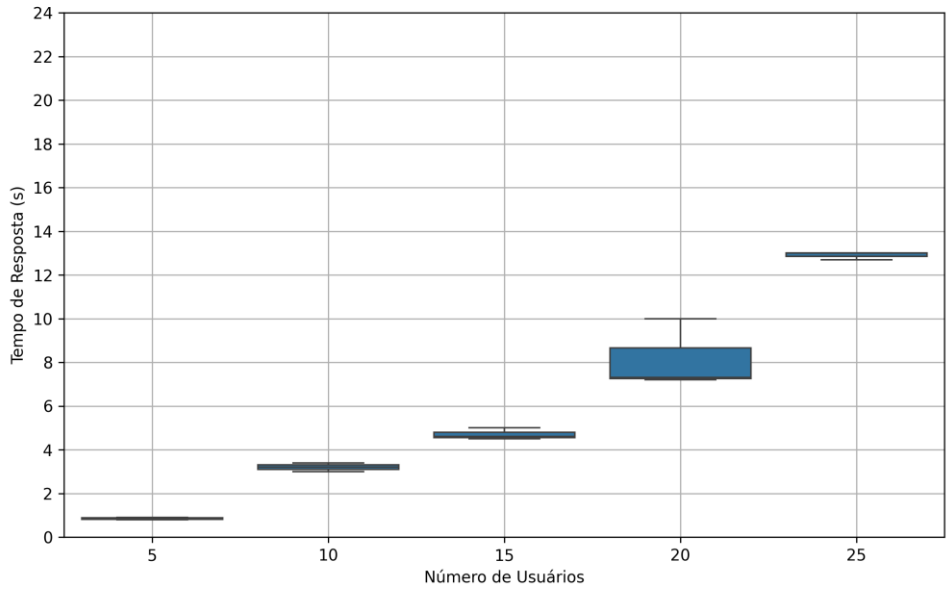
5.4.2. Desempenho do Acesso aos Dados

Durante a modelagem e o armazenamento dos dados, o banco de dados inicial apresentou um tamanho de 35 GB. A conversão dos arquivos para os formatos JSON e CSV resultou em tamanhos que chegaram a até 200 GB. No entanto, a aplicação de uma estratégia de compressão reduziu esses arquivos para menos de 2 GB.

Os testes de simulação realizados para estimar o tempo médio de requisição foram conduzidos em diferentes cenários, nos quais múltiplos usuários acessavam o servidor simultaneamente, solicitando arquivos de variados tamanhos e bases de dados. Nos cenários testados, que incluíram 5, 10, 15, 20 e 25 usuários simultâneos, o tempo médio

de resposta observado foi de 0,8 segundos para 5 usuários, 3,4 segundos para 10, 4,6 segundos para 15, 7,3 segundos para 20 e 12,35 segundos para 25 usuários. Esses resultados refletem o desempenho do servidor sob condições de carga variável, conforme ilustrado na figura 13.

Figura 13: Tempo de resposta do servidor em função do número de usuários simultâneos. *Boxplots* ilustrando a distribuição do tempo de resposta (em segundos) para diferentes cargas de usuários no servidor. Cada *boxplot* representa um conjunto de medidas do tempo de resposta para um número específico de usuários simultâneos.



Fonte: Elaborado pelo autor.

5.4.3. Análise exploratória dos dados de treinamento

A análise exploratória dos dados, conforme apresentado na tabela 7, confirmou um alto desbalanceamento entre as classes de alerta para dengue. As categorias "Normal" (0) e "Resposta Inicial" (1) representavam mais de 99% das entradas, enquanto as classes "Alerta" (2) e "Emergência" (3) apresentavam frequências extremamente baixas, com menos de 0,05% cada uma. Esse padrão de distribuição evidenciado na tabela 7 justificou a fusão das classes minoritárias em uma única categoria de alerta, resultando na nova tabela apresentada na tabela 8.

Tabela 07 - Distribuição das categorias de dengue no conjunto de dados utilizado para a criação de um modelo de *machine learning*. As categorias representam diferentes níveis de alerta: "Normal" (Categoria 0), "Resposta inicial" (Categoria 1), "Alerta" (Categoria 2) e "Emergência"

(Categoria 3). A tabela apresenta o número de ocorrências em cada categoria, a frequência relativa e a porcentagem correspondente para o conjunto de treino.

Categoria	Contagem	Frequência	Porcentagem
Normal	58742	0.4444	44.4379
Resposta inicial	73343	0.5548	55.4834
Alerta	58	0.0004	0.0439
Emergência	46	0.0003	0.0348

Fonte: Elaborado pelo autor.

Tabela 08 - Distribuição das categorias de dengue no conjunto de dados após a fusão das categorias "Resposta inicial" (1), "Alerta" (2) e "Emergência" (3) em uma única classe devido à baixa frequência das últimas duas categorias. A nova tabela mostra a contagem, frequência relativa e porcentagem das duas categorias finais: "Normal" (Categoria 0) e "Não normal" (Categoria 1).

Categoria	Contagem	Frequência	Porcentagem
Normal	58742	0.4444	44.4379
Não normal	73447	0.5556	55.5621

Fonte: Elaborado pelo autor.

5.4.4. Previsão de categorias de surto

No conjunto de Treino, o modelo demonstrou elevada performance, com uma acurácia de 90,40%, F1-*Score* de 0,8920 e ROC AUC de 0,9706, indicando boa capacidade de captura das relações nos dados. Ao mesmo tempo, no conjunto de Teste, as métricas caem ligeiramente – com acurácia de 82,35% e F1-*Score* de 0,7978 –, o que sugere algum nível de *overfitting*, porém o modelo mantém um bom desempenho geral, com ROC AUC de 0,9097, o que indica uma performance robusta na discriminação dos casos. Já no conjunto OOT, o modelo mantém resultados satisfatórios, apresentando acurácia de 84,88% e F1-Score de 0,8550. O alto valor de ROC AUC (0,9325) no conjunto OOT reforça a capacidade de generalização do modelo em períodos futuros, destacando a estabilidade do XGBoost em diferentes janelas temporais. A menor queda de performance em relação ao conjunto de Teste indica que o modelo possui uma boa adaptação a dados não vistos, essencial para a previsão de casos futuros.

Tabela 09 - Tabela de desempenho do modelo XGBoost *Classifier* nos conjuntos de Treino, Teste e *Out-of-Time* (OOT) para previsão de três meses. As métricas apresentadas – Acurácia, F1-*Score*, ROC AUC, *Precision* e *Recall* – permitem avaliar a capacidade do modelo em generalizar e identificar corretamente os casos em diferentes períodos.

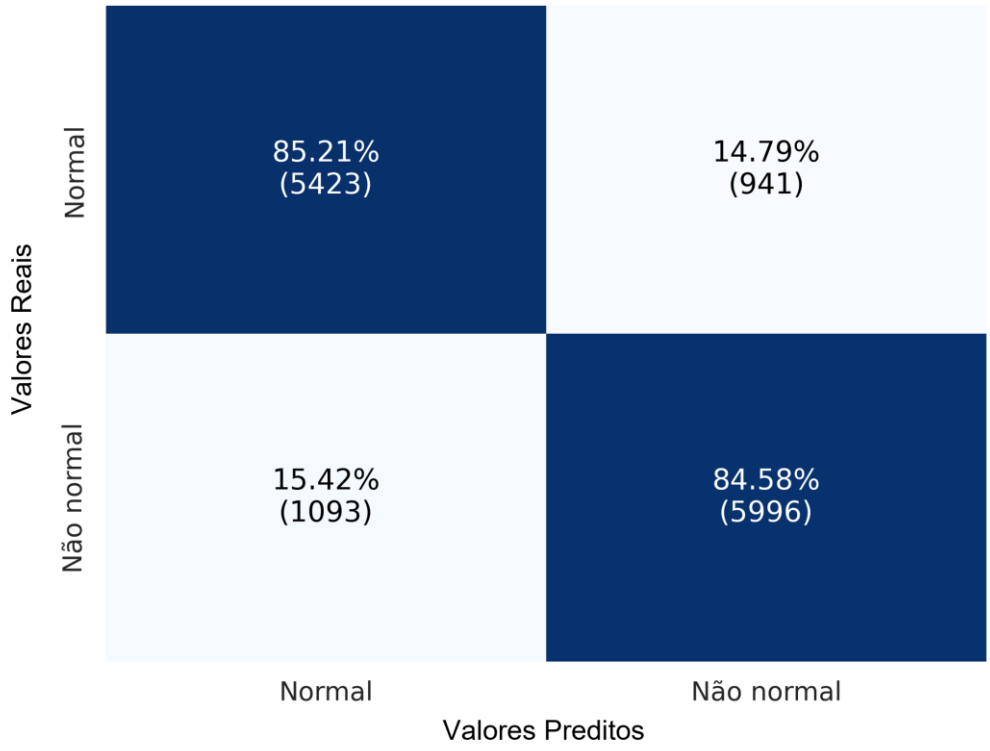
Conjunto	Acurácia	F1-*Score*	ROC AUC	*Precision* (VPP)	*Recall* (Sensibilidade)

Treino	0.904	0.892	0.9706	0.9269	0.8595
Teste	0.8235	0.7978	0.9097	0.8397	0.7598
OOT	0.8488	0.855	0.9325	0.8644	0.8458

Fonte: Elaborado pelo autor.

A matriz de confusão gerada a partir do conjunto de dados OOT, figura 14, revela que o modelo de previsão apresentou um desempenho bom na identificação de situações de aumento de casos prováveis de dengue. O modelo classificou corretamente 84,58% dos casos como 'não normais' (verdadeiros positivos). No entanto, observou-se uma taxa de 14,79% de falsos positivos. Além disso, 15,42% dos casos foram classificados incorretamente como 'normais' (falsos negativos).

Figura 14 - Matriz de confusão para os dados OOT mostrando a comparação entre as previsões do modelo e os rótulos reais.



Fonte: Elaborado pelo autor.

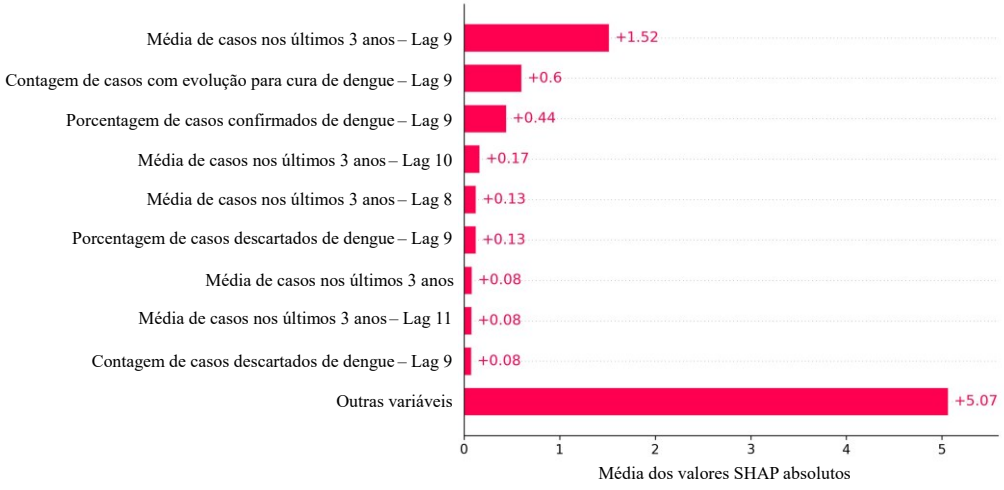
5.4.5. Importância das variáveis para a predição

5.4.5.1. Importância absoluta

A figura 15 mostra a relevância das variáveis mais importantes para a predição. O gráfico de valores SHAP revelou que a variável com maior impacto no modelo preditivo foi a média de casos nos últimos 3 anos com defasagem de 9 meses (*Lag* 9), apresentando um valor médio de SHAP de 1,52. Em seguida, destacaram-se a contagem de casos com evolução para cura de dengue – *Lag* 9 (0,60) e a porcentagem de casos confirmados de dengue – *Lag* 9 (0,44). Outras variáveis relevantes incluíram a média de casos nos últimos 3 anos – *Lag* 10 (0,17), a média de casos nos últimos 3 anos – *Lag* 8 (0,13) e a porcentagem de casos descartados de dengue – *Lag* 9 (0,13). Apesar de menos expressivas individualmente, as demais variáveis somadas na categoria "Outras variáveis" apresentaram um impacto acumulado significativo, totalizando 5,07.

As variáveis apresentaram importâncias próximas e pequenas. Isso indica potencialmente que nenhuma das variáveis possui capacidade preditora satisfatória em relação às demais. Deve se ter cuidado na interpretação destes resultados pois o foco é meramente preditivo, não indicando, em momento algum, relação causal das variáveis encontradas com o aumento anormal de casos.

Figura 15 - Importância das variáveis para a predição. Importância relativa de diversas variáveis no modelo preditivo, com base em *Lags* de diferentes períodos e características relacionadas a dados de casos de dengue.

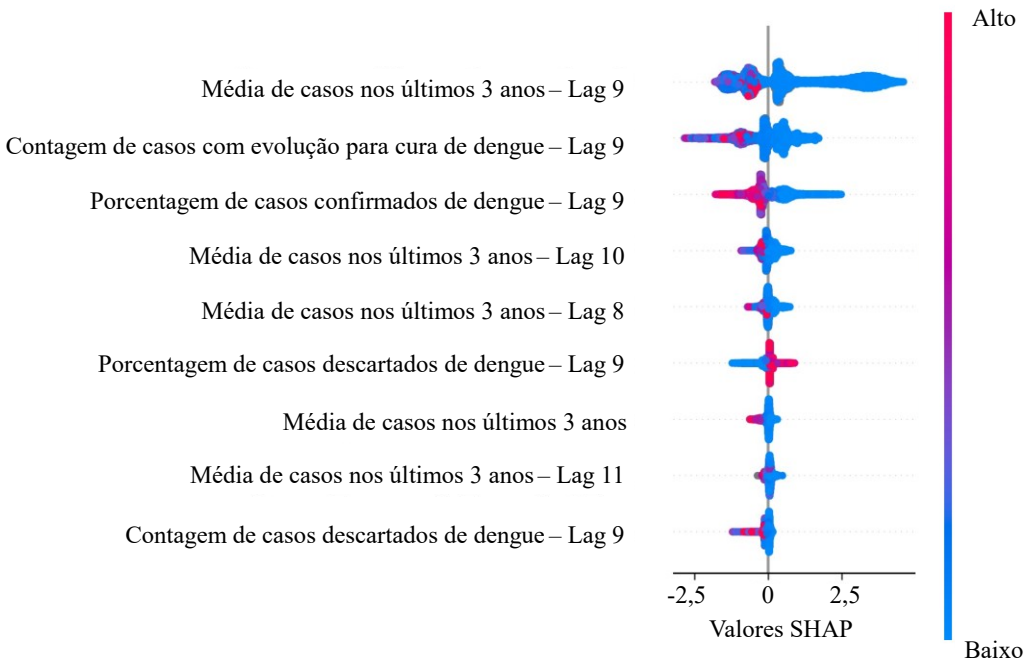


Fonte: Elaborado pelo autor.

5.4.5.2. Importância relativa

O gráfico da figura 16 evidencia que a maioria das variáveis analisadas tem uma relação inversa com a expectativa de ocorrência de um aumento anormal nos casos de dengue (categoria 1). Para variáveis como a média de casos nos últimos 3 anos – *Lag* 9, contagem de casos com evolução para cura de dengue – *Lag* 9, e porcentagem de casos confirmados de dengue – *Lag* 9, valores mais baixos estão associados a uma maior expectativa de ocorrência de casos anormais, conforme indicado pelo agrupamento de pontos azuis (valores baixos) na região positiva dos valores SHAP. No entanto, a porcentagem de casos descartados de dengue – *Lag* 9 se destaca como uma exceção, sugerindo que valores mais altos dessa variável aumentam a expectativa de um aumento anormal nos casos de dengue, representada pelo agrupamento de pontos vermelhos (valores altos) na região positiva dos valores SHAP. Esses resultados indicam, em certo grau, a capacidade indireta do modelo de considerar a informação sobre a suscetibilidade da população após ciclos de aumento casos, indicando que uma menor frequência de casos no ano anterior implica numa maior probabilidade de ocorrer um aumento de casos no ano posterior.

Figura 16 - mostra como as variáveis influenciam a expectativa de aumento anormal de casos de dengue. Cada ponto representa uma observação: a cor indica o valor da variável (azul para baixo, vermelho para alto) e a posição horizontal reflete o impacto no modelo (valores SHAP positivos aumentam a expectativa, negativos reduzem). As variáveis mais importantes estão no topo, indicando maior relevância para o modelo.



Fonte: Elaborado pelo autor.

5.4.6. Atualização de probabilidades com o Teorema de Bayes

Aplicou-se o Teorema de Bayes com base nas métricas do conjunto OOT para atualizar a probabilidade de ocorrência da categoria "Não normal" nos casos em que o modelo a prediz como tal. Considerando as probabilidades a priori das categorias (44,44% para "Normal" e 55,56% para "Não normal") e as métricas de desempenho do modelo OOT, com *Recall* de 0,8458 e taxa de falsos-positivos de 14,79% para a categoria "Não normal," foi obtida uma probabilidade a posteriori de aproximadamente 87,73% para a ocorrência da categoria "Não normal" nos casos classificados como tal pelo modelo, aumentando em aproximadamente 30% a probabilidade de que um município esteja na categoria “Não normal” quando comparado unicamente à frequência histórica.

6. DISCUSSÃO

6.1.Avaliação contextualizada da variabilidade genética de DENV

6.1.1. Potencial impacto da variabilidade genética de DENV

A compreensão dos domínios imunogênicos das proteínas NS1 e envelope é fundamental, dado seu potencial como alvos terapêuticos e de diagnóstico. Estudos anteriores identificaram epítopos na proteína NS1 localizados nas regiões 01-24, 25-38, 21-40, 36-45, 50-65, 57-77, 74-88, 80-89, 103-112, 108-129, 101-130, 121-130, 127-143, 156-175, 166-178, 187-196, 193-209, 206-215, 233-245, 231-255, 264-281, 275-290, 295-304, 315-324, 296-325, 311-330 e 341-351(Bergamaschi *et al.*, 2019; Freire *et al.*, 2017; Hertz *et al.*, 2017; Jiang *et al.*, 2010; Nagar; Savargaonkar; Anvikar, 2020; Rocha *et al.*, 2017). Na proteína de envelope, epítopos relevantes foram identificados nas regiões 02-25, 67-80, 98-109, 201-226, 310-330, 390-409 e 411-428 (Bergamaschi *et al.*, 2019; Nagar; Savargaonkar; Anvikar, 2020). Esses achados ressaltam a complexidade das interações entre o vírus e o sistema imunológico, além de fornecerem subsídios cruciais para o desenvolvimento de estratégias de prevenção e tratamento eficazes contra a dengue. Em virtude da relevância dessas proteínas, a análise das mutações observadas em DENV1 e DENV2, particularmente nas suas respectivas proteínas NS1 e E, revela não

apenas um padrão de variação genética, mas também potenciais implicações na virulência e transmissão do vírus (Chen *et al.*, 2022; Wing *et al.*, 2019). A seguir, serão discutidas as mutações identificadas e suas possíveis consequências biológicas.

Para o DENV1-V, a mutação E:I480V (substituição da isoleucina por valina na posição 480 da proteína E) se destacou como a única diferença estatística significativa em relação às demais regiões. Contudo, esta posição não é descrita como parte dos epítopos imunogênicos conhecidos, e a troca de um aminoácido apolar por outro do mesmo grupo potencialmente não influenciaria a estrutura da proteína E pelo fato de se manterem as interações próximos de um ponto de vista bioquímico. Além disso, a mutação NS1:Y247F (substituição da tirosina por fenilalanina), apesar da substituição também permanecer na mesma categoria bioquímica, uma vez que está localizada próxima a epítopos imunogênicos identificados em regiões como 236-245 (Hertz *et al.*, 2017; Nagar; Savargaonkar; Anvikar, 2020), pode impactar as funções imunológicas da proteína e o reconhecimento por anticorpos.

No DENV2-II, as mutações E:T226I (substituição da treonina por isoleucina), E:T303I (substituição da treonina por isoleucina) e E:S470T (substituição da serina por treonina) foram significativamente mais frequentes no Brasil em comparação com outras regiões, exceto a Europa. A posição 226 é um sítio dentro da região epítopo 201-226 (Nagar; Savargaonkar; Anvikar, 2020), enquanto a posição 303 está localizada no epítopo 310-330 (Nagar; Savargaonkar; Anvikar, 2020), e a posição 470 não é descrita como parte dos epítopos conhecidos. Essas trocas de aminoácidos podem afetar a estabilidade e a conformação da proteína E, influenciando sua capacidade de induzir uma resposta imune robusta. Para a proteína NS1, as mutações NS1:I88V (substituição da isoleucina por valina), NS1:Q175R (substituição da glutamina por arginina) e NS1:R192K (substituição da arginina por lisina) foram mais frequentes no Brasil, indicando uma adaptação local do vírus. A posição 88 não é um epítopo descrito, mas a posição 175 está associada à região 166-178 (Nagar; Savargaonkar; Anvikar, 2020), e a posição 192 também está próxima de regiões reconhecidas por anticorpos (Jiang *et al.*, 2010).

No DENV2-III, a mutação E:S363A (substituição da serina por alanina), embora predominante no Caribe (100%), também se apresentou com frequência significativa no Brasil (73,8%), potencialmente indicando uma emergência independente. Esta mutação embora não caia em uma região definida como epítopo, trata-se da substituição de um aminoácido de um aminoácido polar por um apolar, sugerindo um potencial impacto em sua estabilidade.

Embora essas mutações não sejam necessariamente relevantes por si só, a variabilidade das frequências reforça a ideia da existência de regiões com variabilidade genética e combinações únicas, como evidenciado em estudos que destacam a adaptabilidade do DENV a diferentes hospedeiros e condições ambientais. Isso é relevante ao se considerar que, como exemplo, um pequeno número de mutações pode ocasionar diferença no número de casos: as mutações E1:A226V e E2:L210Q são associadas a adaptabilidade do vírus Chikungunya a um novo vetor, *Aedes albopictus*, sem comprometer a interação do vírus com outro vetor, *Aedes aegypti* (Tsetsarkin; Weaver, 2011). Portanto, o mapeamento de mutações, seja para fins de adaptabilidade ou estabilidade, seja para identificação de epítopos imunogênicos, particularmente nas proteínas NS1 e E, fornece sugestões úteis para o desenvolvimento de estratégias de saúde direcionadas. A proteína NS1, em particular, é alvo de interesse devido a seus epítopos imunológicos identificados em diversas regiões, o que pode auxiliar na formulação de estratégias de diagnóstico e prevenção.

6.1.2. Diversidade genética e surtos

Até a data do presente estudo, o genótipo cosmopolita (DENV2-II) foi relatado como sendo introduzido em uma infecção ocorrida de janeiro a março de 2021 (Amorim *et al.*, 2023) no estado do Acre e em novembro de 2021 no estado de Goiás (Giovanetti *et al.*, 2022b). No caso de Goiás, tratou-se de um indivíduo sem relato de viagens prévias, indicando um caso autóctone. Dessa forma, assumindo que as primeiras infecções com esse genótipo ocorreram ao final de 2021, todas as infecções causadas pelo DENV2 ao longo dos anos anteriores no Brasil têm como provável fonte genótipos diferentes do sorotipo II. Aqui pode-se trabalhar com duas hipóteses para explicar o surto de casos que ocorreu após a introdução do genótipo cosmopolita: (1) embora pertencente ao sorotipo 2, é possível que populações expostas previamente ao sorotipo 2 não estejam adequadamente protegidas contra o DENV2-II; (2) há, ainda, uma vasta quantidade de indivíduos susceptíveis ao sorotipo 2.

Estudos abriram precedente para uma reinfecção por um mesmo sorotipo de DENV (Forshey *et al.*, 2016; Waggoner *et al.*, 2016). Essa hipótese é plausível ao se considerar a distância genética e estrutural entre as proteínas de DENV2-II e DENV2-III. A diferença, como mostrado anteriormente, embora crescente, não é tão grande quando comparada à distância entre sorotipos (Rico-Hesse', 1990). Ainda, é necessário investigar

pois não é clara a existência de uma “linha de corte” que delimite até que porcentagem de diferença possui maior ou menor probabilidade de proporcionar infecções como se fossem de diferentes sorotipos.

Apesar da introdução e do surto de casos de DENV2, a maioria dos casos identificados em 2022 pertencem ao sorotipo DENV1 (Gräf *et al.*, 2023). Por outro lado, há descrições de surtos de casos relacionados à circulação, em 2023 de DENV2-II (Amorim *et al.*, 2023; Ribeiro *et al.*, 2024). Nesse contexto, considera-se que a alternância entre genótipos e sorotipos é um exemplo de um aumento na competência viral de atingir alta viremia em humanos (Amorim *et al.*, 2024). Aliado a isso, não é possível excluir, com os dados atuais, a possibilidade de reinfecções por um mesmo sorotipo, já que houve relatos (Forshey *et al.*, 2016) e o que a nível molecular tem fundamento de estarem envolvidas com a crescente distância genética entre os dois genótipos de DENV2, relatada neste estudo.

6.2. Desenho de *primers* e sequenciamento genômico

6.2.1. Incompatibilidade entre primers e variabilidade genética

A complexidade que envolve a transmissão do DENV permite a ocorrência de genótipos específicos segregados por regiões geográficas. O Brasil, país de dimensões continentais, por si só é passível de possuir uma variabilidade genética que ainda é desconhecida.

Os *mismatches* encontrados virtualmente indicam uma diferença da predominância de mutações regiões específicas. Como é esperado, foram encontradas maiores divergências entre o painel internacional e as sequências do Brasil, quando comparado com os *primers* desenvolvidos utilizando-se unicamente as sequências brasileiras.

Pode até ser que os *primers* funcionem, mas os *primers* com mutações podem causar falsas reversões de mutações (Au *et al.*, 2017; Ulhuq *et al.*, 2023), o que dificulta o acompanhamento da evolução do vírus, não permitindo claramente a distinção da fixação de uma mutação em um genótipo.

6.2.2. Sequenciamento genético

Os resultados indicam um bom desempenho geral dos painéis testados, com alta cobertura genômica e profundidade de sequenciamento suficiente para a maioria das amostras. No entanto, algumas limitações precisam ser consideradas. O desempenho inferior da amostra RJ005 sugere possível degradação do material genético ou eficiência reduzida na etapa de enriquecimento da biblioteca. Além disso, o controle negativo no protocolo de 800 bases apresentou um resultado positivo, levantando questões sobre a especificidade dos *primers* e a possibilidade de contaminação cruzada durante o preparo das amostras.

Os dados mostram que é viável utilizar *primers* para DENV1 com *amplicons* de 800 pares de base sem perdas significativas na qualidade do sequenciamento, o que pode ampliar a estimativa de vida útil desse painel de *primers*.

Contudo, algumas limitações devem ser consideradas na interpretação dos resultados. A representatividade geográfica das amostras foi restrita a duas regiões do Brasil (Amazonas e Rio de Janeiro), o que limita a generalização dos achados para áreas com perfis epidemiológicos e genéticos distintos do DENV. Assim, ainda não é possível determinar claramente a quantidade de *primers* com possíveis *mismatches* entre os painéis utilizados e as sequências encontradas em outras regiões do país. Além disso, a dificuldade na quantificação impediu a análise comparativa com os dados dos painéis de *primers* de (Vogels *et al.*, 2023).

Adicionalmente, a validação inicial foi realizada apenas com amostras de DENV1, o que impossibilita uma avaliação completa do desempenho dos painéis para DENV2. Essa limitação restringe a comparação direta entre os sorotipos e pode afetar a robustez das conclusões sobre a aplicabilidade dos *primers* desenvolvidos. Por fim, a validação não incluiu a comparação entre diferentes faixas de valores de *Cycle Threshold* (CT), o que impede uma análise mais detalhada do desempenho dos painéis em amostras com diferentes cargas virais. Essa limitação pode impactar a avaliação da sensibilidade dos *primers*, especialmente em amostras com baixa concentração de RNA viral.

6.3. Aprimoramento da atualização de painéis de *primers*

6.3.1. Tempo de vida útil dos primers

A escolha do tamanho ideal do *amplicon* em um protocolo de sequenciamento por *amplicon* deve considerar a situação financeira do laboratório e o tempo de envio de reagentes e *primers*. É fundamental encontrar um equilíbrio entre o número de *primers* necessários para lidar com amostras mais degradadas e a opção por *amplicons* maiores, que podem exigir menos *primers* e, consequentemente, oferecer um tempo de vida útil mais prolongado. Para laboratórios com recursos limitados, a opção por *amplicons* maiores pode ser vantajosa, desde que não haja um impacto significativo na qualidade do sequenciamento. Essa abordagem não apenas economiza na quantidade de *primers* utilizados, mas também permite uma vigilância contínua com menor probabilidade de necessidade de troca dos painéis, otimizando assim a eficiência do laboratório.

6.3.2. Hydra: software para acompanhamento de painéis de amplicons

O desenvolvimento de um software capaz de resumir e compilar métricas de qualidade para a avaliação dos arquivos de mapeamento das *reads* constitui um modo de simplificar o trabalho de analistas nas fases após a montagem dos genomas, servindo como uma automação para checagens minuciosas e laboriosas de outra forma. Embora existam outros softwares capazes de permitir a análise detalhada de dados oriundos de montagem de genomas virais por referência, como o IGV (Thorvaldsdóttir; Robinson; Mesirov, 2013) e o BamView (Carver *et al.*, 2013) a implementação do software Hydra nasce da necessidade de uma análise direcionada, integrada e específica para o acompanhamento do funcionamento de *primers* por longos períodos de vigilância genômica.

A visualização dos dados do mapeamento das *reads* em relação a uma referência, por meio do gráfico de profundidade do genoma, oferece uma avaliação direta da cobertura de sequenciamento. Isso permite identificar áreas com baixa cobertura, indicando possíveis falhas no processo de sequenciamento. Além disso, o sistema de métricas em formato de semáforo fornece uma avaliação quantitativa de diversos aspectos dos dados genômicos. Isso inclui a identificação de *amplicons* com cobertura média abaixo de 50, a detecção de grandes deleções no genoma e a variação da estabilidade da

profundidade. Tais informações proporcionam uma compreensão abrangente da qualidade dos dados.

Os scores de variabilidade e estabilidade do mapeamento das *reads* contra o genoma de referência proporcionam uma análise mais aprofundada da qualidade dos dados de sequenciamento, permitindo uma avaliação mais precisa da biologia e epidemiologia, por exemplo, dos sorotipos de DENV. Essas métricas são fundamentais para identificar potenciais problemas durante o sequenciamento, como baixa cobertura ou regiões deletadas, garantindo a confiabilidade das análises realizadas e a interpretação correta dos resultados obtidos.

Não obstante inicialmente concebido para análise de genomas virais, a aplicabilidade do método não se restringe a esse contexto, estendendo-se à análise de fragmentos de DNA de tamanhos variados, desde que sua montagem tenha uma referência. Até o momento, não foi identificada uma ferramenta disponível capaz de realizar a avaliação da presença de variações intra-hospedeiro em regiões de *primers* após sequenciamento por *amplicon*, significando possíveis *primers* posicionados em regiões de mutações e capazes de gerar resultados falso-negativos de mutações sinapomórficas de linhagens (Au *et al.*, 2017), e apresentar os resultados de forma sucinta por meio de dados tabulares. O Hydra se destaca, ainda, pois sua implementação em um repositório de imagens Docker permite sua utilização tanto como uma ferramenta web quanto localmente. Além disso, o *software* é facilmente integrável com quaisquer tecnologias para montagem de genomas virais, utilizando apenas arquivos intermediários do processamento ou arquivos de entrada já utilizados pelas *pipelines*.

6.4. Arquitetura e modelagem de dados para vigilância

O Brasil tem se esforçado para integrar dados públicos com o intuito de fortalecer a pesquisa, formular políticas mais eficazes e aprimorar os serviços prestados à população. Diversas iniciativas foram implementadas para conectar informações sociais e de saúde, aprimorar a infraestrutura digital e expandir o uso de tecnologias móveis. Essas ações visam disponibilizar dados abrangentes e de alta qualidade, promovendo melhores resultados e uma gestão mais eficiente dos recursos disponíveis (Fillinger *et al.*, 2019; Sheikh *et al.*, 2021).

Entretanto, a integração desses dados enfrenta desafios significativos (Parimala *et al.*, 2017). Esse processo exige a harmonização de formatos e a superação de barreiras

tecnológicas, incluindo problemas de armazenamento (Chen *et al.*, 2013) e questões de segurança dos dados (Wu *et al.*, 2014). A centralização de grandes volumes de informações requer soluções robustas para garantir tanto a proteção contra acessos não autorizados quanto a integridade dos dados. Além disso, a análise integrada pode resultar em correlações espúrias, o que torna crucial o envolvimento de especialistas para a formulação de hipóteses plausíveis e a prevenção de conclusões equivocadas. A falta de um entendimento profundo das interações entre variáveis pode levar a interpretações inadequadas, ressaltando a importância da expertise na análise correta das informações, o que, por sua vez, melhora a eficácia dos sistemas de saúde (Smith, 2020).

Portanto, apesar dos avanços nas tecnologias e na disponibilidade de dados, a integração eficaz desses dados ainda enfrenta desafios críticos que devem ser abordados para otimizar a pesquisa e a resposta a crises de saúde pública.

6.4.1. Interoperabilidade e padronização

As múltiplas fontes de dados coletadas por diferentes órgãos no Brasil frequentemente carecem de uma comunicação eficaz, fator essencial para a integração eficiente dessas informações (De Soarez *et al.*, 2022). Iniciativas bem-sucedidas, como o Centro de Integração de Dados e Conhecimentos para Saúde (CIDACS), fundado em 2016, têm contribuído significativamente para a integração de dados de saúde pública (Barreto *et al.*, 2019). Este estudo, voltado para a predição de surtos e epidemias, busca tornar as bases públicas disponíveis no Brasil mais integradas e acessíveis. Embora essas bases sejam de livre acesso, muitas vezes requerem conhecimentos técnicos ou infraestrutura apropriada para acessar grandes volumes de dados históricos.

Para mitigar a falta de comunicação entre as bases de dados, pode-se adotar um nível comum de granularidade entre várias fontes, como município, mês e ano (MMA). Essa abordagem permite a agregação de informações em níveis individuais sem perdas significativas, facilitando a concatenação de diferentes bases e a avaliação de surtos de múltiplas doenças, sem a necessidade de desanonimização dos dados ou a implementação de regras complexas de *linkage* entre as bases. A granularidade MMA é compatível com o objetivo de identificar surtos de casos em nível municipal, utilizando uma definição temporal de um mês.

Adicionalmente, inconsistências nas bases de dados podem surgir devido a diferentes versões dos bancos de dados públicos e atualizações nas nomenclaturas por

organizações externas. Um exemplo é a alteração na classificação da severidade do agravo de Dengue pela OMS em 2009 (WHO, 2009b), que impactou a base de dados do SINAN. Essas mudanças podem resultar em discrepâncias e dificultar a comparação e análise de dados ao longo do tempo.

Além das inconsistências relacionadas a versionamentos, a especificação inadequada nos dicionários de dados também representa um desafio. Muitos dicionários de dados carecem de descrições detalhadas sobre as colunas (Saldanha; Bastos; Barcellos, 2019), como os disponíveis para a base CNES ST (Estabelecimentos). Isso leva os usuários a buscarem informações adicionais em fontes variadas, que muitas vezes não são claras ou não oferecem orientações adequadas sobre onde encontrar as informações compiladas.

6.4.2. Acesso e transparência

A dificuldade de acesso a dados, mesmo que públicos e integrados na plataforma DATASUS, compromete a tomada de decisões por analistas e cientistas de dados que carecem de capacidade computacional para compilar séries históricas extensas, essenciais para identificar padrões associados a surtos. Esses padrões são fundamentais para previsões e predições de surtos e epidemias.

Além disso, o formato DBC utilizado pelo DataSUS impõe várias limitações, como o acesso e a manipulação dos dados, devido à sua natureza binária e à necessidade de decodificação de variáveis numéricas em categorias frequentemente mal documentadas. Embora existam alternativas eficazes, como os pacotes PySUS (Coelho *et al.*, 2021) e microdatasus (Saldanha; Bastos; Barcellos, 2019), estas requerem conhecimentos em linguagens de programação e recursos computacionais adequados para converter arquivos DBC, além da instalação de dependências em nível de servidor (Linux), o que pode não ser viável em determinados ambientes, dadas as especificações dos sistemas operacionais ou as limitações dos usuários.

Dessa forma, a disponibilização de uma base de dados já curada, decodificada e prontamente acessível pode acelerar o acesso às informações e a tomada de decisões. Essa abordagem elimina a necessidade de os usuários enfrentarem a etapa de extração de dados e as complexidades associadas ao uso de linguagens de programação específicas, ampliando o público-alvo que pode se beneficiar desses dados. Este trabalho visa superar essas limitações ao fornecer um repositório que facilita o acesso e a utilização de bases

de dados públicas. A integração eficiente dos dados permitirá que analistas identifiquem padrões de disseminação de doenças de maneira mais precisa, sem a necessidade de lidar com dados individuais sensíveis ou não anonimizados. O objetivo é possibilitar análises mais robustas e predições mais confiáveis, contribuindo para uma resposta mais eficaz a crises de saúde pública no Brasil.

Iniciativas semelhantes, como o ÆSOP (https://cidacs.bahia.fiocruz.br/plataforma/aesop-predicao-de-pandemias/), que se dedica à predição de pandemias, e o Mosqlimate (https://mosqlimate.org/), focado em modelagem climática e saúde, também trabalham na integração de dados para melhorar a resposta a crises de saúde pública. Entretanto, o foco deste repositório é complementar essas iniciativas, facilitando o acesso a bases públicas e permitindo a inclusão modular de novas bases ao longo do tempo. O repositório apresentado pode facilitar o uso de outras bases de dados ainda não abordadas por iniciativas similares, promovendo uma gestão mais eficiente dos recursos de saúde pública no Brasil.

6.4.3. Segurança e privacidade dos dados

As medidas de segurança implementadas no sistema, incluindo o uso de *prepared statements*, validação de entradas e limitação de taxa, mostraram-se eficazes na mitigação de ameaças críticas, como SQL *Injection* e ataques DDoS. Essas práticas, amplamente reconhecidas na literatura, foram essenciais para garantir a integridade e a disponibilidade dos dados. Estudos anteriores, demonstraram que os *prepared statements* são uma das abordagens mais seguras para prevenir ataques de injeção SQL, corroborando os resultados obtidos nesta implementação (Nasereddin *et al.*, 2023). Da mesma forma, os mecanismos de limitação de taxa empregados para reduzir os impactos de ataques DDoS têm suporte em pesquisas (Zargar *et al.*, 2013), que enfatizam a importância de estratégias proativas de defesa. A adoção dessas práticas fortalece a eficácia das medidas preventivas na manutenção da segurança e disponibilidade de sistemas baseados em dados.

6.4.4. Capacidade de Fluxo de Acesso de Indivíduos

A gestão eficaz de grandes volumes de dados é um desafio considerável, especialmente no que se refere à disponibilização ágil para diversos usuários. Durante as etapas de modelagem e armazenamento, constatou-se que o banco de dados original

ocupava 35 GB. No entanto, a conversão dos arquivos para os formatos JSON e CSV resultou em tamanhos que chegaram a 200 GB, o que impossibilitou tanto a transferência quanto o armazenamento em servidores. Para solucionar esse problema, foi implementada uma estratégia de compressão, que se mostrou extremamente eficiente, reduzindo o tamanho dos arquivos de 200 GB para menos de 2 GB. Essa abordagem, aliada à criação de um sistema de cache, otimizou significativamente o tempo de resposta, permitindo o acesso rápido e eficiente aos dados por múltiplos usuários. Essa solução não apenas facilita a análise de grandes volumes de dados, mas também assegura que a infraestrutura possa atender à crescente demanda por acessibilidade e interação com conjuntos de dados extensos.

6.4.5. Previsão de surtos

Estudos têm demonstrado que fatores climáticos e de mudanças climáticas estão entre os mais frequentemente empregados em modelos preditivos. Em uma revisão de 64 estudos, incluindo 99 modelos, 94,7% utilizaram dados climáticos de agências governamentais, e 77,8% incluíram informações sobre mudanças climáticas. No entanto, apenas 59,6% dos modelos ajustaram para atrasos no tempo de relato. Todos os modelos revisados incluíram preditores climáticos, e 70,7% utilizaram exclusivamente essas variáveis. Em outros casos, fatores climáticos foram combinados com mudanças climáticas (13,4%), fatores demográficos (5,2%) ou vetoriais (6,3%) (Leung *et al.*, 2023). Outro trabalho elenca a precipitação, a temperatura e a umidade como principais preditores utilizados para modelagem da previsão de casos (Sylvestre *et al.*, 2022).

Em contrapartida, técnicas de *machine learning*, empregadas em 39,4% dos modelos analisados, permitem captar relações complexas e não lineares entre as variáveis. Entre os modelos mais utilizados estão *Random Forest* (15,4%), redes neurais (23,1%) e modelos ensemble (10,3%). Nos modelos estatísticos (60,6%), a regressão linear (18,3%), a regressão de Poisson (18,3%) e os modelos autorregressivos (26,7%) foram os mais frequentes. Cerca de 20,2% dos modelos não relataram nenhum tipo de validação, e apenas 5,2% utilizaram validação externa (Leung *et al.*, 2023). Os modelos com melhor desempenho foram as redes neurais e as árvores de decisão (52%), seguidos pela máquina de vetor de suporte (17%) (Sylvestre *et al.*, 2022).

Além disso, os modelos clássicos de predição de séries temporais, como o modelo autorregressivo (AR), o modelo de média móvel (MA), sua combinação (ARMA), bem

como suas variações (ARIMA e SARIMA), são comumente utilizados para explorar a correlação interna das séries temporais. Essas abordagens, no entanto, utilizam combinações lineares dos cálculos de erros do próprio modelo juntamente de correlações entre o valor previsto e valores passados (Chen; Moraga, 2024).

Embora as técnicas de *machine learning* capturem interações complexas, elas geralmente não utilizam explicitamente métodos clássicos de séries temporais. A combinação de ambas as abordagens pode aumentar a robustez dos modelos preditivos. O modelo de *Gradient Boosting* (Friedman, 2001) foi selecionado devido à sua capacidade de incorporar erros durante o processo de modelagem, uma característica semelhante à dos modelos MA. Adicionalmente, a introdução de *Lags* permite capturar dinâmicas semelhantes às exploradas pelos modelos AR, e características sazonais podem ser derivadas de forma semelhante a um modelo SARIMA.

Não obstante capazes de gerar relações complexas e não lineares entre as variáveis preditoras, não explicitamente utilizam de métodos clássicos de séries temporais. Logo, a incorporação de ambas as metodologias possui grande potencial. Foi escolhido o modelo de *Gradient Boosting* pois ele é capaz de incorporar implicitamente os erros cometidos ao longo da criação do modelo, mimetizando o que é aplicado em modelos MA. Ao mesmo tempo, a criação de *Lags* ou atrasos na série temporal como forma de capturar a ideia por trás do modelo AR. Ainda, podemos, de forma similar, derivar as características sazonais presentes de um modelo SARIMA.

Embora tenha sido aplicada uma complexa metodologia para capturar fatores relevantes para a predição de surtos de dengue, dentro dos parâmetros definidos pelo Ministério da Saúde, os resultados não foram satisfatórios, refletindo a complexidade do fenômeno biológico em questão. O aumento no número de casos eleva as chances de o vírus acumular mutações e, à medida que mais mutações se acumulam, aumenta a probabilidade de ocorrerem alterações em regiões de *primer*. A análise de predição de surtos visa antecipar os picos de casos, gerando uma maior demanda por atenção na substituição de *primers* específicos. Os resultados indicam que o modelo possui alta confiança ao prever casos como "Não normal". A probabilidade a posteriori de 87,73% sugere que, ao incorporar as métricas do conjunto OOT, as previsões para "Não normal" são robustas, com confiança na classificação.

Esse valor elevado na probabilidade a posteriori reflete que o modelo tem um bom desempenho em capturar a ocorrência de casos "Não normal", o que é crucial para aplicações de monitoramento, onde antecipar condições atípicas pode orientar decisões

preventivas e corretivas. Além disso, a metodologia utilizada destaca a importância de incorporar lógicas bayesianas em modelos preditivos, ajustando as probabilidades com base na precisão específica do modelo em novos dados (OOT), que serve como uma proxy da generalização do modelo para dados reais.

A análise dos valores SHAP oferece uma visão ainda mais profunda sobre a contribuição das variáveis no modelo. O gráfico de barras (figura 15) revela que variáveis históricas de casos, como a média de casos nos últimos 3 anos com *Lags* específicos, são determinantes para a previsão de surtos de dengue. Essa tendência reforça a importância de incluir informações temporais detalhadas para capturar padrões sazonais e tendências ao longo do tempo. Variáveis relacionadas à evolução clínica e à classificação dos casos também desempenham um papel importante, sugerindo que características epidemiológicas complementam os dados históricos na construção de modelos robustos. Esses achados indicam a necessidade de manter registros detalhados e atualizados para melhorar a acurácia de previsões futuras.

A análise do presenta na figura 16, por sua vez, revela que a predominância de variáveis com relação inversa ao aumento anormal de casos sugere que fatores como a redução na média de casos e na contagem de casos confirmados nos períodos anteriores podem estar ligados a cenários propensos a surtos subsequentes. Esse padrão reforça a necessidade de atenção em contextos de redução aparente na atividade de dengue, já que esses cenários podem mascarar um aumento iminente. A porcentagem de casos descartados de dengue – *Lag* 9 como exceção é particularmente interessante, indicando que um aumento na proporção de casos descartados pode ser um sinal de alerta para surtos futuros. Esse comportamento peculiar pode estar associado a possíveis subnotificações ou erros no diagnóstico que, em conjunto com outros fatores, contribuem para a dinâmica dos surtos. Essa interpretação reforça o papel dos modelos preditivos na identificação precoce de padrões epidemiológicos.

7. CONCLUSÃO

Este estudo evidenciou que a diversidade genética do vírus da dengue no Brasil potencialmente impacta diretamente a eficácia dos métodos de vigilância epidemiológica, destacando a importância de considerar a variabilidade genética regional no desenvolvimento de painéis de sequenciamento para garantir precisão e confiabilidade. A análise temporal realizada sugeriu que a introdução do genótipo cosmopolita (DENV2-

II) contribuiu para o aumento de casos devido à suscetibilidade populacional, embora a reinfecção pelo mesmo sorotipo ainda requeira investigação adicional. O *software* Hydra, desenvolvido para monitorar a qualidade do sequenciamento, demonstrou eficácia ao avaliar a cobertura e detectar variações nas regiões de *primer*, mostrando-se essencial para a vigilância genômica do DENV além de ser aplicável para outros vírus. Ademais, a criação de uma base de dados modular, integrando informações epidemiológicas, sociodemográficas, ambientais e econômicas, desdobra-se como um passo fundamental para a vigilância e o controle da dengue, apoiando decisões estratégicas. Com suporte nessa infraestrutura, o modelo desenvolvido apresentou alta precisão na identificação de casos atípicos de dengue, mostrando-se promissor para o monitoramento de surtos e antecipação de demandas de atualização de *primers*. Contudo, uma abordagem de classificação refinada pode aprimorar ainda mais a generalização dos resultados.

REFERÊNCIAS

ALCON, S. et al. Enzyme-Linked Immunosorbent Assay Specific to Dengue Virus Type 1 Nonstructural Protein NS1 Reveals Circulation of the Antigen in the Blood during the Acute Phase of Disease in Patients Experiencing Primary or Secondary Infections. **Journal of Clinical Microbiology**, v. 40, n. 2, p. 376–381, 2002.

ALIDJINOU, E. K. et al. Prospective Evaluation of a Commercial Dengue NS1 Antigen Rapid Diagnostic Test in New Caledonia. **Microorganisms**, v. 10, n. 2, 2022.

ALLICOCK, O. M. et al. Phylogeography and Population Dynamics of Dengue Viruses in the Americas. **Molecular Biology and Evolution**, v. 29, n. 6, p. 1533–1543, 2012.

AMORIM, M. T. et al. Detection of a Multiple Circulation Event of Dengue Virus 2 Strains in the Northern Region of Brazil. **Tropical Medicine and Infectious Disease**, v. 9, n. 1, 2024.

AMORIM, M. T. et al. Emergence of a New Strain of DENV-2 in South America: Introduction of the Cosmopolitan Genotype through the Brazilian-Peruvian Border. **Tropical Medicine and Infectious Disease**, v. 8, n. 6, 2023.

ARANA, C. et al. A Short Plus Long-Amplicon Based Sequencing Approach Improves Genomic Coverage and Variant Detection in the SARS-CoV-2 Genome. **PLoS ONE**, v. 17, n. 1, 2022.

ARYATI, A. et al. Performance of Commercial Dengue NS1 ELISA and Molecular Analysis of NS1 Gene of Dengue Viruses Obtained During Surveillance in Indonesia. **BMC Infectious Diseases**, v.13, n. 611, 2013.

AU, C. H. et al. BAMClipper: Removing Primers from Alignments to Minimize False-Negative Mutations in Amplicon Next-Generation Sequencing. **Scientific Reports**, v. 7, n. 1, 2017.

AYUKEPI AYUKEKBONG, J. et al. Value of Routine Dengue Diagnosis in Endemic Countries. **World Journal of Virology**, v. 6, n. 1, p. 16, 2017.

BARRETO, M. L. et al. The Centre for Data and Knowledge Integration for Health (CIDACS): Linking Health and Social Data in Brazil. **International Journal of Population Data Science**, v. 4, n. 2, 2019.

BEHJATI, S.; TARPEY, P. S. What is Next Generation Sequencing? **Archives of Disease in Childhood: Education and Practice Edition**, v. 98, n. 6, p. 236–238, 2013.

BERGAMASCHI, G. et al. Computational Analysis of Dengue Virus Envelope Protein (E) Reveals an Epitope with Flavivirus Immunodiagnostic Potential in Peptide Microarrays. **International Journal of Molecular Sciences**, v. 20, n. 8, 2019.

BHATT, S. et al. The Global Distribution and Burden of Dengue. **Nature**, v. 496, n. 7446, p. 504–507, 2013.

BRONNER, I. F.; QUAIL, M. A. Best Practices for Illumina Library Preparation. **Current Protocols in Human Genetics**, v. 102, n. 1, p. 1–48, 2019.

BROWN, W. C. et al. Extended Surface for Membrane Association in Zika Virus NS1 Structure. **Nature Structural and Molecular Biology**, v. 23, n. 9, p. 865–867, 2016.

CARVER, T. et al. BamView: Visualizing and Interpretation of Next-Generation Sequencing Read Alignments. **Briefings in Bioinformatics**, v. 14, n. 2, p. 203–212, 2013.

CHANG, F.; LI, M. M. Clinical Application of Amplicon-Based Next-Generation Sequencing in Cancer. **Cancer Genetics**, v. 206, n. 12, p. 413–419, 2013.

CHEN, J. et al. Big Data Challenge: A Data Management Perspective. **Frontiers of Computer Science**, v. 7, n. 2, p. 157–164, 2013.

CHEN, L. et al. Neighboring Mutation-Mediated Enhancement of Dengue Virus Infectivity and Spread. **EMBO Reports**, v. 23, n. 11, 2022.

CHEN, S. et al. Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor. **Bioinformatics**, v. 34, n. 17, p. i884–i890, 2018.

CHEN, X.; MORAGA, P. Assessing Dengue Forecasting Methods: A Comparative Study of Statistical Models and Machine Learning Techniques in Rio de Janeiro, Brazil. **medRxiv**, 2024.

CHIARA, M. et al. Next Generation Sequencing of SARS-CoV-2 Genomes: Challenges, Applications and Opportunities. **Briefings in Bioinformatics**, v. 22, n. 2, p. 616–630, 2021.

CLINCY, V.; SHAHRIAR, H. Web Application Firewall: Network Security Models and Configuration. **International Computer Software and Applicaions Conference**, p. 835-836, 2018.

COCK, P. J. A. et al. The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants. **Nucleic Acids Research**, v. 38, n. 6, p. 1767–1771, 2009.

COELHO DE SOAREZ, P. et al. Electronic Health Records in Brazil: Prospects and Technological Challenges. **Frontiers**, v. 10, 2022.

COELHO, F. C. et al. AlertaDengue/PySUS: Vaccine. **Zenodo**, 2021. Disponível em: https://doi.org/10.5281/zenodo.4883502.

DELRIEU, M. et al. Temperature and Transmission of Chikungunya, Dengue, and Zika Viruses: A Systematic Review of Experimental Studies on Aedes aegypti and Aedes albopictus. **Current Research in Parasitology and Vector-Borne Diseases**, v. 3, 2023.

DUONG, V. et al. Asymptomatic Humans Transmit Dengue Virus to Mosquitoes. **Proceedings of the National Academy of Sciences**, v. 112, n. 47, p. 14688–14693, 2015.

FAHIMI, H. et al. Dengue Viruses and Promising Envelope Protein Domain III-Based Vaccines. **Applied Microbiology and Biotechnology**, v. 102, n. 7, p. 2977–2996, 2018.

FILLINGER, S. et al. Challenges of Big Data Integration in the Life Sciences. **Analytical and Bioanalytical Chemistry**, v.411, 2019.

FORSHEY, B. M. et al. Incomplete Protection Against Dengue Virus Type 2 Re-infection in Peru. **PLoS Neglected Tropical Diseases**, v. 10, n. 2, 2016.

FREIRE, M. C. L. C. et al. Mapping Putative B-Cell Zika Virus NS1 Epitopes Provides Molecular Basis for Anti-NS1 Antibody Discrimination Between Zika and Dengue Viruses. **ACS Omega**, v. 2, n. 7, p. 3913–3920, 2017.

FREITAS SALDANHA, R. et al. Microdatasus: A Package for Downloading and Preprocessing Microdata from Brazilian Health Informatics Department (DATASUS). **Cadernos de Saúde Pública**, v. 35, n. 9, 2019.

GIOVANETTI, M. et al. Emergence of Dengue Virus Serotype 2 Cosmopolitan Genotype, Brazil. **Emerging Infectious Diseases**, v. 28, n. 8, p. 1725–1727, 2022.

GRÄF, T. et al. Multiple Introductions and Country-Wide Spread of DENV-2 Genotype II (Cosmopolitan) in Brazil. **Virus Evolution**, v. 9, n. 2, 2023.

GRUBAUGH, N. D. et al. An Amplicon-Based Sequencing Framework for Accurately Measuring Intrahost Virus Diversity Using PrimalSeq and iVar. **Genome Biology**, v. 20, n. 1, 2019.

GUTSCHE, I. et al. Secreted Dengue Virus Nonstructural Protein NS1 is an Atypical Barrel-Shaped High-Density Lipoprotein. **Proceedings of the National Academy of Sciences**, v. 108, n. 19, p. 8003–8008, 2011.

GUZMAN, M. G. et al. Dengue: A Continuing Global Threat. **Nature Reviews Microbiology**, v. 8, n. 12, p. S7–S16, 2010.

GUZMAN, M. G. et al. Multi-Country Evaluation of the Sensitivity and Specificity of Two Commercially-Available NS1 ELISA Assays for Dengue Diagnosis. **PLoS Neglected Tropical Diseases**, v. 4, n. 8, 2010.

HERMANN, L. L. et al. Evaluation of a Dengue NS1 Antigen Detection Assay Sensitivity and Specificity for the Diagnosis of Acute Dengue Virus Infection. **PLoS Neglected Tropical Diseases**, v. 8, n. 10, 2014.

HERTZ, T. et al. Antibody Epitopes Identified in Critical Regions of Dengue Virus Nonstructural 1 Protein in Mouse Vaccination and Natural Human Infections. **The Journal of Immunology**, v. 198, n. 10, p. 4025–4035, 2017.

HOLMES, E. C.; TWIDDY, S. S. The Origin, Emergence and Evolutionary Genetics of Dengue Virus. **Infection, Genetics and Evolution**, v. 3, n. 1, p. 19–28, 2003.

HOPP, M. J.; FOLEY, J. A. Global-Scale Relationships Between Climate and the Dengue Fever Vector, Aedes aegypti. **Climatic Change**, v.48, 2001.

HUTCHISON, C. A. DNA Sequencing: Bench to Bedside and Beyond. **Nucleic Acids Research**, v. 35, n. 18, p. 6227–6237, 2007.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

JIANG, L. et al. Selection and Identification of B-Cell Epitope on NS1 Protein of Dengue Virus Type 2. **Virus Research**, v. 150, n. 1–2, p. 49–55, 2010.

KATOH, K.; STANDLEY, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. **Molecular Biology and Evolution**, v. 30, n. 4, p. 772–780, 2013.

KENT, W. J. et al. The Human Genome Browser at UCSC. **Genome Research**, v. 12, n. 6, p. 996–1006, 2002.

KINGSMORE, K. M. et al. An Introduction to Machine Learning and Analysis of Its Use in Rheumatic Diseases. **Nature Reviews Rheumatology**, v. 17, n. 12, p. 717–731, 2021.

KOSKELA VON SYDOW, A. et al. Comparison of SARS-CoV-2 Whole Genome Sequencing Using Tiled Amplicon Enrichment and Bait Hybridization. **Scientific Reports**, v. 13, n. 1, 2023.

KREATSOULAS, C.; SUBRAMANIAN, S. V. Machine Learning in Social Epidemiology: Learning from Experience. **SSM - Population Health**, v. 4, p. 270–276, 2018.

LAMBRECHTS, L. et al. Consequences of the Expanding Global Distribution of Aedes albopictus for Dengue Virus Transmission. **PLoS Neglected Tropical Diseases**, v. 4, n. 5, 2010.

LEITMEYER, K. C. et al. Dengue Virus Structural Differences That Correlate with Pathogenesis. **Journal of Virology**, v. 73, n. 6, p. 4738–4747, 1999.

LEUNG, X. Y. et al. A Systematic Review of Dengue Outbreak Prediction Models: Current Scenario and Future Directions. **PLoS Neglected Tropical Diseases**, v. 17, n. 2, 2023.

LI, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. **arXiv**, 2013.

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

LIMA, M. R. Q. et al. Comparison of Three Commercially Available Dengue NS1 Antigen Capture Assays for Acute Diagnosis of Dengue in Brazil. **PLoS Neglected Tropical Diseases**, v. 4, n. 7, 2010.

LOH, W. Y. Classification and Regression Trees. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 1, n. 1, p. 14–23, 2011.

MALDONADO, S. et al. Out-of-Time Cross-Validation Strategies for Classification in the Presence of Dataset Shift. **Applied Intelligence**, v. 52, n. 5, p. 5770–5783, 2022.

MEDAR, R. et al. Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning. **International Conference on Computing, Communication, Control and Automation**, 2017.

METZKER, M. L. Emerging Technologies in DNA Sequencing. **Genome Research**, v. 15, n. 12, p. 1767–1776, 2005.

METZKER, M. L. Sequencing Technologies - The Next Generation. **Nature Reviews Genetics**, v. 11, n. 1, p. 31–46, 2010.

MINISTÉRIO DA SAÚDE (BRASIL). **Plano de Contingência para Resposta às Emergências em Saúde Pública por Dengue, Chikungunya e Zika**. Brasília: Ministério da Saúde, 2023.

MODIS, Y. et al. Structure of the Dengue Virus Envelope Protein After Membrane Fusion. **Nature**, v. 427, n. 6972, p. 313–319, 2004.

MODIS, Y. et al. Variable Surface Epitopes in the Crystal Structure of Dengue Virus Type 3 Envelope Glycoprotein. **Journal of Virology**, v. 79, n. 2, p. 1223–1231, 2005.

MOUSSON, L. et al. Phylogeography of Aedes (Stegomyia) aegypti (L.) and Aedes (Stegomyia) albopictus (Skuse) (Diptera: Culicidae) Based on Mitochondrial DNA Variations. **Genetical Research**, v. 86, n. 1, p. 1–11, 2005.

MULLER, D. A. et al. Clinical and Laboratory Diagnosis of Dengue Virus Infection. **Journal of Infectious Diseases**, v. 215, p. S89–S95, 2017.

NAEEM, M. et al. Comparative Analysis of Machine Learning Approaches to Analyze and Predict the COVID-19 Outbreak. **PeerJ Computer Science**, v. 7, 2021.

NAGAR, P. K. et al. Detection of Dengue Virus-Specific IgM and IgG Antibodies Through Peptide Sequences of Envelope and NS1 Proteins for Serological Identification. **Journal of Immunology Research**, v. 2020, 2020.

NASEREDDIN, M. et al. A Systematic Review of Detection and Prevention Techniques of SQL Injection Attacks. **Information Security Journal**, v. 32, n. 3, p. 198–216, 2023.

PAPAGEORGIOU, S. N. On Correlation Coefficients and Their Interpretation. **Journal of Orthodontics**, v. 49, n. 3, p. 347–351, 2022.

PARIMALA, K. et al. Challenges and Opportunities with Big Data. **International Journal of Scientific Research in Computer Science and Engineering**, v. 5, n. 5, p. 16–20, 2017.

PEARSON, W. R.; LIPMAN, D. J. Improved Tools for Biological Sequence Comparison. **Proceedings of the National Academy of Sciences**, v. 85, n. 8, p. 2444–2448, 1988.

PHADUNGSOMBAT, J. et al. Unraveling Dengue Virus Diversity in Asia: An Epidemiological Study Through Genetic Sequences and Phylogenetic Analysis. **Viruses**, v. 15, n. 7, 2023.

POWERS, D. M. W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37–63, 2011.

QUICK, J. et al. Multiplex PCR Method for MinION and Illumina Sequencing of Zika and Other Virus Genomes Directly from Clinical Samples. **Nature Protocols**, v. 12, n. 6, p. 1261–1266, 2017.

RIBEIRO, J. R. et al. DENV-2 Outbreak Associated With Cosmopolitan Genotype Emergence in Western Brazilian Amazon. **Bioinformatics and Biology Insights**, v. 18, 2024.

RICO-HESSE, R. Molecular Evolution and Distribution of Dengue Viruses Type 1 and 2 in Nature. **Virology**, v. 194, n. 2, p. 479–493, 1993.

ROCHA, L. B. et al. Epitope Sequences in Dengue Virus NS1 Protein Identified by Monoclonal Antibodies. **Antibodies**, v. 6, n. 4, 2017.

SANDERSON, T.; BARRETT, J. C. Variation at Spike Position 142 in SARS-CoV-2 Delta Genomes is a Technical Artifact Caused by Dropout of a Sequencing Amplicon. **Wellcome Open Research**, v. 6, 2021.

SEN, P. C. et al. Supervised Classification Algorithms in Machine Learning: A Survey and Review. **Advances in Intelligent Systems and Computing**, v. 1224, p. 3–18, 2020.

SHEIKH, A. et al. Health Information Technology and Digital Innovation for National Learning Health and Care Systems. **The Lancet Digital Health**, v. 3, n. 6, p. e383–e396, 2021.

SILVA, A. F. et al. ViralFlow v1.0 - A Computational Workflow for Streamlining Viral Genomic Surveillance. **NAR Genomics and Bioinformatics**, v. 6, n. 2, 2024.

SIMMONS, C. P. et al. Dengue. **New England Journal of Medicine**, v. 366, n. 15, p. 1423–1432, 2012.

SLIUSARENKO, T.; FILATOV, V. Relational vs Non-Relational Databases. **Grail of Science**, n. 23, p. 269–271, 2023.

SMITH, G. The Paradox of Big Data. **SN Applied Sciences**, v. 2, n. 6, 2020.

SU, W. et al. A Serotype-Specific and Multiplex PCR Method for Whole-Genome Sequencing of Dengue Virus Directly from Clinical Samples. **Microbiology Spectrum**, v. 10, n. 5, 2022.

SYLVESTRE, E. et al. Data-Driven Methods for Dengue Prediction and Surveillance Using Real-World and Big Data: A Systematic Review. **PLoS Neglected Tropical Diseases**, v. 16, n. 1, 2022.

TAGHAVI ZARGAR, S. et al. A Survey of Defense Mechanisms Against Distributed Denial of Service (DDoS) Flooding Attacks. **IEEE Communications Surveys & Tutorials**, v. 15, n. 4, p. 2046-2069, 2013.

THERMOFISHER SCIENTIFIC. DNA sequencing by capillary electrophoresis chemistry guide. **Waltham: Thermo Fisher Scientific**, 2009.

THORVALDSDÓTTIR, H. et al. Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. **Briefings in Bioinformatics**, v. 14, n. 2, p. 178–192, 2013.

TRACZ, P. M.; PLECHAWSKA-WÓJCIK, M. Comparative Analysis of the Performance of Selected Database Management System. **Journal of Computer Science Institute**, v. 31, 2024

TSETSARKIN, K. A.; WEAVER, S. C. Sequential Adaptive Mutations Enhance Efficient Vector Switching by Chikungunya Virus and Its Epidemic Emergence. **PLoS Pathogens**, v. 7, n. 12, 2011.

TWIDDY, S. S. et al. Phylogenetic Relationships and Differential Selection Pressures Among Genotypes of Dengue-2 Virus. **Virology**, v. 298, n. 1, p. 63–72, 2002.

ULHUQ, F. R. et al. Analysis of the ARTIC V4 and V4.1 SARS-CoV-2 Primers and Their Impact on the Detection of Omicron BA.1 and BA.2 Lineage-Defining Mutations. **Microbial Genomics**, v. 9, n. 4, 2023.

VOGELS, C. B. F. et al. DengueSeq: A Pan-Serotype Whole Genome Amplicon Sequencing Protocol for Dengue Virus. **medRxiv**, 2023.

WAGGONER, J. J. et al. Homotypic Dengue Virus Reinfections in Nicaraguan Children. **Journal of Infectious Diseases**, v. 214, n. 7, p. 986–993, 2016.

WEI, K.; LI, Y. Global Evolutionary History and Spatio-Temporal Dynamics of Dengue Virus Type 2. **Scientific Reports**, v. 7, 2017.

WORLD HEALTH ORGANIZATION (WHO). **Dengue Guidelines for Diagnosis, Treatment, Prevention and Control**. Geneva: WHO, 2009.

WIENS, J.; SHENOY, E. S. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. **Clinical Infectious Diseases**, v. 66, n. 1, p. 149–153, 2018.

WING, K. et al. A T164S Mutation in the Dengue Virus NS1 Protein is Associated with Greater Disease Severity in Mice. **Science Translational Medicine**, v. 8, n. 330, 2016.

WONG WEI XIANG, B. et al. Dengue Virus Infection Modifies Mosquito Blood-Feeding Behavior to Increase Transmission to the Host. **Proceedings of the National Academy of Sciences**, v. 118, n. 21, 2021.

WU, X. et al. Data Mining with Big Data. **IEEE Transactions on Knowledge and Data Engineering**, v. 26, n. 1, p. 97–107, 2014.

YENAMANDRA, S. P. et al. Evolution, Heterogeneity and Global Dispersal of Cosmopolitan Genotype of Dengue Virus Type 2. **Scientific Reports**, v. 11, n. 1, 2021.