UNIVERSIDADE FERDERAL DO RIO GRANDE DO NORTE
CENTRO DE BIOCIÊNCIAS
REDE NORDESTE DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

GIOVANNA MELO MARTINS SILVA

IDENTIFICAÇÃO DE SEQUÊNCIAS DA SUBFAMÍLIA *ORTHOCORONAVIRINAE* EM METAGENOMAS E METATRANSCRIPTOMAS E ANÁLISE DA DIVERSIDADE GENÔMICA DO CORONAVÍRUS HUMANO NL63

Natal

2023

GIOVANNA MELO MARTINS SILVA

IDENTIFICAÇÃO DE SEQUÊNCIAS DA SUBFAMÍLIA *ORTHOCORONAVIRINAE* EM METAGENOMAS E METATRANSCRIPTOMAS E ANÁLISE DA DIVERSIDADE GENÔMICA DO CORONAVÍRUS HUMANO NL63

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal do Rio Grande do Norte (UFRN), fazendo parte da Rede Nordeste de Biotecnologia (RENORBIO), como requisito parcial para a obtenção do título de Doutora em Biotecnologia, área de concentração: Biotecnologia em Saúde.

Orientadora: Prof.ª Dr.ª Riva de Paula Oliveira

Coorientador: Prof. Dr. Thiago Bruce Rodrigues

Natal

2023

GIOVANNA MELO MARTINS SILVA

IDENTIFICAÇÃO DE SEQUÊNCIAS DA SUBFAMÍLIA *ORTHOCORONAVIRINAE* EM METAGENOMAS E METATRANSCRIPTOMAS E ANÁLISE DA DIVERSIDADE GENÔMICA DO CORONAVÍRUS HUMANO NL63

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal do Rio Grande do Norte (UFRN), fazendo parte da Rede Nordeste de Biotecnologia (RENORBIO), como requisito parcial para a obtenção do título de Doutora em Biotecnologia, área de concentração: Biotecnologia em Saúde.

Aprovada em: 29/09/2023

BANCA EXAMINADORA

Prof.ª Dr.ª Riva de Paula Oliveira (Orientadora)

Prof. Dr. Thiago Bruce Rodrigues (Coorientador)

Prof. Dr. Daniel Carlos Ferreira Lanza

Prof. Dr. Josélio Maria Galvão de Araújo

Prof.ª Dr.ª Gloria Regina Franco

Prof.ª Dr.ª Stela Mirla da Silva Felipe Acácio

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Setorial Prof. Leopoldo Nelson - -Centro de Biociências - CB

Silva, Giovanna Melo Martins.
Identificação de sequências da subfamília Orthocoronavirinae em metagenomas e metatranscriptomas e análise da diversidade genômica do coronavírus humano NL63 / Giovanna Melo Martins Silva. - 2023.
96 f.: il.

Tese (doutorado) - Universidade Federal do Rio Grande do Norte, Centro de Biociências, Programa de Pós-graduação em Biotecnologia. Natal,RN, 2023.
Orientação: Profa. Dra. Riva de Paula Oliveira.
Coorientação: Prof. Dr. Thiago Bruce Rodrigues.

1. Coronavírus - Tese. 2. Metagenômica - Tese. 3. Taxonomia - Tese. 4. HCoV-NL63 - Tese. 5. Filogenética - Tese. 6. Recombinação - Tese. I. Oliveira, Riva de Paula. II. Rodrigues, Thiago Bruce. III. Título.

RN/UF/BSECB CDU 616.2-053.9

Elaborado por KATIA REJANE DA SILVA - CRB-15/351

AGRADECIMENTOS

À professora Dra. Riva de Paula Oliveira, pela sua paciência e dedicação, por ser um exemplo para todos os alunos que buscam seguir a carreira acadêmica, por sua orientação contínua e atenciosa, por sempre me incentivar a superar novos desafios, por ter me ensinado a beleza da docência e a arte da pesquisa científica, por ter formado a profissional que sou hoje, serei eternamente grata.

Ao professor Dr. Thiago Bruce Rodrigues, por ter se disponibilizado a me ensinar e ajudar no momento mais desafiador do curso de doutorado. Muito obrigada por toda a sua paciência e incentivo.

À professora Dra. Lucymara Fassarella Agnez Lima, por me dar a oportunidade de aprender com sua experiência, muito obrigada.

Ao doutorando Diego Arthur de Azevedo Morais, por toda paciência, dedicação e ensinamentos voltados para a área de Bioinformática.

Ao professor Dr. José Miguel Ortega, da Universidade Federal de Minas Gerais (UFMG), por gentilmente nos ceder acesso ao uso dos computadores de sua instituição, pela troca de conhecimentos e por sua disponibilidade e atenção contínuos.

Ao professor Dr. Jorge Estefano Santana de Souza e ao doutorando Raul Maia Falcão, por toda dedicação, paciência e orientações. Sou muito grata pela oportunidade de aprendizado.

A toda a equipe do Laboratório de Genética Bioquímica, por toda a ajuda, ensinamentos, companheirismo e amizade. Especialmente a César Muñoz, Ana Moniz, Liliane de Castro, Aian Viana e Maria Luíza. Obrigada por tornarem o ambiente de trabalho mais leve e divertido.

Agradeço imensamente a Flávia Roberta, Cynthia Haynara, Weslley Souza, Francisco Carlos, Júlia Freitas, Lídia Rodrigues e Gizélia Rodrigues. Obrigada por me ouvirem e compartilharem suas experiências comigo.

Às minhas queridas amigas Giulia Morais, Ingriddy Caroline e Nathália Hernandes. Obrigada por todo o apoio, conselhos e acolhimento. Serei eternamente grata por tão valiosas amizades.

À Universidade Federal do Rio Grande do Norte (UFRN), minha *alma mater*, pela formação acadêmica de qualidade.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro.

À minha família, por todo o apoio, compreensão, incentivo e conselhos.

Aos meus irmãos Victor e Samuel, obrigada por me apoiarem incondicionalmente.

À minha mãe, Marcia Maria Dias de Melo Silva, por sempre acreditar em mim e nunca me deixar desistir, por ser um exemplo de mãe e determinação, por me ensinar o que é amar incondicionalmente.

RESUMO

O desenvolvimento de tecnologias de sequenciamento de alto rendimento na última década resultou em grandes repositórios de dados brutos de sequenciamento e projetos metagenômicos. Paralelamente, a biologia computacional e a bioinformática se tornaram aliadas importantes, resultando na criação de diversas ferramentas de classificação taxonômica. A pandemia da COVID-19 (Coronavirus Disease 2019) demonstrou a importância de investigar melhor os coronavírus, especialmente os coronavírus humanos que impactam a saúde e os serviços públicos. Entre eles, o HCoV-NL63 é um coronavírus sazonal que afeta os seres humanos, está distribuído globalmente, causa doenças respiratórias leves a moderadas e compartilha o mesmo receptor para entrada no hospedeiro que o SARS-CoV e o SARS-CoV-2. Este trabalho utiliza bancos de dados públicos de sequências para 1) comparar duas ferramentas de bioinformática, Kaiju e Burrows-Wheeler Aligner (BWA), na identificação de sequências virais da subfamília Orthocoronavirinae em metagenomas humanos e 2) realizar uma análise filogenética abrangente do HCoV-NL63. Foram selecionadas aleatoriamente 1169 amostras humanas (de um total de 3670) e a classificação taxonômica foi realizada com o uso do Kaiju (análise de dois passos com diferentes bancos de dados de referência). Usando o Kaiju, foram encontradas 150 amostras positivas para CoVs, no entanto, não foi possível montar nenhum genoma. Além disso, todas as 3670 amostras foram analisadas usando o BWA, com resultados negativos para CoVs, exceto uma amostra que continha sequências do HCoV NL63. Nossos resultados sugerem que as amostras investigadas não continham coronavírus, apenas sequências conservadas e similares entre organismos, o que foi superestimado pela ferramenta Kaiju. Por outro lado, o BWA pode ser utilizado para identificar viromas em diversos produtos de sequenciamento. Neste estudo, também utilizamos 173 sequências de genes de espícula disponíveis publicamente do HCoV-NL63 para realizar a análise filogenética. A análise de máxima verossimilhança resultou em oito subgenótipos (A1, A2, A3, B1, B2, C1, C2 e C3), consolidando a divisão da linhagem B em B1 e B2. Os subgenótipos B2 (20,1%), B1 (19,5%) e C3 (16,1%) foram os mais prevalentes. Foi encontrada seleção positiva nas posições S1 (V57, G96, N431) e S2 (V1177, E1206). Foram detectadas sete substituições não sinônimas na região BLR. Em comparação com outros genótipos, o genótipo B apresenta um número significativo de substituições de aminoácidos na região S1. Nossa análise também identificou uma alta prevalência de eventos de recombinação dentro da região S1 (12-307 aminoácidos). O genótipo A, que apresenta o maior número de eventos de recombinação, mostra um período restrito de surgimento tanto em termos de tempo quanto de local, exibindo um padrão de seleção purificadora. Nossos resultados destacam a importância da vigilância epidemiológica e filogenética contínua, sendo fortemente recomendada para prever a evolução de futuras variantes.

Palavras-chave: coronavírus; metagenômica; taxonomia; HCoV-NL63; filogenética; recombinação.

ABSTRACT

The development of high-throughput sequencing technologies in the last decade resulted in large repositories of raw sequencing data and metagenomic projects. Alongside, computational biology and bioinformatics became important allies and many taxonomic classification tools were created. The pandemic of COVID-19 (Coronavirus Disease 2019) has shown the importance of better investigating coronaviruses, specially human coronaviruses that impact health and public services. Among them, HCoV-NL63 is a seasonal human coronavirus that spreads worldwide, causes mild to moderate respiratory disease, and shares the same receptor for host entry as SARS-CoV and SARS-CoV-2. This study uses public databases of sequences to 1) compares two bioinformatic tools, Kaiju and Burrows-Wheeler Aligner (BWA), in identifying from Orthocoronavirinae viral sequences in human metagenomes and 2) conduct a comprehensive phylogenetic analysis of HCoV-NL63. 1169 human samples were randomly selected (from 3670), and taxonomic classification was performed with Kaiju (double-step analysis with different reference databases). 150 samples were found positive for CoVs using Kaiju, however, it was not possible to assembly any genomes. In addition, all 3670 samples were analyzed using BWA, with negative results for CoVs, except for one sample containing sequences from HCoV NL63. Our findings suggest that the investigated samples did not have coronaviruses, only conserved and similar sequences between organisms, which was overestimated by the Kaiju tool. On the other hand, BWA can be used to identify viromes in various sequencing products. In this study, we also used 173 publicly available spike genes sequences from HCoV-NL63 to perform phylogenetic analysis. Maximum likelihood analysis resulted in eight subgenotypes (A1, A2, A3, B1, B2, C1, C2, and C3), consolidating the lineage B division into B1 and B2. Subgenotypes B2 (20.1%), B1 (19.5%), and C3 (16.1%) were the most prevalent. Positive selection was found in S1 (V57, G96, N431) and S2 (V1177, E1206) residues. Seven non-synonymous substitutions were detected in the RBD region. Compared to other genotypes, genotype B has a remarkable number of amino acid substitutions in the S1 region. Our analysis also detected a high prevalence of recombination events within the S1 region (12–307 amino acids). Genotype A, which has the highest number of recombination events, shows a constricted period of emergence in both time and place, displaying a pattern of purifying selection. Our results highlight the importance of continuous epidemiologic and phylogenetic surveillance, which is strongly recommended in order to predict the evolution of future variants.

Keywords: coronavirus; metagenomics; taxonomy; HCoV-NL63; phylogenetics; recombination.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES E SIGLAS

ACE2	Enzima Conversora de Angiotensina 2
+ssRNA	RNA Fita Simples Senso Positivo
229E	Coronavírus Humano HCoV-229E
aa	aminoácidos
ACoV	Alpaca Coronavírus
APN	Aminopeptidase N
APNPMA	Agrupamento por Pares Não Ponderados com Média Aritmética
BatCoV	Coronavírus de Morcego
BCV	Coronavírus Bovino
BWA	Em inglês, Burrows-Wheeler Aligner
BWT	Em inglês, Burrows-Wheeler Transform ou algoritmo de transformação de dados de Burrows-Wheeler
CCoV	Coronavírus Canino
COVID-19	Em inglês, Coronavirus Disease 2019 ou doença causada por coronavírus 2019
CoVs	Coronavírus
CRCoV	Coronavírus Respiratório Canino
CRR	Complexo de Replicação e Transcrição
DLR	Domínio de Ligação ao Receptor
dN	Substituições não sinônimas por sítio não sinônimo
DPP4	Dipeptidil-peptidase 4
dS	Substituições sinônimas por sítio sinônimo
E	Proteína do Envelope Viral
EMBL-EBI	Em inglês, European Bioinformatics Institute ou Instituto Europeu de Bioinformática
EUA	Estados Unidos da América
FCoV	Coronavírus Felino
HCoVs	Coronavírus capazes de infectar os seres humanos
HKU1	Coronavírus Humano HCoV-HKU1
IB	Inferência Bayesiana

IBV	Vírus da Bronquite Infecciosa
ICTV	Em inglês, the International Committee on Taxonomy of Viruses ou Comitê Internacional de Taxonomia dos Vírus
JV ou NJ	Junção de vizinhos ou do inglês, Neighbour-Joining
LSD	"Least Square Dating", do inglês
M	Proteína da Membrana Viral
MCCM	Algoritmo Monte Carlo via Cadeia de Markov
MCMC	Em inglês, Markov Chain Monte Carlo
MERS-CoV	Middle East Respiratory Syndrome Coronavirus, do inglês, ou Coronavírus da Síndrome Respiratória do Oriente Médio
MHV	Vírus da Hepatite do Camundongo
MV	Máxima Verossimilhança
N	Proteína do Nucleocapsídeo Viral
NCBI	National Center for Biotechnology Information
NL63	Coronavírus Humano HCoV-NL63
nt	Nucleotídeo(s)
OC43	Coronavírus Humano HCoV-OC43
OMS	Organização Mundial da Saúde
ORF	Open Reading Frame, do inglês, ou janela de leitura
Par	Parcimônia
PNE ou nsp	Proteína não estrutural ou, do inglês, non-structural protein
poli-A	cauda poliadenilada
RNAmsg	RNA mensageiro -RNAm- subgenômicos
RNAsg	RNAs subgenômicos
RpdR	RNA polimerase dependente de RNA
RRT-C	Sequência regulatória de transcrição presentes ao longo do corpo do genoma viral
RRT-L	Sequência regulatória de transcrição líder
S	Proteínas de superfície ou Spike
S1	Subdomínio 1 da proteína Spike
S1-CTD	Domínio C-terminal da proteína
S1-NTD	Domínio N-terminal da proteína

S2	Subdomínio 2 da proteína Spike
SADS-CoV	Coronavírus Causador da Síndrome Suína de Diarreia Aguda
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus, do inglês, ou Coronavírus da Síndrome Respiratória Aguda Grave
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2, do inglês, ou Coronavírus da Síndrome Respiratória Aguda Grave 2
SARSr-CoV	Coronavírus Similares ou Relacionados ao SARS
SDAV	Vírus da Sialodacrioadenite de Ratos
TCoV	Coronavírus de Peru
TGEV	Coronavírus de Gastroenterite Transmissível
TMPRSS2	Proteína serina protease transmembranar 2
UFBoot	UltrafastBootstraps ou método ultrarápido Bootstrap
UPGMA	Em inglês, Unweighted Pair-Group Method with Arithmetic Mean

SUMÁRIO

1 INTRODUÇÃO

1.1 ANÁLISES METAGENÔMICAS E METATRANSCRIPTÔMICAS NA IDENTIFICAÇÃO DE PATÓGENOS

À medida que as técnicas de sequenciamento se tornaram cada vez mais eficientes e acessíveis, a biologia computacional e a bioinformática tornaram-se importantes aliadas na análise da grande quantidade de informações genômicas e metadados biológicos associados gerados (LEVY; BOONE, 2019). Nesse contexto, diversas subáreas relacionadas às ciências biológicas e da saúde puderam se beneficiar de forma inovadora. A metagenômica e metatranscriptômica clínicas, em especial, prometem revolucionar a medicina de precisão. Enquanto a metagenômica investiga a diversidade de microrganismos presentes em amostras de DNA, a metatranscriptômica é a área que analisa a diversidade de microrganismos de acordo com a diversidade de expressão gênica identificada em amostras de RNA. Em ambos os casos o objetivo é a identificação e caracterização das comunidades de microrganismos (TERRÓN-CAMERO et al., 2022).

Um relevante exemplo de como a bioinformática veio a se tornar uma importante aliada a problemas biológicos, com aplicação metagenômica, foi o diagnóstico clínico de um menino de 14 anos com imunodeficiência combinada grave que sofreu por quatro meses sem diagnóstico, mesmo com biópsia do cérebro. Ele apresentava febre, dores de cabeça, hidrocefalia e estado epiléptico, e somente após o sequenciamento de nova geração do líquido cefalorraquidiano ele foi diagnosticado com *Leptospira santarosai*, devidamente tratado e recebeu alta 32 dias depois (LEVY; BOONE, 2019; WILSON et al., 2014).

Um dos maiores desafios para se entender melhor fatores ecológicos e a diversidade dos vírus está na dificuldade em isolá-los e cultivá-los. Nesse contexto, as tecnologias de sequenciamento metagenômico permitiram a descoberta de novos vírus independentemente da necessidade de cultivo ou da sequência genômica apresentada (NG; TAN, 2017). A possibilidade de se conduzir sequenciamentos de alto rendimento de genomas virais considerados grandes, a partir mínimas quantidades de amostras, foi revolucionária. Especialmente para os vírus de RNA, como os CoVs, os avanços de técnicas moleculares e dos sistemas de genética reversa permitiram a síntese de vírus recombinantes para estudos de ciclo celular, infecção e de como as alterações genômicas influenciam o fenótipo e comportamento

virais (DRIOUICH et al., 2019; NG; TAN, 2017; TORII et al., 2021). É nesse contexto que as ferramentas de bioinformática ou as análises *in silico*, ganham destaque por facilitarem comparações de sequências de genomas, estudos preditivos de estruturas e de funções proteicas, análises taxonômicas, análises de filogenética e filodinâmica, dentre outras.

Como consequência à quantidade de dados gerados pela acessibilidade às tecnologias de sequenciamento, muitas ferramentas de bioinformática foram desenvolvidas, porém, ainda não é tão fácil escolher a melhor ferramenta ou estratégia. A identificação de patógenos em metagenomas pode usar uma abordagem baseada em sequências de aminoácidos (do inglês, *gene-based*) ou sequências de nucleotídeos (do inglês, *reference-based*) (SONG et al., 2021). Ambos têm vantagens e desvantagens e podem funcionar de forma diferente dependendo das características dos dados/amostras.

Entre as ferramentas de classificação taxonômica mais utilizadas estão Centrifuge, Clark, Kaiju, Kraken 2, Genome Detective, Diamond e BWA (CARBO et al., 2022; MORAIS et al., 2022; YE et al., 2019). O Kaiju é um programa de classificação taxonômica baseado na comparação de sequências com um banco de dados de referência de proteínas (MENZEL; NG; KROGH, 2016). E o Burrows-Wheeler Aligner (BWA) é um programa que usa o algoritmo de transformação de dados de Burrows-Wheeler (Burrows-Wheeler Transform - BWT) para alinhar sequências de nucleotídeos curtas contra uma sequência de referência grande, permitindo incompatibilidades e lacunas (LI; DURBIN, 2009). Já o Genome Detective está disponível online nas versões paga e livre de custos, e contempla toda a execução de um pipeline desde o download da amostra, análise de qualidade e identificação taxonômica de vírus, até a montagem dos genomas virais e a chamada de variantes (VILSKER et al., 2019). Esse software utiliza ambas as estratégias de alinhamento baseadas em aminoácidos e em nucleotídeos.

Em um trabalho recente, Carbo et al.(CARBO et al., 2022) compararam o desempenho de cinco classificadores metagenômicos (Centrifuge, Clark, Kaiju, Kraken 2, Genome Detective) para a detecção de vírus utilizando amostras clínicas respiratórias. Nesse estudo, eles destacam que parâmetros como especificidade, sensibilidade e tipo de sequenciamento devem ser levados em consideração ao se implementar a detecção metagenômica viral no diagnóstico clínico. No contexto da pandemia da COVID-19, um estudo recente de Wahba et al. (WAHBA et al., 2020) realizaram uma busca por sequências de SARS-CoV-2 em conjuntos de dados metagenômicos públicos com a palavra chave “virome” usando a ferramenta BWA. Este estudo foi um exemplo dentre os trabalhos que mostraram a possibilidade de se “reanalisar” algumas das sequências brutas com acesso público para a busca de patógenos em metagenomas.

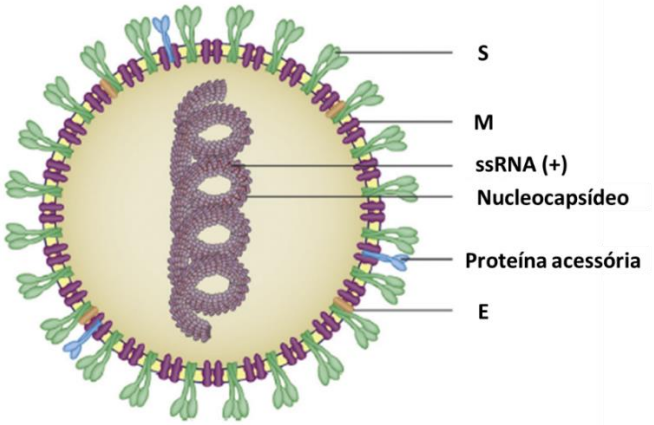
Com uma abordagem parecida, Kawasaki et al. (KAWASAKI et al., 2021) analisaram mais de 46.000 amostras de RNA-seq públicas de mamíferos e aves com o objetivo de reconhecer possíveis infecções por vírus de RNA. Eles conseguiram identificar aproximadamente 900 infecções e novos genomas virais zoonóticos com potencial de causar doenças em humanos. Um outro grupo, de Melnick et al. (MELNICK et al., 2021), em estudos anteriores, também utilizou dados públicos de RNA-seq para estabelecer um modelo de análises de dados para identificar sequências virais em amostras de humanos. Também envolvendo o recente SARS-CoV-2, Grimaldi et al. (GRIMALDI et al., 2022) utilizaram RNA extraído de amostras clinicas (swab nasal) para estabelecer uma estratégia eficaz e acessível de monitoramento da variabilidade genética do vírus em uma região da Itália. Com um objetivo parecido, o grupo de Karthikeyan et al. (KARTHIKEYAN et al., 2022) buscou identificar variantes do SARS-CoV-2 em águas residuais no campus da Universidade da Califórnia em San Diego. Nesse estudo, eles foram capazes de identificar variantes emergentes de preocupação em amostras de águas residuais até 14 dias antes, quando comparadas com as amostras de vigilância clínica.

Diante do que foi apresentado, o presente trabalho tem a finalidade de identificar sequências de coronavírus em metagenomas e transciptomas de hospedeiros, e realizar análises filogenéticas e de caracterização molecular, de forma a contribuir para um melhor esclarecimento da distribuição desses vírus e suas características filogenéticas.

1.2 OS CORONAVÍRUS E SUAS CARACTERÍSTICAS MOLECULARES

Os coronavírus (CoVs) são membros da família *Coronaviridae*, constituída por vírus de genoma não segmentado, de tamanho entre 25 e 32 kb, RNA fita simples senso positivo (+ssRNA) e com estruturas 5’-cap e 3’-cauda poliadenilada /poli-A (estruturas que contribuem para a estabilidade da molécula). Os vírions (forma extracelular completa) são esféricos, envelopados, e contém glicoproteínas proeminentes denominadas S ou “Spike” (em inglês) ou espícula (em português), formando uma estrutura similar a uma “corona”/“coroa”, característica que dá nome à família (**Figura 1**) (CHEN; LIU; GUO, 2020; PAYNE, 2017).

Figura 1: Estrutura Geral de Vírion da Subfamília *Orthocoronavirinae*



Fonte: Adaptado de PAYNE, SUSAN, 2017.

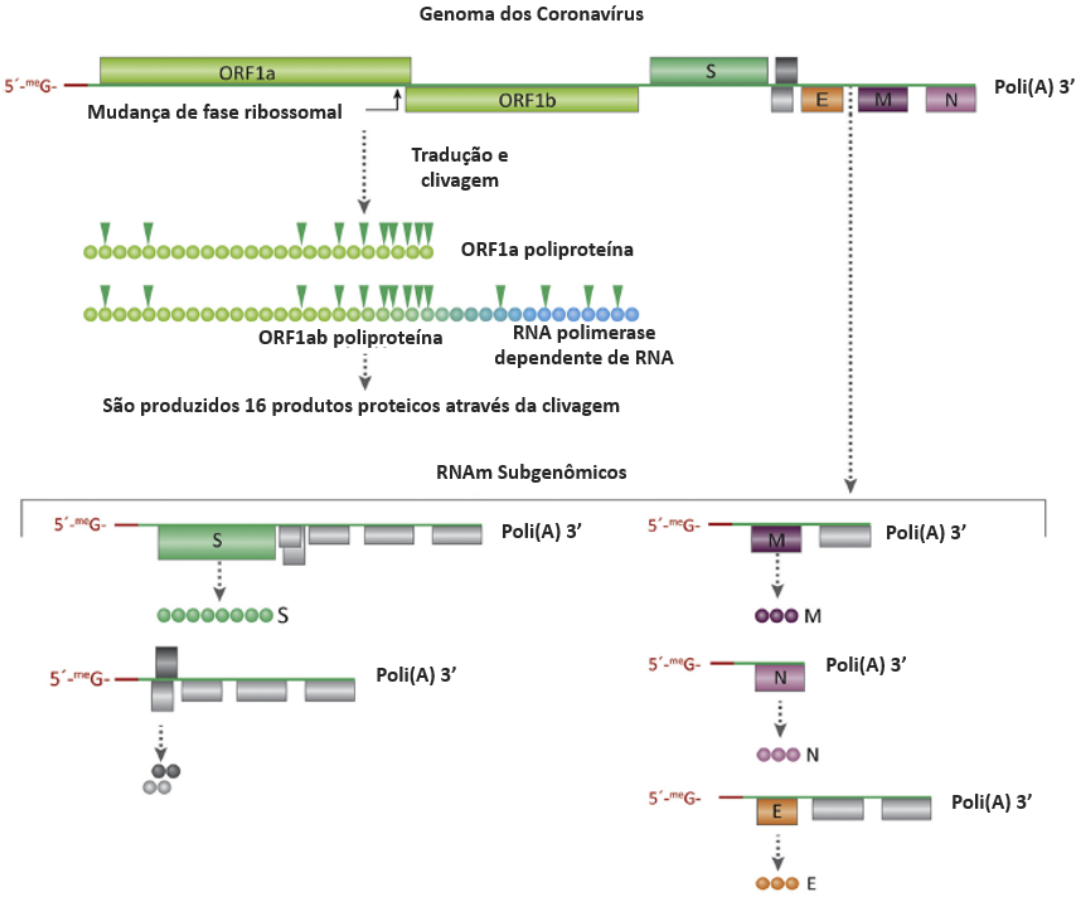
O genoma desses vírus apresenta uma organização padrão, dividida nas sequências ou ORFs (do inglês, *Open Reading Frame*/ "janelas de leitura"), que dão origem à poliproteína 1a/1ab e às proteínas de superfície (S ou Spike), do envelope (E), da membrana (M) e do nucleocapsídeo (N), nessa ordem, além das sequências de algumas proteínas acessórias intercaladas (**Figura 2**) (CHEN; LIU; GUO, 2020; PAYNE, 2017). Quase dois terços do genoma desses vírus é composto pela OFR1ab, responsável por dar origem a várias proteínas não estruturais (PNEs) e à RNA polimerase dependente de RNA (RpdR). Ela pode ser dividida nos segmentos 1a e 1b por causa de um sítio que provoca uma mudança na janela de leitura do ribossomo (denominado *ribossomal frameshifting*, do inglês), no momento da tradução (síntese proteica) (**Figura 2**). Essa mudança na janela de leitura é como se fosse um "deslise" do ribossomo que permite que ele ignore ou pule o códon de terminação da sequência 1a e siga traduzindo a sequência 1b, produzindo uma proteína mais longa, correspondente ao seguimento 1ab. Esse processo só ocorre em cerca de 20 a 25% dos casos, e o produto 1b gerado é o que contém a RpdR, além de outras PNEs (CHEN; LIU; GUO, 2020; PAYNE, 2017). Embora o processo de replicação viral seja comum a todos os CoVs, cada um deles apresenta características distintas, seja em relação aos seus hospedeiros (final, intermediário ou reservatório), aos receptores e proteínas acessórias (molecular), ao período de incubação e patogenicidade (clínica) (KESHEH et al., 2022).

O ciclo desses vírus inicia-se com a ligação entre um receptor, presente na superfície das células do hospedeiro, e o domínio de ligação ao receptor (DLR) do vírus, que está presente na proteína Spike. Quando ocorre a ligação DLR-receptor, a Spike sofre mudanças conformacionais e é clivada de forma a expor o chamado "peptídeo de fusão" presente em sua

estrutura. É dessa forma que a fusão das membranas ocorre e o genoma viral é internalizado (PAYNE, 2017). Vale ressaltar que tanto os receptores, quanto a clivagem da Spike, os sítios e condições moleculares para o processo de fusão variam entre as espécies de CoVs (PAYNE, 2017).

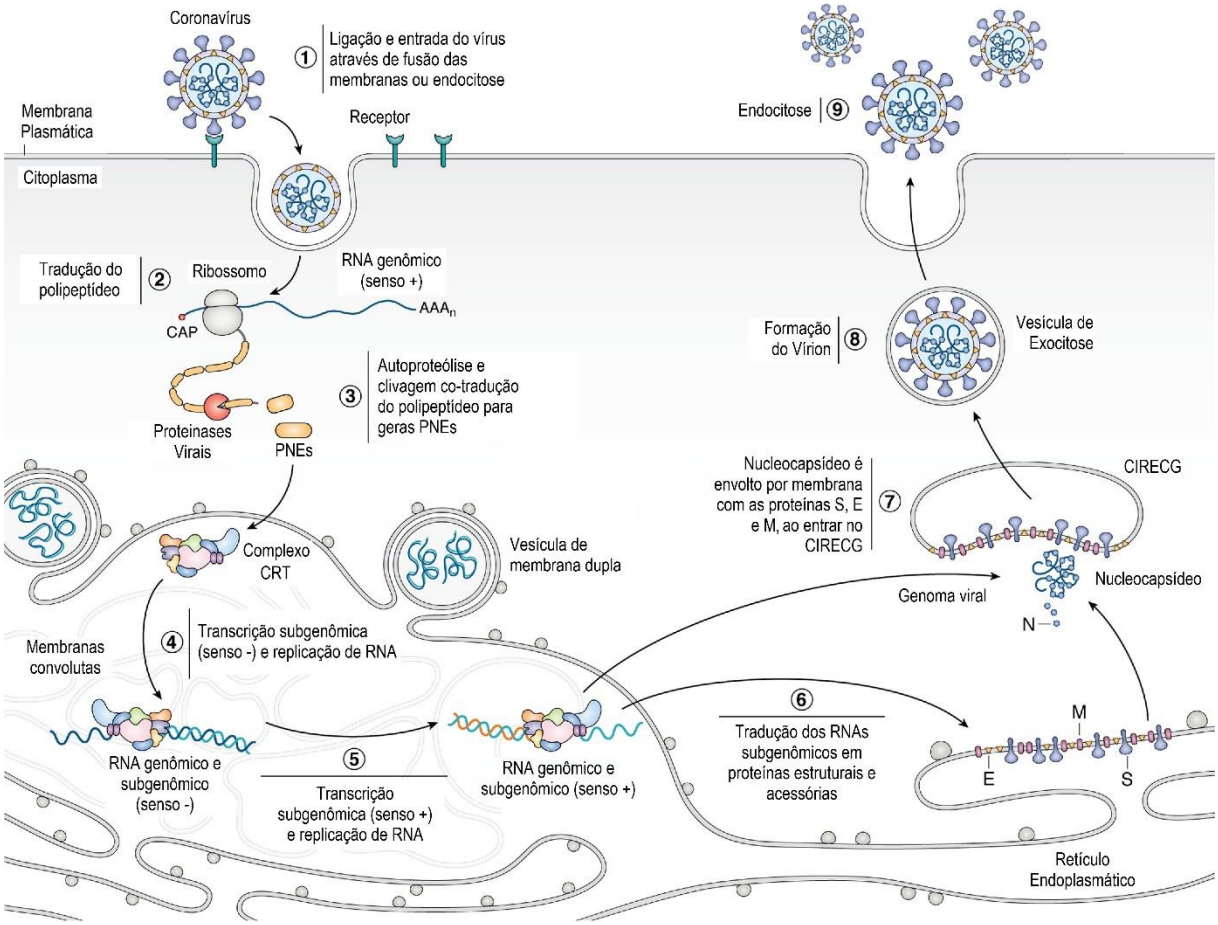
Como o genoma é de RNA fita positiva, a molécula é utilizada diretamente para a tradução de proteínas, mais especificamente a poliproteína 1a/1ab, correspondente às PNEs, responsáveis pela formação do complexo de replicação e transcrição (CRT) junto à RpdR (CHEN; LIU; GUO, 2020; PAYNE, 2017). O complexo CRT sintetiza os RNAs subgenômicos (RNAsg) através de um processo de transcrição descontinuada, que contribui para as altas taxas recombinação entre esses vírus. Nesse processo de transcrição descontinuada, o CRT precisa de desprender/desacoplar da fita molde de RNA para conseguir sintetizar os RNAsg, isso ocorre porque as sequências principais dessas moléculas se encontram próximo à extremidade 3’ do genoma viral e, além disso, todos os RNAsg compartilham uma sequência reguladora comum que se localiza próximo à extremidade 5’ do genoma. Dessa forma, cada RNAsg apresenta uma sequência comum 5’ e distintas sequências terminais 3’ que resultam nas proteínas estruturais (S ou Spike, E, M e N) e proteínas acessórias. Os RNAsg são fitas negativas e servem de moldes para RNAs mensageiros -RNAm- subgenômicos (RNAmsg) (CHEN; LIU; GUO, 2020; PAYNE, 2017). As representações do processo expressão gênica e do ciclo viral completo estão nas **Figura 2** e **Figura 3**.

Figura 2: Organização do Genoma e Estratégia de Expressão Gênica dos Coronavírus



Fonte: Adaptado de PAYNE, SUSAN, 2017.

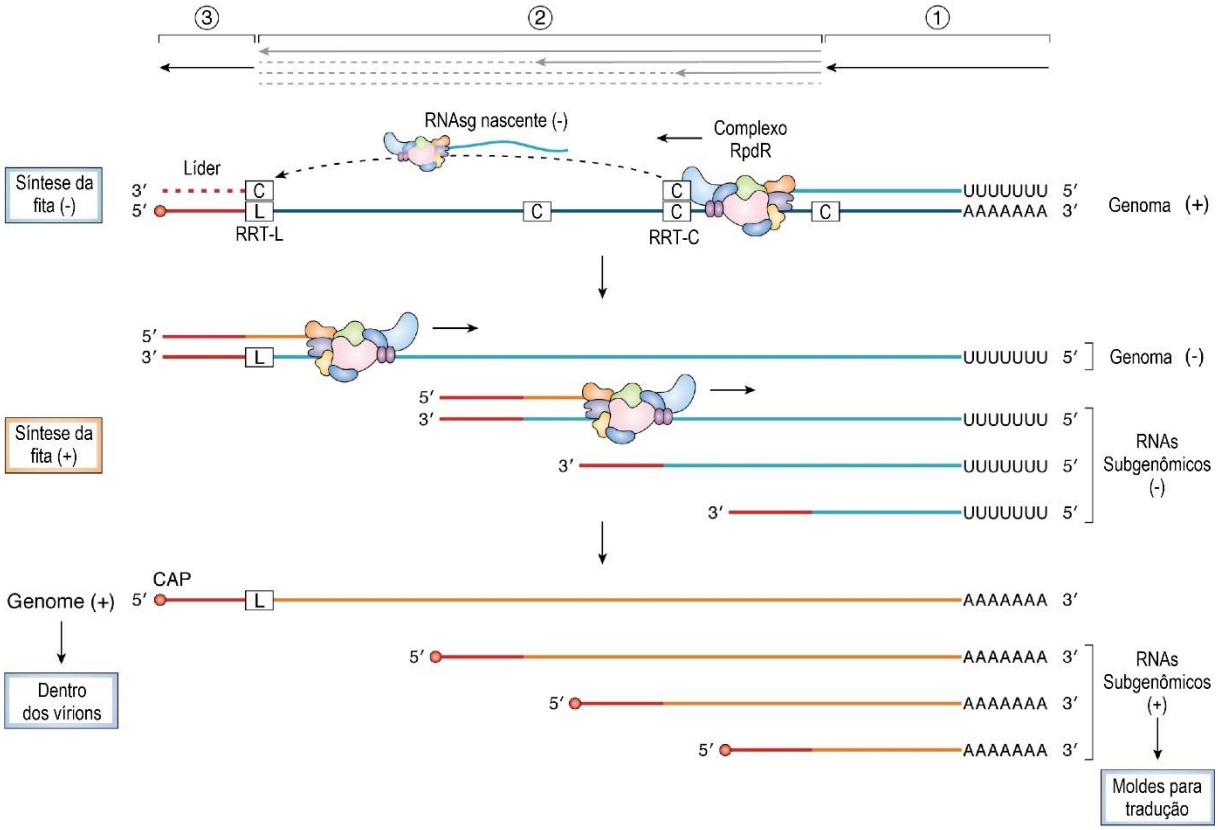
Figura 3: Ciclo viral completo dos coronavírus



Os coronavírus interagem com um receptor na superfície das células do hospedeiro e depositam seus genomas de RNA(+) no citoplasma por endocitose ou fusão das membranas (1). O genoma de RNA(+) é traduzido pela maquinaria de tradução do hospedeiro (2) para produzir poliproteínas que são clivadas co-traducionalmente por proteases codificadas na própria poliproteína e gerar componentes do CRT (3). O CRT usa o genoma como molde para gerar RNAsgs e RNAs genômicos de comprimento integral senso (-) (4), que por sua vez são usados como moldes para a síntese de genomas completos e RNAmsg senso (+) (5). A transcrição e a replicação ocorrem em membranas convolutas adjacentes às vesículas de membrana dupla, ambas derivadas do retículo endoplasmático rugoso. Os RNAmsg são traduzidos em proteínas estruturais e acessórias (6). O RNA genômico senso (+) é ligado ao nucleocapsídeo e brota no compartimento intermediário entre retículo endoplasmático e complexo de Golgi (CIRECG), que apresenta as proteínas estruturais S, E e M traduzidas de RNAsg de senso (+) (etapas 6 e 7). O vírion envelopado é então exportado da célula por exocitose (etapas 8 e 9). Fonte: Adaptado de HARTENIAN et al. 2020.

Dessa forma, são gerados dois tipos de RNAs nas células dos hospedeiros: os RNAsg, fitas negativas, que dão origem a RNAs complementares positivos e são traduzidos para gerar as proteínas estruturais e acessórias; e os RNAs complementares aos genomas, também fitas negativas e que servem de molde para novos genomas virais (**Figura 4**) (CHEN; LIU; GUO, 2020; HARTENIAN et al., 2020; KESHEH et al., 2022; PAYNE, 2017).

Figura 4: Resumo dos diferentes tipos de RNA envolvidos no ciclo dos coronavírus



O CRT inicia a transcrição na extremidade 3’ do genoma senso (+) (1). Ao copiar as sequências regulatórias de transcrição presentes em locais específicos ao longo do corpo do genoma RRT-C (2), o CRT pode “saltar” para a sequência regulatória de transcrição líder RRT-L (3) devido à complementaridade entre a sequência RRT-C no RNAsg nascente e Sequência RRT-L no genoma. A transcrição é retomada no novo molde e a sequência líder (mostrada em vermelho) é copiada para completar o RNAsg de fita (-). O CRT nem sempre troca de molde nas sequências RRT-C, resultando na síntese de RNA de fita (-) com comprimento de genoma. Os RNAs de fita (-) servem como modelos para a síntese de RNAs de fita (+) com comprimento de genoma ou RNAm subgenômicos. Fonte: Adaptado de HARTENIAN et al. 2020.

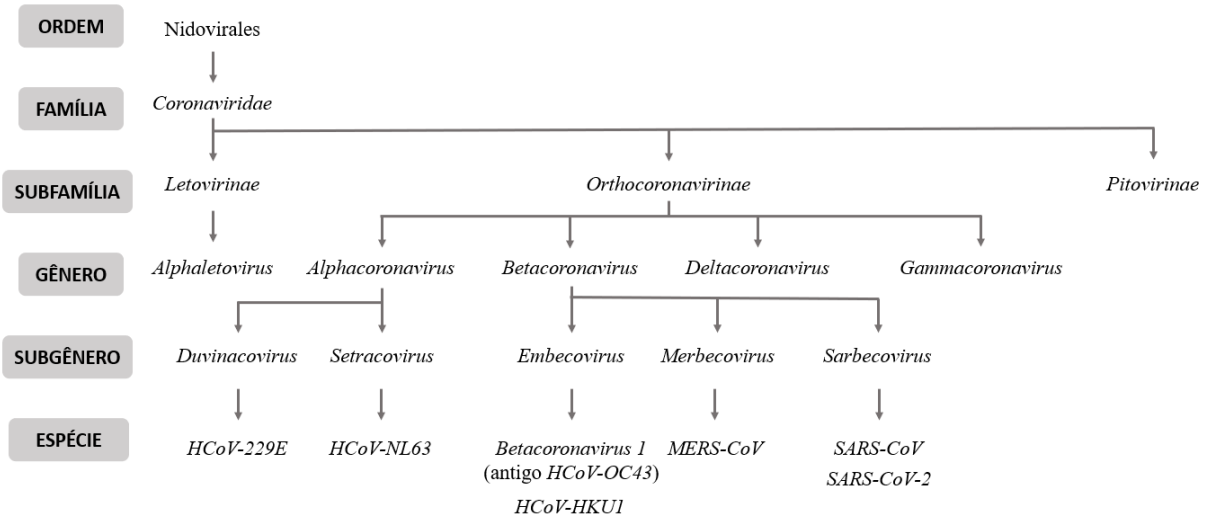
Acredita-se que o genoma de RNA dos CoVs, por ser considerado um genoma grande quando comparado a outros vírus de RNA, favorece um maior número de mutações e recombinações, o que pode contribuir para uma maior variabilidade intraespécie, para o “salto entre espécies” de hospedeiros e para o surgimento de novas espécies ou variantes (DHAMA et al., 2020; SU et al., 2016). Em resumo, são basicamente três fatores que parecem contribuir para esses processos: as altas taxas de erros permitidos pela RpdR, a maior propensão a ocorrer recombinação homóloga entre os genomas de diferentes espécies de CoVs durante o processo de transcrição descontinuada, e a infecção de um mesmo hospedeiro por diferentes CoVs (DHAMA et al., 2020; SU et al., 2016).

Os CoVs podem causar doenças em uma variedade de animais domésticos e silvestres, incluindo cavalos, camelos, gado, suínos, cachorros, gatos, roedores, aves, morcegos, coelhos,

cobras e vários outros (DHAMA et al., 2020). Eles são conhecidos e estudados desde a década de 60, principalmente por causarem doenças em animais domésticos e sintomas respiratórios leves em humanos (SU et al., 2016). Em animais, os sintomas provocados por esses vírus são principalmente gastrointestinais, a exemplo do coronavírus de gastroenterite transmissível (TGEV), coronavírus bovino (BCV), coronavírus felino (FCoV), coronavírus canino (CCoV) e o coronavírus de peru (TCoV). Alguns exemplos de vírus dessa família que causam sintomas respiratórios em animais são o vírus da bronquite infecciosa (IBV), que causa bronquite infecciosa aviária em galinhas, o coronavírus respiratório canino (CRCoV), que causa doença respiratória em cães, e o vírus da hepatite do camundongo (MHV), que pode causar uma encefalite desmielinizante progressiva nesses animais (SU et al., 2016). Mais recentemente, em 2017, também foi identificado o coronavírus causador da síndrome suína de diarreia aguda (SADS-CoV) (CUI; LI; SHI, 2019).

Segundo o Comitê Internacional de Taxonomia dos Vírus (do inglês, *the International Committee on Taxonomy of Viruses – ICTV*), os CoVs são membros da família *Coronaviridae*, subfamília *Orthocoronavirinae*, e estão distribuídos em quatro gêneros: *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus* e *Gammacoronavirus* (**Figura 5**). Os gêneros Alpha e Beta comumente infectam mamíferos, incluindo os humanos, enquanto Delta e Gamma infectam aves e mamíferos (DHAMA et al., 2020). Existem sete CoVs capazes de infectar os seres humanos (HCoVs) e causar sintomas respiratórios: HCoV-229E (229E), HCoV-OC43 (OC43), HCoV-NL63 (NL63), HCoV-HKU1 (HKU1), SARS-CoV (do inglês, S*evere Acute Respiratory Syndrome Coronavirus*/coronavírus da síndrome respiratória aguda grave), SARS-CoV-2 (do inglês, S*evere Acute Respiratory Syndrome Coronavirus 2*/ coronavírus da síndrome respiratória aguda grave 2) e MERS-CoV (do inglês, *Middle East Respiratory Syndrome Coronavirus*/coronavírus da síndrome respiratória do oriente médio) (SU et al., 2016). É importante destacar que o ICTV considera o HCoV-OC43 como um dos representantes da espécie *Betacoronavírus 1*, e ambos SARS-CoV e SARS-CoV-2 como uma só espécie: SARS-CoV, como representado na **Figura 5**.

Figura 5: Esquema Representativo da Classificação Taxonômica dos Coronavírus Humanos



Fonte: Elaborada pela autora.

1.3 CORONAVÍRUS HUMANOS: PRINCIPAIS CARACTERÍSTICAS, HISTÓRIA E HOSPEDEIROS

Entender a origem e evolução dos coronavírus é a melhor forma de solidificar estratégias eficazes para lidar com o surgimento de novas espécies e variantes que possam vir a causar consequências negativas na saúde humana e animal. Embora não seja uma tarefa fácil, um excelente ponto de partida é comparar informações como descoberta, características biológicas e moleculares, relação com hospedeiros e histórico de infecções. Dentre os HCoVs, o 229E e o OC43 (alfa e betacoronavírus) já eram bastante estudados por causarem gripes e resfriados comuns, e o SARS-CoV e o MERS-CoV foram responsáveis por surtos importantes de síndrome respiratória aguda grave na China, em 2002, e no Oriente Médio em 2012, respectivamente (SU et al., 2016). Além disso, o SARS-CoV-2 foi o causador da maior crise de saúde pública da história humana causada por um coronavírus. As principais características dos HCoVs estão resumidas na **Tabela 1:** Principais características dos coronavírus humanos**Tabela 1**.

Tabela 1: Principais características dos coronavírus humanos

Vírus	Descoberta do vírus (ano/país)	Receptor	Modo de internalização	Genes acessórios	Período de incubação	Severidade da doença	Origem	Hospedeiros intermediários
HCoV-229E	1962/Estados Unidos	APN	Endocitose	ORF4a, ORF4b	2–4 dias	Leve em indivíduos imunocompetentes e grave em bebês, idosos, e indivíduos imunocomprometidos	Morcegos	Morcegos
HCoV-OC43	1967/Estados Unidos	N-acetil-9-O-acetilneuramínico	Endocitose dependente de caveolina-1	NS12.9, NS2	2–4 dias		Morcegos	Bovinos
SARS-CoV	2002/China	ACE2	Mediado por receptor, independente de clatrina, endocitose independente de claveolina, provável por balsas lipídicas	ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8a, ORF8b, ORF9b	2–10 dias	Leve a grave com quase 10% de mortalidade	Morcegos	Civeta de palmeira asiática
HCoV-NL63	2004/Holanda	ACE2	Via endossomal (mediada por clatrina)	ORF3	2–4 dias	A mesma do HCoV-229E	Morcegos	
HCoV-HKU1	2005/China/Hong Kong	ácido siálico O-acetilado	TMPRSS2 de superfície e catepsinas endossomais	ORF4, NS2	2–4 dias		Morcegos	Ratos
MERS-CoV	2012/Arábia Saudita	DPP4 (CD26)	Via endossomal dependente de catepsina L, e fusão da membrana celular dependente de TMPRSS2	ORF3, ORF4a, ORF4b, ORF5, ORF8b	2–14 dias (em média 5,5-6,5 dias)	Leve a grave com quase 35% de mortalidade	Morcegos	Dromedários
SARS-CoV-2	2019/China	ACE2	Via endossomal e por fusão de membranas	ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF9c, ORF10	2–12 dias (em média 5,1 dias)	Leve a grave (pandemia atual)	Morcegos	Pangolim (provável)

Abreviações: ACE2, enzima conversora de angiotensina 2; APN, aminopeptidase N; DPP4, dipeptidil-peptidase 4; TMPRSS2, serina protease transmembranar 2. Fonte: Adaptado de KESHEH et al.,2021.

Uma importante característica dos HCoVs é que todos apresentam origem animal, a exemplos de morcegos (229E, NL63, SARS, SARS-2 e MERS) e roedores (OC43 e HKU1) (CUI; LI; SHI, 2019). Além disso, alguns pesquisadores e estudos epidemiológicos apontaram evidências de que o SARS e o MERS-CoV foram transmitidos para humanos diretamente de animais (das civetas de mercados e dos dromedários, respectivamente), atuando como hospedeiros intermediários (**Figura 6**) (CUI; LI; SHI, 2019; SU et al., 2016). Um outro caso que corrobora com essa linha de raciocínio são casos de isolamento de CoVs caninos e felinos em humanos com sintomas respiratórios, como o novo *Alphacoronavírus* recombinante canino-felino, denominado coronavírus canino CCoV-HuPn-2018, que foi detectado em crianças de áreas indígenas na Malásia (**Figura 6**) (TANG; LIU; CHEN, 2022; TORTORICI et al., 2022). Este vírus pode vir a ser o oitavo HCoV a ser identificado, embora o seu padrão de transmissão ainda não esteja totalmente esclarecido. No entanto, esse é mais um exemplo da capacidade dos CoVs de transmissão entre espécies e de como a circulação deles é subestimada. Dessa forma, não seria possível compreender esses vírus sem entender a complexa relação que apresentam com seus hospedeiros.

Segundo Ye et al. (2020), existem diferentes tipos de hospedeiros: evolutivo, natural, reservatório, intermediário e amplificador. O hospedeiro evolutivo é aquele que abriga um vírus ancestral (intimamente relacionado e que compartilha alta identidade de sequência de nucleotídeos), enquanto o vírus ancestral (filogeneticamente relacionado, porém sem a mais alta identidade de sequência), geralmente é bem adaptado e não patogênico nesse hospedeiro. Já o hospedeiro reservatório é aquele capaz de abrigar o vírus continuamente e por um longo período. Em ambos os casos, os hospedeiros são naturalmente infectados e são considerados hospedeiros naturais (YE et al., 2020). Quando o vírus foi recém introduzido em seres humanos, antes ou próximo à infecção de hospedeiros considerados intermediários, então ele não estará bem adaptado ao novo hospedeiro e, com frequência, será patogênico. O hospedeiro intermediário normalmente serve como fonte zoonótica de infecção humana e desempenha o papel de um hospedeiro amplificador, o que permite que o vírus se replique transitoriamente e, em seguida, seja transmitido aos seres humanos para ampliar a escala da infecção humana (YE et al., 2020). Em uma outra possibilidade, o vírus também pode se adaptar ao hospedeiro intermediário e até mesmo estabelecer uma endemia de longo prazo, nesse caso, o hospedeiro intermediário se torna um reservatório natural (YE et al., 2020).

Os HCoVs humanos podem ser divididos, conforme a gravidade de sintomas causados, em: coronavírus de alta patogenicidade (SARS-CoV, MERS-CoV e SARS-CoV-2) e de baixa patogenicidade ou comunitários (HCoV-229E, HCoV-OC43, HCoV-NL63 e HCoV-HKU1).

1.3.1 Coronavírus humanos de alta patogenicidade: SARS-CoV, MERS-CoV e SARS-CoV-2

O SARS-CoV, MERS-CoV e SARS-CoV-2 são os coronavírus que tiveram maior impacto negativo na saúde humana por terem sido causa de epidemias e uma pandemia (no caso do SARS-CoV-2) preocupantes, com altos números de internações e mortes. No caso do SARS-CoV, quando incialmente identificado na China, em 2002, vários estudos subsequentes foram realizados para se investigar a sua origem, afinal, tratava-se do primeiro coronavírus com potencial de causar mortes e crises significativas no sistema de saúde.

O SARS-CoV ou anticorpos anti-SARS-CoV foram encontrados em civetas de palmeira mascaradas (*Paguma larvata*) em um mercado na província de Guandong, China, enquanto os mesmos organismos de vida selvagem, quando capturados em outras regiões, eram negativos (CUI; LI; SHI, 2019; SU et al., 2016). Esses achados levaram à conclusão de que as civetas eram apenas hospedeiros intermediários e não reservatórios dos vírus (SU et al., 2016). Além disso, posteriormente, foram descobertos vários coronavírus similares ou relacionados ao SARS (SARSr-CoV) em morcegos de ferradura (do gênero *Rhinolophus*) da China (inclusive de mercados), da Europa e do sudeste da Ásia (CUI; LI; SHI, 2019; SU et al., 2016). Uma interessante descoberta foi também a presença de diversos SARSr-CoVs coexistindo em populações de morcegos de uma caverna em Yunnan, província da China, onde as linhagens virais apresentavam uma alta diversidade genética (CUI; LI; SHI, 2019).

Embora ainda não se saiba com certeza a origem do SARS, todas essas informações levaram à suposição de que ele tenha se originado em morcegos e pode ter apresentado algum hospedeiro intermediário antes de apresentar o padrão de infecção de pessoa para pessoa (**Figura 6**). Essa característica de transmissão remete ao chamado “salto entre espécies” que provavelmente permitiu o surgimento do SARS em humanos (NG; TAN, 2017).

Especula-se o mesmo para o mais recente SARS-CoV-2, causador da COVID-19 (*Coronavirus Disease 2019*/ doença causada por coronavírus 2019) que foi anunciada como pandemia pela Organização Mundial da Saúde (OMS) em 11 de março de 2020. Segundo Indranil Chakraborty e Prasenjit Maity (2020), foi a sexta emergência de saúde pública internacional declarada pela organização depois da H1N1 em 2009, poliomielite em 2014, Ebola em 2014 (na África Ocidental), Zika em 2016 e Ebola novamente em 2019 (Democrata República do Congo).

Ela foi identificada pela primeira vez em dezembro de 2019, na cidade de Wuhan, na província de Hubei, China, através de pacientes com sintomas de pneumonia de etiologia desconhecida e que haviam frequentado o mercado de frutos do mar de Hunan, conhecido por vender animais silvestres vivos, como morcegos, aves e outros (CHAKRABORTY; MAITY, 2020; JIANG et al., 2020). O novo coronavírus foi identificado a partir de uma amostra de lavado bronco-alveolar de um paciente que trabalhava no mercado de Hunan, através de sequenciamento de nova geração, e a análise filogenética do seu genoma demonstrou que havia 89.1% de similaridade de nucleotídeos com coronavírus do tipo SARS, do gênero *Betacoronavirus* (WU et al., 2020; YANG; WANG, 2020). O novo vírus foi então nomeado SARS-CoV-2 (do inglês, S*evere Acute Respiratory Syndrome Coronavirus 2*) pelo ICTV em 11 de fevereiro de 2020 (Y et al., 2020). Até agosto de 2023, a doença havia ocasionado 6.956.173 milhões de mortes, com cerca de 770.085.713 milhões de casos confirmados em todo o mundo, segundo dados da OMS (https://covid19.who.int/table).

Análises de similaridade de sequência genômica revelaram que esse vírus apresenta 96.2% de identidade de sequência com um CoV de morcego BatCoV RaTG13 (ZHOU et al., 2020a), e cerca de 89% de similaridade de sequência com outros dois coronavírus também derivados de morcegos e relacionados ao SARS, o bat-SL-CoVZC45 e o bat-SL-CoVZXC21, enquanto ele parece ser um pouco mais distinto em relação ao SARS, com 80% de similaridade, e ao MERS, com quase 51.8% (DHAMA et al., 2020). Análises filogenéticas também confirmaram uma maior proximidade do SARS-CoV-2 com linhagens relacionadas a SARS presentes em morcegos (DHAMA et al., 2020). Embora ainda sejam necessários estudos adicionais para confirmar possíveis reservatórios do SARS-CoV-2, alguns pesquisadores identificaram em coronavírus isolados do pangolim-malaio (*Manis javanica*) uma altíssima identidade de sequências de aminoácidos para as proteínas estruturais desse vírus (proteína E do envelope: 100%, M da membrana: 98.2%, N do nucleocapsídeo: 96.7%, e S Spike: 90.4%.), incluindo a diferença de apenas um aminoácido para o domínio de ligação ao receptor da

proteína Spike (DHAMA et al., 2020). Além disso, os genomas de coronavírus do pangolim e de morcegos também foram comparados, sugerindo a possibilidade de recombinação gênica entre eles (**Figura 6**) (DHAMA et al., 2020).

O MERS-CoV foi identificado em 2012, a partir de amostras de um paciente com pneumonia aguda na Arábia Saudita, e análises filogenéticas agrupam esse vírus aos coronavírus CoV-HKU4, de morcegos *Tylonycteris*, e CoV-HKU5, de morcegos *Pipistrellus* (S et al., 2012; TANG; LIU; CHEN, 2022; YE et al., 2020). Esse vírus também apresenta possível origem em morcegos, com a identificação de diversos vírus relacionados (MERSr-CoV), como o Ii-MERSr-CoV, Ve-MERSr-CoV e Hy-MERSr-CoV (China), que compartilham até cerca de 85% de identidade de sequência genômica com os MERS-CoVs de humanos e camelos (CUI; LI; SHI, 2019). O MERS-CoV foi encontrado em dromedários no Oriente Médio, dessa forma, compreende-se que esses animais possam ser reservatórios do vírus, embora ainda não se compreenda muito bem como se deu a transmissão para os humanos ou se houveram outros hospedeiros envolvidos (**Figura 6**) (JF et al., 2015; TANG; LIU; CHEN, 2022; YE et al., 2020)

1.3.2 Coronavírus humanos de baixa patogenicidade: HCoV-229E, HCoV-OC43, HCoV-NL63 e HCoV-HKU1

O HCoV-229E foi isolado de um paciente com infecção respiratória e descoberto em 1966, enquanto que em 2007, na Califórnia, foi descoberto o Alpaca Coronavírus (ACoV), que continha 92,2% de identidade de nucleotídeos com o genoma do HCoV-229E (CROSSLEY et al., 2012; TANG; LIU; CHEN, 2022). Estudos posteriores também demonstraram que esse vírus apresentava parentesco molecular com morcegos, estabelecendo as alpacas como hospedeiros intermediários (**Figura 6**) (CORMAN et al., 2015; KESHEH et al., 2022; TANG; LIU; CHEN, 2022). O HCoV-229E apresenta 63% de semelhança genômica com o HCoV-NL63, de quem é mais próximo segundo análises filogenéticas e este também apresenta provável origem em morcegos (CROSSLEY et al., 2012; KESHEH et al., 2022; TANG; LIU; CHEN, 2022; YE et al., 2020). Evidências apontam que o HCoV-229E apresenta parentesco genético com um CoV de morcego denominado Hipposideros/GhanaKwam/19/2008, detectado em Gana (CORMAN et al., 2015; HUYNH et al., 2012), enquanto o HCoV-NL63 está

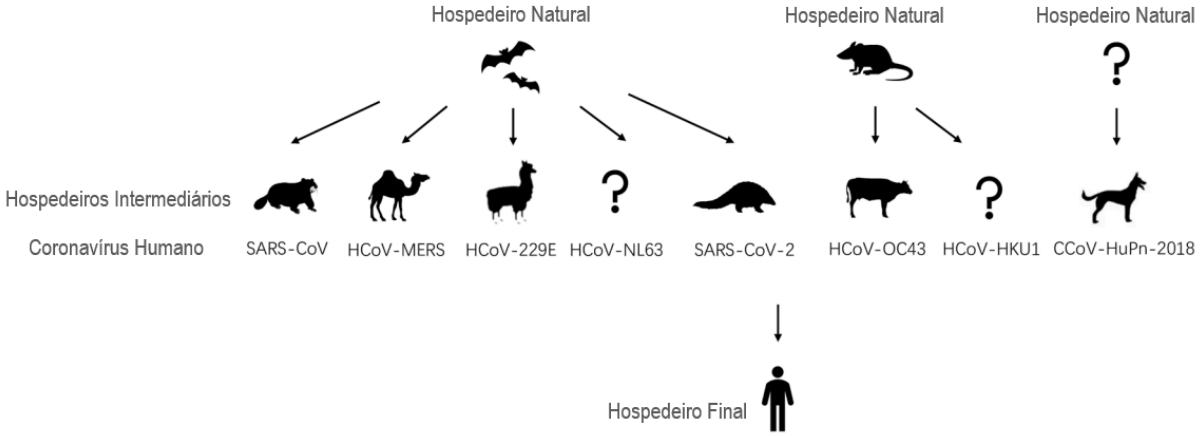
relacionado ao ARCoV.2 (*Appalachian Ridge CoV*), detectado em um morcego tricolor norte-americano (HUYNH et al., 2012; YE et al., 2020).

O HCoV-NL63 foi descoberto oficialmente em 2004 em amostras clínicas de uma criança na Holanda (KESHEH et al., 2022; VAN DER HOEK et al., 2004). No entanto, por se tratar de um vírus não emergente, com circulação contínua entre os humanos estimada em séculos (PYRC et al., 2006), sequências do HCoV-NL63 de antes de sua descoberta oficial estão disponíveis em bancos de dados públicos. Esse vírus também compartilha o mesmo receptor que o SARS-CoV e o SARS-CoV-2, a proteína ACE2 (enzima conversora de angiotensina 2) (HOFMANN et al., 2005; KESHEH et al., 2022). Um trabalho de vigilância em morcegos no Quênia também foi capaz de demonstrar que HCoV-NL63 humano é produto de eventos de recombinação entre vírus semelhantes ao NL63 que circulam nos morcegos *Triaenops* e vírus semelhantes ao 229E que circulam nos morcegos *Hipposideros*, com regiões de quebra localizadas próximo às extremidades 5' e 3' do gene S da proteína Spike (**Figura 6**) (TAO et al., 2017). O mesmo grupo também identificou outros eventos de recombinação interespécies envolvendo o gene S, de forma a defini-lo como uma região "*hotspot*" para recombinações (TANG; LIU; CHEN, 2022; TAO et al., 2017). O mesmo gene também já foi implicado em eventos de recombinação para o SARS-CoV e vírus SARS relacionados, e também entre cepas de HCoV-OC43 (CC et al., 2008; LAU et al., 2011; SK et al., 2010; TANG; LIU; CHEN, 2022).

O HCoV-OC43 foi identificado em 1967 (KESHEH et al., 2022; LAU et al., 2011) e análises filogenéticas e moleculares apontam que ele apresenta maior similaridade com o coronavírus bovino BCoV, provavelmente compartilhando um ancestral por volta de 1890, antes de se adaptar aos humanos (**Figura 6**). Uma evidência que dá suporte a essa hipótese é a ausência de uma sequência de 290 nucleotídeos no HCoV-OC43 quando comparado ao BCoV ou a outros vírus próximos, o vírus da hepatite murina (MHV) e no vírus da sialodacrioadenite de ratos (SDAV) (TANG; LIU; CHEN, 2022; VIJGEN et al., 2005). O HCoV-229E e o HCoV-OC43 causam cerca de 15 a 29% dos quadros gripais comuns e já são bastante estudados (MONTO, 1974; SU et al., 2016). Além disso, o HCoV-229E, o HCoV-OC43 e o HCoV-NL63 são distribuídos globalmente e apresentam padrão de transmissão aumentada durante o inverno em países de clima temperado, com o HCoV-NL63 também podendo apresentar maior transmissão no período de primavera-verão (CHIU et al., 2005, p. 63; HENDLEY; FISHBURNE; GWALTNEY, 1972; SU et al., 2016).

O HCoV-HKU1 foi primeiro identificado em Hong Kong em 2005 (KESHEH et al., 2022; TANG; LIU; CHEN, 2022; VABRET et al., 2006), e apresenta evidências de que se originou em roedores, porém, assim como o HCoV-NL63, ainda não se sabe da existência de hospedeiros intermediários ou reservatórios, e como se deu a transmissão até os humanos (KESHEH et al., 2022; MULABBI; TWEYONGYERE; BYARUGABA, 2021; TANG; LIU; CHEN, 2022). A **Figura 6** representa um resumo das origens zoonóticas dos HCoVs.

Figura 6: Origens dos Coronavírus Humanos



Acredita-se que o surto epidêmico SARS Antes da SARS surgiu quando vírus de morcegos infectaram civetas e então evoluíram para se adaptar aos humanos. O MERS-CoV parece ter se originado de morcegos e se espalhado para camelos há cerca de 30 anos, e tem sido prevalente em camelos dromedários. O HCoV-NL63 e HCoV-229E não causam doenças respiratórias graves e o ancestral de ambos foi encontrado em morcegos africanos. Os camelídeos podem ser o hospedeiro intermediário do HCoV-229E. Como os genomas do coronavírus do pangolim malaio têm alta similaridade de sequência com o SARS-CoV-2, eles poderiam ser considerados o possível hospedeiro intermediário dessa espécie de vírus. HCOV-OC43 e HCOV-HKU1 podem ser originados de roedores. Os humanos são os hospedeiros finais da transmissão e as setas pretas indicam a direção da propagação. Fonte: Adaptado de TANG; LIU; CHEN, 2022.

1.4 IMPORTÂNCIA DE ESTUDOS RELACIONADOS AOS CORONAVÍRUS

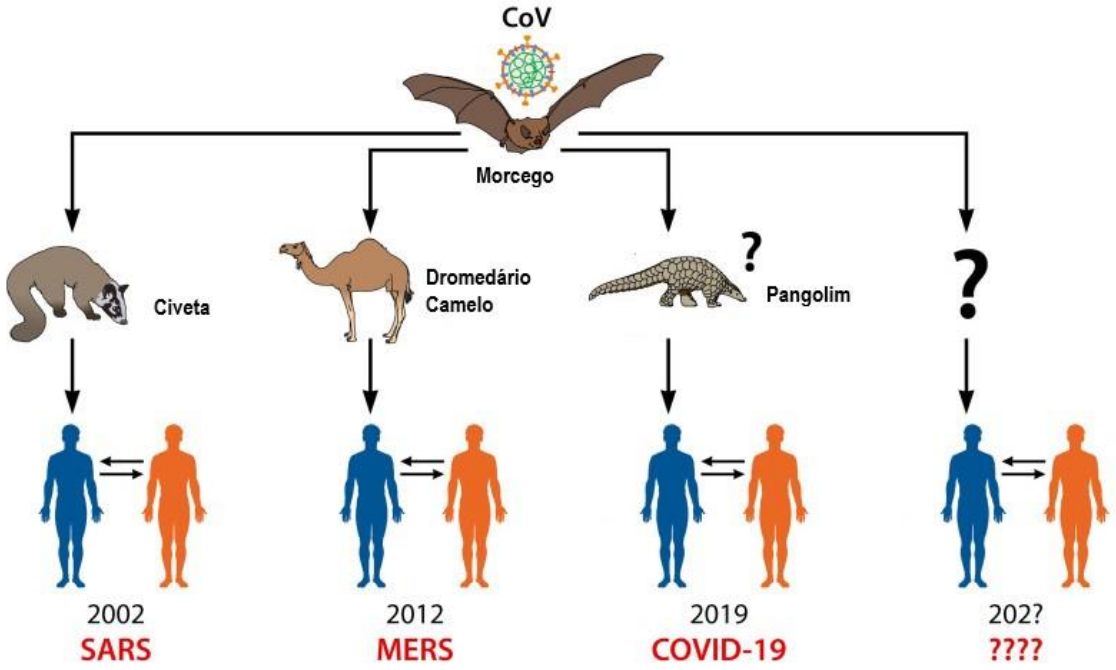
Antes do surgimento do SARS-CoV-2 e da ocorrência da COVID-19 como uma emergência de saúde pública mundial, vários pesquisadores e virologistas já haviam publicado trabalhos sugerindo que o próximo CoV a afetar os seres humanos era apenas uma questão de tempo. O artigo de Shuo Su et al. (SU et al., 2016) foi um deles. Oi-Wing Ng e Yee-Joo Tan (NG; TAN, 2017), Patrick C. Y. Woo et al. (WOO et al., 2009) e Vincent C. C. Cheng et al. (CHENG et al., 2007), foram outros grupos a fazer o alerta.

Embora previsível, a atual pandemia do SARS-CoV-2 ainda apresenta várias incertezas em relação à sua origem, possíveis hospedeiros iniciais e intermediários, e à base molecular e imunológica da patogênese da COVID-19. No entanto, sem dúvidas, a mobilização mundial da comunidade científica no enfrentamento do vírus emergente foi essencial para a rápida identificação do novo patógeno, esclarecimento dos meios de transmissão e elaboração de políticas públicas correspondentes, além da obtenção de alguns conhecimentos sobre a base molecular do vírus e de sua patogênese para a criação de vacinas e terapias alternativas.

A prevenção e o controle de novos surtos e doenças emergentes causadas por CoVs depende de estudos continuados para melhor esclarecimento da replicação, dinâmica de transmissão e patogênese em humanos. Além disso, alguns pesquisadores já especulam sobre o surgimento de novos HCoVs devido a alterações climáticas e ecológicas que possam favorecer a interação ou exposição dos humanos a animais (DHAMA et al., 2020). A **Figura 7** apresenta uma progressão temporal resumida da dinâmica de transmissão dos principais HCoVs com importância médica conhecidos.

Considerando que os CoVs são vírus zoonóticos e a possível origem do SARS-CoV e SARS-CoV-2 envolvendo o contato direto e o consumo de animais silvestres em mercados chineses, torna-se crucial investigar a prevalência de CoVs em circulação em populações humanas e animais, a ocorrência de recombinação entre CoVs existentes, potenciais hospedeiros desconhecidos e, principalmente, estabelecer uma rede de monitoramento com o potencial de prever a emergência de novos CoVs recombinantes de alta virulência a partir de animais (SU et al., 2016). Além disso, ainda existem muitas lacunas de informação sobre os HCoVs conhecidos e mesmo os de baixa patogenicidade não são frequentemente diagnosticados e monitorados.

Figura 7: Progressão Temporal e Dinâmica de Transmissão dos Principais Surtos Respiratórios Causados por Coronavírus Humanos



Os coronavírus são exemplos de vírus emergentes que atravessaram a barreira entre espécies de animais silvestres para humanos, como SARS e o MERS. A origem do SARS-CoV-2 também é suspeita de ter ocorrido através de um hospedeiro intermediário. A possibilidade de o salto entre espécies acontecer uma quarta vez não pode ser dispensada. Fonte: Adaptado de Dhama, Kuldeep et al., 2020.

1.5 ANÁLISES MOLECULARES EVOLUTIVAS EM CORONAVÍRUS HUMANOS

A filogenética é uma grande área da biologia que contempla os processos evolutivos de uma ou múltiplas espécies e suas relações entre si, ao longo de gerações. A representação gráfica dessa trajetória é denominada árvore filogenética e é constituída por nodos e ramos. As extremidades mais externas dos ramos, o que seria análogo às “folhas”, são os nodos externos e eles podem representar espécies ou sequências (DNA, RNA ou proteína). As árvores filogenéticas podem ser baseadas em análises morfológicas e/ou moleculares, embora estas últimas tenham sido bastante exploradas devido ao avanço das plataformas de sequenciamento e da biologia computacional. Diante disso, diversos tipos de árvores podem ser obtidos dependendo das sequências utilizadas (DNA, RNA ou proteína; genomas completos, genes, fragmentos de genes ou conjuntos de genes), da pergunta/hipótese evolutiva, e do método utilizado. Para isso, diversas ferramentas e modelos probabilísticos foram desenvolvidos para melhor caracterizar as possibilidades de cada situação biológica/espécie/sequência (HALL, 2017; LEMEY; SALEMI; VANDAMME, 2009).

Segundo Barry G. Hall (2017), existem basicamente duas estratégias para se obter uma árvore filogenética: a algorítmica e a "busca de árvore". Enquanto a primeira abordagem utiliza um algoritmo específico para obter a árvore com base nos dados fornecidos, é um método rápido, e resulta em apenas um modelo de árvore; a segunda usa critérios de decisão para obter o melhor modelo, é um pouco mais lento (embora o atual poder computacional disponível torne essa diferença desprezível), e resulta na análise de diversos modelos (HALL, 2017).

Os métodos algorítmicos são: a junção de vizinhos (JV, ou do inglês, NJ, *Neighbour-Joining*), muito mais comum atualmente, e método de agrupamento por pares não ponderados com média aritmética (APNPMA ou do inglês, Unweighted Pair-Group Method with Arithmetic Mean, UPGMA) (HALL, 2017). Ambos são considerados "métodos de distância" por se basearem em uma matriz de distância criada com base nas diferenças entre pares de sequências do arquivo de alinhamento múltiplo. No entanto, a diferença entre eles é que o UPGMA utiliza agrupamentos ou "clustering", e assume que a árvore é ultramétrica, ou seja, que a distância da raiz a qualquer extremidade/ponta é a mesma para todas as extremidades, enquanto o NJ calcula diretamente as distâncias das extremidades para os nodos (HALL, 2017). Os métodos de busca de árvore são: Parcimônia (Par), Máxima Verossimilhança (MV) e Inferência Bayesiana (IB). Tais métodos são denominados "métodos baseados em caracteres" por se basearem no alinhamento múltiplo das sequências, comparando os caracteres de cada coluna ou cada sítio/posição dos nucleotídeos/aminoácidos (HALL, 2017). Enquanto a Par procura obter a(s) árvore(s) com o mínimo de divergência entre elas, considerando-se um mesmo número de eventos, a MV utiliza modelos matemáticos evolutivos com base no que é mais provável de acontecer. Já a IB é uma variação da MV que considera maximizar a probabilidade de se obter determinada árvore com base no conjunto de dados fornecido (HALL, 2017).

Para que as análises filogenéticas sejam realizadas, é necessário obter antes o alinhamento múltiplo das sequências de estudo, pois as ferramentas filogenéticas utilizadas apenas reconhecem arquivos de alinhamento. Alguns dos programas mais comuns utilizados para alinhamentos são o MUSCLE (EDGAR, 2004), Clustal Omega (SIEVERS; HIGGINS, 2014) e o MAFFT (KATOH et al., 2002). O MUSCLE (do inglês, *Multiple Sequence Comparison by Log-Expectation*) é um alinhador progressivo que utiliza uma função denominada de escore de expectativa logarítmica para alinhar as sequências, ele é bastante rápido e pode ser utilizado para um grupo de médio ou grande de sequências, porém não é o mais adequado para sequências de proteínas que apresentem baixa homologia nas regiões N-

terminal (S1-NTD) e o C-terminal (EDGAR, 2004). Já o Clustal Omega é rápido e preciso devido aos algoritmos mBed (BLACKSHIELDS et al., 2010) e HMM (do inglês, *Hidden Markov Model*), respectivamente. É indicado para grupos muitos grandes de dados, adequado para sequências de proteínas que apresentem baixa homologia nas regiões N-terminal (S1-NTD) e o C-terminal, e não é a melhor opção para sequências com muitas inserções e deleções de bases (SIEVERS; HIGGINS, 2014). O MAFFT também é um alinhador progressivo, é uma ferramenta versátil com múltiplas estratégias de alinhamento, também é indicado para grupos muitos grandes de dados, adequado para sequências de proteínas que apresentem baixa homologia nas regiões N-terminal (S1-NTD) e o C-terminal, e ainda recomendado para sequências com lacunas (KATOH et al., 2002; "Which multiple alignment algorithm should I use?", 2023). Uma outra ferramenta de alinhamento chamada webPRANK (LÖYTYNOJA; GOLDMAN, 2010) também se destaca como sensível à janela de leitura ou sensível aos códons, trincas de nucleotídeos que correspondem aos aminoácidos.

Para as árvores filogenéticas, algumas das ferramentas mais utilizadas são o MEGA, IQ-Tree, MrBayes, PhyML. O MEGA é um programa com interface amigável que oferece várias opções para análises filogenéticas e evolutivas moleculares, desde o alinhamento a diferentes métodos para a montagem das árvores (TAMURA; STECHER; KUMAR, 2021). Desse modo, representa uma opção bastante completa e fácil de manusear. O IQ-Tree é voltado para análises de Máxima Verossimilhança (MV), apresenta uma variedade de métodos e modelos, além de ser fácil de utilizar e rápido (MINH et al., 2020). Já o MrBayes utiliza Inferência Bayesiana (IB) e o algoritmo Monte Carlo via Cadeia de Markov (MCCM ou, do inglês, *Markov Chain Monte Carlo, MCMC*) para obter as árvores filogenéticas (RONQUIST et al., 2012), e o PhyML, assim como o IQ-Tree, utiliza análises de Máxima Verossimilhança (MV) (GUINDON et al., 2005). FigTree ("FigTree", [s.d.]) e iTOL (LETUNIC; BORK, 2021) são as duas ferramentas utilizadas para desenhar e personalizar a representação gráfica das árvores.

No caso dos vírus, que não deixam registros fósseis, a única alternativa de estudar seu passado é por meio das relações filogenéticas entre os vírus já existentes (LEMEY; SALEMI; VANDAMME, 2009). E ainda, no caso dos coronavírus, as mutações esporádicas e os eventos de recombinação são os principais mecanismos envolvidos nos processos adaptativos-evolutivos (LAU et al., 2011; TAO et al., 2017; TEMMAM et al., 2022). Dessa forma, para os HCoVs, a maior parte dos trabalhos publicados costuma apresentar as análises filogenéticas como base para a classificação de grupos ou genótipos, com posteriores análises de relógio

molecular e caracterização das sequências usadas em relação a mutações, recombinações e alterações de aminoácidos.

Dois modelos de árvores são comumente utilizados para esses vírus: a junção de vizinhos (JV ou do inglês, NJ, *Neighbour-Joining*) e a Máxima Verossimilhança (MV). Os trabalhos costumam utilizar sequências de nucleotídeos (RNA) tanto do genoma completo, quanto de genes específicos, como a ORF1ab, a sequência da RpdR, e o gene S ou N, embora os genomas completos e as sequências do gene S sejam os mais utilizados, como foi feito em dois trabalhos recentes que classificaram sequências do HCoV-229E, HCoV-NL63, HCoV-OC43 e HCoV-HKU1 (SHAO et al., 2022; YE et al., 2023). Nesses trabalhos, cada um utilizou um método diferente (JV ou MV) e os vírus apresentaram as seguintes classificações de genótipos: HCoV-229E (1,2,3,4,5 e 6), HCoV-NL63 (A1, A2, A3, B, C1, C2 e C3), HCoV-OC43 (A, B, C, D, E, F, G, H, I, J e K) e HCoV-HKU1 (A, B e C), fora novas linhagens identificadas e ainda não nomeadas (SHAO et al., 2022; YE et al., 2023).

Um outro tipo específico de análise filogenética que representa tempos de divergência entre as amostras utilizadas é denominado relógio molecular ou datação filogenética. Os relógios moleculares são importantes para podermos estimar o tempo de divergência de uma espécie em relação a um ancestral (LEMEY; SALEMI; VANDAMME, 2009). Exemplos de programas mais utilizados para esse tipo de análise datada são o MEGA, BEAST e IQ-Tree. O MEGA oferece a opção de datação molecular baseada no método RelTime de Tamura et al. (2012) (TAMURA; STECHER; KUMAR, 2021), já o BEAST utiliza o algoritmo Monte Carlo via Cadeia de Markov (MCCM) (SUCHARD et al., 2018), e o IQ-Tree usa um modelo gaussiano relacionado ao relógio molecular Langley–Fitch (MINH et al., 2020; TO et al., 2016). Todas as ferramentas citadas apresentam várias opções de calibração e ajustes nas análises, e permitem que os modelos de relógio molecular sejam mais restritos, quando o modelo assume que todas as linhagens envolvidas apresentam similares taxas de alterações/substituições em suas sequências (processo de divergência constante), ou relaxados, quando as taxas de alterações/substituições das linhagens podem ser diferentes.

Sendo assim, informações sobre a possível data de surgimento de ancestrais dos HCoVs permitem uma melhor compreensão do processo evolutivo dessas espécies, a exemplo do HCoV-OC43 que compartilha um ancestral com o BCoV por volta de 1890 (VIJGEN et al., 2005). Outros trabalhos identificaram, em relação ao HCoV-NL63, um ancestral mais recente e comum a todos os genótipos, com surgimento em torno de 1970 (ROCHARS et al., 2017), e

também um possível ancestral do genótipo mais recente (C3) com aparecimento provável entre 2012-2013 (WANG et al., 2020).

Como as mutações/polimorfismos e as recombinações são eventos que contribuem para a diversidade viral, análises considerando esses fatores podem esclarecer aspectos do processo evolutivo. Dentro dessas análises, alguns conceitos são essenciais para se entender a relevância das alterações nas sequências investigadas. Com relação às mutações, elas podem ser primariamente classificadas como mutações sinônimas, quando as alterações de nucleotídeos não são capazes de causar alterações estruturais na proteína, ou não sinônimas quando afetam a estrutura das proteínas correspondentes. Além disso, do ponto de vista evolutivo, essas alterações também podem ser classificadas como contribuintes para os processos de seleção purificadora/negativa, quando causam danos estruturais e/ou funcionais e contribuem para a eliminação, ou de seleção positiva, quando são mais rapidamente fixadas na população por causarem algum benefício (LEMEY; SALEMI; VANDAMME, 2009). Em termos de análises moleculares virais, algumas ferramentas podem auxiliar na identificação desses eventos, como o BioAider (mutações) e o RDP5 (recombinação). O BioAider é capaz de identificar e sumarizar todas as mutações identificadas, tanto em nucleotídeos quanto proteínas, classificá-las em sinônimas e não sinônimas, e calcular suas frequências por sítio/posição. Além de trazer essas informações no formato de tabelas ou gráficos (ZHOU et al., 2020b). Já o RDP5 apresenta sete ferramentas com métodos diferentes para a identificação de possíveis eventos de recombinação (RDP, GENECONV, BootScan, MaxChi, Chimaera, SiScan e 3Seq) (MARTIN et al., 2021).

As análises de recombinação e de mutações são realizadas com mais frequência em relação ao gene S, dada a sua grande importância para a internalização do vírus na célula hospedeira. A proteína Spike é subdividida em dois subdomínios, S1 e S2, e o domínio de ligação ao receptor (DLR) está localizado dentro do S1 que, por abrigá-lo, tem um papel fundamental na determinação dos hospedeiros e do tropismo de células (LI, 2016; TANG et al., 2023). O domínio S1 pode ainda ser dividido entre o subdomínio N-terminal (S1-NTD) e o C-terminal (S1-CTD), de acordo com os grupamentos químicos de suas extremidades (grupamentos amina -NH2 ou carboxila -COOH, respectivamente), como representado pela proteína Spike do SARS-CoV-2 na **Figura 8**. Ambos S1-NTD e S1-CTD podem funcionar como DLR, a depender da espécie de coronavírus e da proteína ligante no hospedeiro (LI, 2016; TANG et al., 2023).

Retomando a informação de que eventos de recombinação interespécies envolvendo o gene S já foram identificados entre vírus ancestrais de morcegos do HCoV-229E e do HCoV-NL63 (TANG; LIU; CHEN, 2022; TAO et al., 2017. Além disso, Tang et al., 2023, através de cálculos envolvendo o dN (número de substituições não-sinônimas por sítio não-sinônimo) e o dS (número de substituições sinônimas por sítio sinônimo), demonstrou que o gene S, particularmente o subdomínio S1, está sob pressão positiva nos HCoVs.

Figura 8: Diagrama Esquemático da Proteína S do SARS-CoV-2

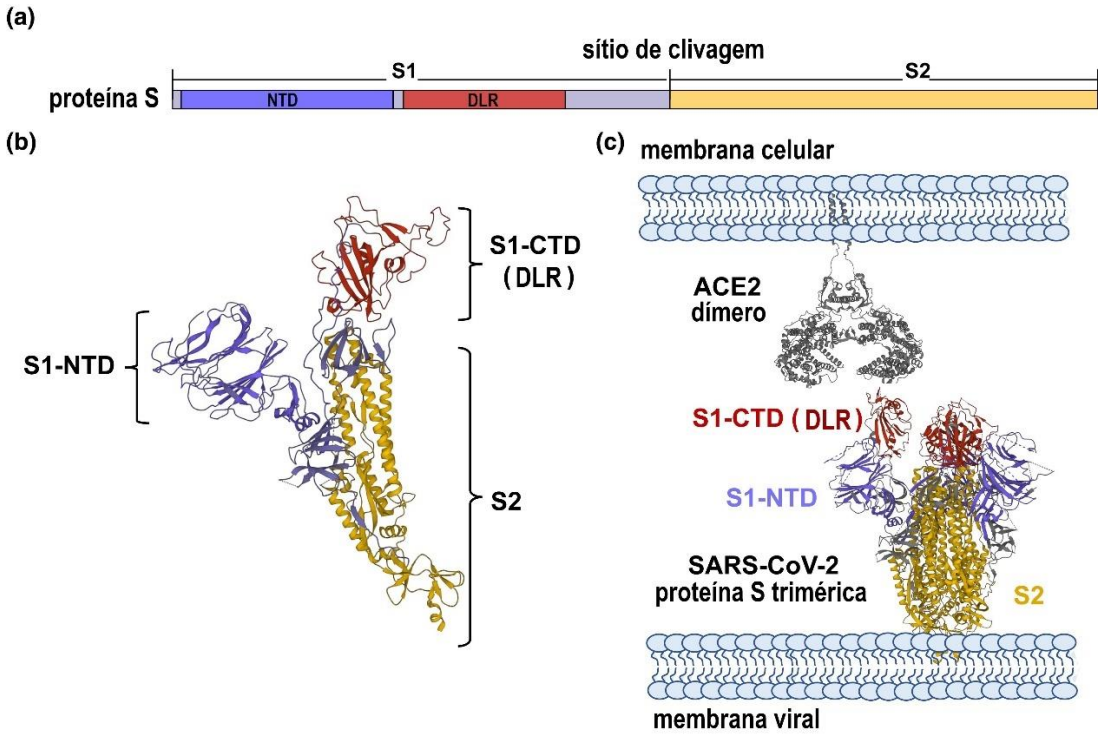


Diagrama esquemático das proteínas S do SARS-CoV-2. (a) Delineamento dos domínios baseado na proteína S do SARS-CoV-2. NTD, domínio N-terminal; DLR, domínio de ligação ao receptor. O domínio C-terminal (CTD) do SARS-CoV-2 funciona como RBD. (b) Estrutura monomérica da proteína S do SARS-CoV-2 (Wrobel et al. 2020). (c) A ligação entre o dímero ACE2 humano e a proteína S trimérica do SARS-CoV-2 (Cai et al. 2020). Fonte: Adaptado de TANG et al. 2023.

1.5.1 Análises moleculares evolutivas do HCoV-NL63

Considerando as características e dinâmica biológica dos coronavírus humanos, não é difícil concluir a importância de melhorar seu monitoramento epidemiológico. Especialmente porque mesmo sendo bastante estudados, os HCoVs de baixa patogenicidade não são frequentemente diagnosticados, sequenciados e acompanhados, apesar de amplamente

distribuídos no mundo. Um desses exemplos é o HCoV-NL63. O grupo de Van Der Hoek (2004) foi o primeiro a identificar e caracterizar o genoma do HCoV-NL63 em 2003, a partir de um aspirado da nasofaringe de uma criança na Holanda (KESHEH et al., 2022; VAN DER HOEK et al., 2004). Nesse trabalho, o genoma do NL63 foi descrito com estrutura conservada e semelhante à da família viral, incluindo uma proteína acessória entre as sequências dos genes S e E (correspondente à ORF3) que só apresentava semelhança com o 229E e o PEDV. Eles também utilizaram as sequências gênicas 1a, 1b, S, M e N para realizar análises filogenéticas junto a sequências de outras espécies de CoVs conhecidos e identificaram que o NL63 sempre ficava agrupado junto ao 229E. De forma semelhante, o grupo de Ron A. M. Fouchier (2004) também identificou o novo vírus e realizou análises filogenéticas com as sequências do genoma completo, do CRT 1ab, das proteínas Spike, E, M e N. Para isso eles utilizaram o método MV e ferramentas como o alinhador ClustalW e o software Phylip para montar a árvore (RA et al., 2004).

Com o vírus identificado, seu genoma sequenciado, estrutura esclarecida, testes diagnósticos propostos e genoma de referência estabelecido previamente, Nathalie Bastien e colaboradores (2005) realizaram um estudo mais amplo no Canadá. Eles utilizaram 13 amostras positivas de pacientes, além das sequências dos trabalhos anteriores, para gerar uma árvore filogenética do tipo NJ utilizando apenas sequências do gene 1a e o programa MEGA. Dessa forma, eles foram capazes de identificar dois grupos distintos do NL63 que não se correlacionavam com a geografia das amostras (Holanda/Canadá) (BASTIEN et al., 2005). Krzysztof Pyrc e seu grupo de pesquisa também sequenciaram e montaram dois genomas do HCoV-NL63 a partir de amostras de uma criança e de uma mulher idosa com sintomas respiratórios na Holanda. Nesse trabalho, eles identificaram duas regiões hipervariáveis: uma na parte 5′ do gene 1a que codifica as PNEs 1 a 3 (nucleotídeos -nt- 170-5000) e outra na parte 5′ do gene Spike (nt 20.300-22.000). Esta última englobava a região S1 e também apresentava uma sequência exclusiva no HCoV-NL63 quando comparada com vírus mais próximo, HCoV-229E, enquanto as regiões correspondentes ao gene 1b (da RpdR) e ORF3 se mantiveram bastante conservadas (PYRC et al., 2006). Eles ainda realizaram análises filogenéticas (ferramentas ClustalX e MEGA) utilizando diferentes regiões do gene 1a e a região N-terminal do gene da Spike, e encontraram muitas discrepâncias nos agrupamentos das sequências. Por isso, realizaram análises de variabilidade, de substituições sinônimas e não sinônimas (SimPlot, DnaSP e PAML), e concluíram que essas variações de classificação eram produto de recombinação e de mosaicos das sequências genômicas analisadas. Esse trabalho também foi o

primeiro a realizar um relógio molecular (software BEAST) e propor um ancestral comum entre HCoV-NL63 e HCoV-229E 900 anos antes (PYRC et al., 2006).

O trabalho de Katherine E. Arden e sua equipe sequenciou vários genomas do NL63 na Austrália e identificou que as sequências do gene 1b eram de fato as menos variáveis (99%), enquanto as do gene 1a permitiram o agrupamento filogenético das amostras em A (com subgrupos A1 e A2) e B (B1 e B2) (ARDEN et al., 2005). Chiu, Susan S. et al (2005), de forma semelhante, também utilizou sequências do gene 1a e identificaram dois grupos (A e B) ao classificar suas amostras de Hong Kong (utilizaram o MEGA) (CHIU et al., 2005). Um diferencial desse trabalho foi a identificação de circulação do HCoV-NL63 nos períodos de primavera e verão, enquanto todos os outros perceberam um maior número de infecções durante o período de inverno. Outros trabalhos também realizaram análises filogenéticas de novas sequências, no entanto, apenas as compararam a algumas sequências de referência e não classificaram os genótipos em A ou B. A. Koetz et al. investigou sequências S do vírus na Suécia (MEGA) (A et al., 2006). Astrid Vabret et al. trabalharam com sequências parciais do gene S e analisaram 12 amostras da França que demonstraram bastante heterogeneidade entre si (VABRET et al., 2005). C. Minosse et al. identificou o mesmo com mais 10 amostras da Itália e, ao analisar a filogenia das sequências da ORF1a e do gene S, observaram diferentes padrões de agrupamento. Esse grupo utilizou um número maior de sequências previamente identificadas em outros trabalhos para comparação mas também não as classificou em genótipos, embora tenha destacado que provavelmente diversos genótipos estariam em circulação em diferentes regiões do mundo, mais uma vez indicando uma falta de padrão geográfico de dispersão (C et al., 2008). Lili Ren et al. foi mais um dos trabalhos que identificou sequências na China e utilizou a ORF1ab para análises filogenéticas (REN et al., 2011).

O trabalho do grupo de Samuel R. Dominguez comparou as sequências de 16 genomas do vírus HCoV-NL63 de pacientes pediátricos do Colorado, nos Estados Unidos (EUA), com as sequências da Holanda. Nesse trabalho, eles identificaram três genótipos distintos através da filogenia dos genomas completos (A, B e C), além de sequências não agrupadas que foram denominadas de recombinantes pois apresentavam regiões de sequência similares aos genótipos B e C. Esse trabalho foi o primeiro a identificar o novo genótipo C (DOMINGUEZ et al., 2012). Além disso, ao realizarem a comparação entre sequências do gene da PNE 3 (nsp3) e do gene S, observaram que os genótipos A (NL63/DEN/2005/1876) e C (NL63/DEN/2009/20) eram semelhantes entre si até por volta do nucleotídeo (nt) 1030 da PNE 3, com o genótipo B (NL63/DEN/2009/14) exibindo divergência na mesma região. Após o nt 1030, os genótipos B

e C apresentavam sequências semelhantes, enquanto o A era mais divergente. No gene S, os genótipos A e C eram semelhantes até cerca do nt 920, enquanto do nt 920 ao 1710, os três genótipos apresentaram sequências relativamente divergentes. Após o nt 1710, os genótipos B e C apresentaram sequências relativamente semelhantes. Essa segregação de sequências de nucleotídeos foi mantida em relação aos aminoácidos na PNE 3 e somente para a primeira parte do gene S. Um outro achado importante foi que a região N-terminal do gene S (nt 1 a 600) se mostrou a mais variável (DOMINGUEZ et al., 2012).

A maior parte das publicações posteriores passou a consistentemente classificar as novas sequências em relação aos genótipos e considerar um maior número de amostras previamente identificadas para comparação. Rapeepun Soonnarong e colaboradores identificaram novas sequências do vírus NL63 na Tailândia e classificaram-nas como pertencentes ao grupo B, segundo árvore filogenética JV (MEGA) (SOONNARONG et al., 2016). Maryam Nabiel Al-Khannaq et al. trabalharam com o sequenciamento de amostras clínicas em Kuala Lumpur, na Malásia. Esse grupo também confirmou divergências entre os modelos filogenéticos utilizando os genes S, N e 1a, dessa forma, priorizaram as análises utilizando a sequência do subdomínio S1 do gene S, que demonstraram maior coerência em termos de agrupamento. A maior parte das sequências desse trabalho ficou distribuída entre os genótipos B e C, com algumas sequências não agrupadas classificadas como recombinantes. No entanto, o grupo B aparentou ter uma divisão que não foi mencionada de forma clara pelo trabalho. Eles também realizaram uma análise temporal indicando a possível origem do NL63 por volta de 1920, com surgimento do genótipo A em 1975, do B em 1995 e do C em 2002 (AL-KHANNAQ et al., 2016). O trabalho de Valery Madsen Beau De Rochars, identificou sequências do vírus NL63 no Haiti e classificou os genes S como pertencentes ao genótipo C, com agrupamento próximo a algumas das sequências da Malásia, com quem possivelmente compartilharam um ancestral comum por volta de 2013. Além disso, identificaram possível surgimento do NL63 com ancestral mais recente comum por volta de 1970, com as cepas mais antigas tendo origem nos EUA (ROCHARS et al., 2017).

Posteriormente, Su-Hua Huang et al. utilizaram sequências parciais do gene 1a para agrupar sequências de Taiwan, sem a classificação de genótipos estabelecida previamente (SH et al., 2017). Patience K Kiyuka et al. utilizaram tanto as sequências da Spike quanto dos genomas completos para classificar amostras do Quênia em relação a amostras globais. As linhagens foram divididas entre os genótipos A e B e eles ainda sugeriram novas classificações dos subgenótipos baseadas na Spike: A (A0, A1 e A2) e B (B0, B1 e B2). Além disso, eles

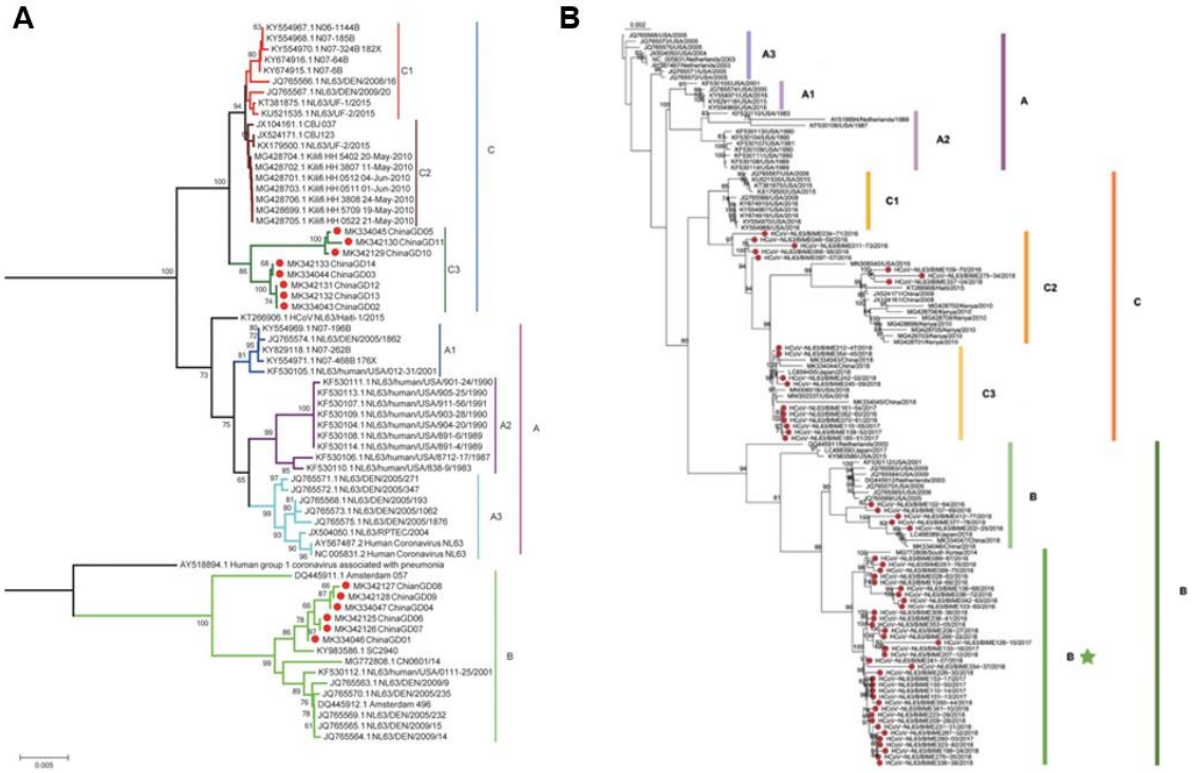
também pontuaram polimorfismos de aminoácidos dentro da região do DLR entre as cepas (A: I507L; B: E471D, I478V, P536A, G534V, H503R, E572Ae S576G) (KIYUKA et al., 2018). O trabalho de Lu Zhang et al. também contribuiu com 5 novas sequências do NL63 da China (ZHANG et al., 2020).

Trabalhando com amostras pediátricas de pacientes com HcoV-NL63 em Guangzhou na China, em 2018, Yanqun Wang e seu grupo utilizaram 53 genomas do NL63 de domínio público para classificar filogeneticamente novas amostras. Eles encontraram coerência de classificação quando compararam as estruturas dos modelos para os genomas completos, genes S e ORF1ab, classificando as novas sequências chinesas entre o grupo B (sem subdivisão) e um novo subgenótipo derivado do C que foi denominado C3 (utilizaram MAFFT e PhyML) (WANG et al., 2020). Além disso, eles realizaram análises de recombinação (software SimPlot) e concluíram que o novo subgrupo surgiu por meio de mutações e não recombinação. Os resultados do relógio molecular (BEAST) também indicaram que o C3 havia surgido entre 5 e 6 anos antes de sua identificação em 2012-2013, e eles ainda caracterizaram os polimorfismos de aminoácidos na Spike, destacando a mutação I507L no DLR como característica do C3 e potencialmente associada a um aumento de virulência ou transmissão (WANG et al., 2020). Em concordância com o trabalho de Wang, Yanjun Zhang et al. também identificaram novas amostras da China, da província de Zhejiang e do período de 2017 a 2019, como B e C com base na análise filogenética dos genes S (ZHANG et al., 2021).

Como estudos posteriores demonstraram que a presença de recombinação e de seleção purificadora afetam a datação molecular, Diego Forni et al. decidiram realizar análises de relógio molecular para os vírus 229E, OC43, HKU1 e NL63. Considerando todas as amostras disponíveis em bancos de dados públicos, eles utilizaram a ferramenta 3SEQ para identificar eventos de recombinação nos genomas e escolheram as maiores regiões possíveis não recombinantes. Para o NL63, essas regiões foram as menores e eles conseguiram estimar um ancestral comum datando de 50 anos antes. Apesar disso, o trabalho não conseguiu estimar o tempo de emergência do NL63 e do HKU1 por apresentarem hospedeiros intermediários desconhecidos (FORNI et al., 2022b). Ao analisar duas novas sequências do HCoV-NL63 na China, Nan Shao e seu grupo identificaram os genótipos B e C2, além de eventos de recombinação na região 1ab e que se estendiam da região do gene S ao N (SHAO et al., 2022). Em seu trabalho, Modeste Name Faye também identificou novas sequências do vírus no Senegal de 2012 a 2020 e classificou-as entre os genótipos A e B (FAYE et al., 2023).

Em um dos trabalhos mais recentes envolvendo o HcoV-NL63, Run-Ze Ye e colaboradores realizaram uma vigilância envolvendo diversos centros de pesquisa/da saúde na China entre 2016 e 2019. Nesse trabalho, eles obtiveram 58 novas sequências genômicas e suas análises filogenéticas destacaram um novo subgrupo dentro do genótipo B, ao qual muitas de suas amostras pertenciam. Eles o denominaram de novo grupo B ou grupo emergente B, dentro do qual identificaram até 31 substituições na região de 1 a 304 aminoácidos (aa) da proteína S. 26 dessas 31 substituições apareceram na região 1-201 aa do subgênero A2 em 1988, e as outras cinco na região 1-231 aa foram semelhantes às do subgênero C3 (YE et al., 2023). A **Figura 9** apresenta dois modelos de árvores filogenéticas dos trabalhos de Yanqun Wang, que identificou o subgrupo C3, e de Run-Ze Ye, que melhor identificou a divisão do grupo B.

Figura 9: Árvores filogenéticas S do vírus HCoV-NL63 dos trabalhos de Wang et al. e Ye et al.



A. Árvore filogenética das sequências parciais do gene S do vírus HcoV-NL63 pelo método JV. B. Árvore filogenética de genomas completos do vírus HcoV-NL63 pelo método MV. As novas amostras de cada trabalho estão marcadas por círculos vermelhos e os genótipos estão indicados por diferentes cores. O novo grupo B identificado no trabalho de Ye et al. está marcado pelo símbolo estrelado. Fonte: Adaptado de Wang et al. 2020 e Ye et al. 2023.

Todos os trabalhos acima citados contribuíram para gerar sequências virais revisadas do HCoV-NL63 que foram depositadas em bancos de dados públicos, com potencial para novas análises e melhor exploração dos dados disponíveis. Somando-se a isso, o HCoV-NL63: 1) não é frequentemente e amplamente diagnosticado ou sequenciado por causar sintomas leves; 2) compartilha o mesmo receptor (ACE2) com os vírus SARS-CoV e SARS-CoV-2; 3) já foi erroneamente diagnosticado como H1N1 2009 e SARS-CoV-2 (TK; N, 2012); 4) e ainda apresenta hospedeiro intermediário não identificado. Dessa forma, esse vírus apresenta ainda muitos pontos em sua história evolutiva a serem esclarecidos, além da sua contribuição para casos clínicos de sintomas gripais. Por isso, diante do que foi apresentado, o presente trabalho também tem a finalidade de realizar análises filogenéticas e de caracterização molecular de todas as sequências do gene S do vírus HCoV-NL63 presentes em bancos de dados públicos, de forma a contribuir para um melhor esclarecimento da dinâmica dessa espécie em relação à sua distribuição e características moleculares.

2 OBJETIVOS

2.1 OBJETIVO GERAL

Identificar sequências de coronavírus em metagenomas e transciptomas de hospedeiros, e realizar análises filogenéticas e de caracterização molecular dos genes S do vírus HCoV-NL63.

2.2 OBJETIVOS ESPECÍFICOS

2.2.1 Análises de Metagenomas e Transcriptomas

– Identificar as amostras que apresentam sequências de coronavírus em metagenomas e transcriptomas de hospedeiros e realizar a classificação taxonômica das sequências virais utilizando mais de uma ferramenta;

– Comparar os resultados entre as ferramentas utilizadas.

2.2.2 Análises Filogenéticas

– Fazer a árvore filogenética e a genotipagem das amostras do gene S do HCoV-NL63;

– Identificar e caracterizar mutações nas amostras do gene S;

– Caracterizar amostras quanto a eventos de recombinação no gene S;

– Realizar a datação filogenética das amostras em relação ao gene S.

1 MATERIAIS E MÉTODOS

1.1 Análises de Metagenomas e Transcriptomas

1.1.1 OBTENÇÃO DAS SEQUÊNCIAS DE TRABALHO

Amostras de DNA

A primeira etapa do trabalho consiste na obtenção das sequências ou amostras a serem utilizadas nas análises. Para tanto, primeiramente foi realizada uma pesquisa bibliográfica para identificar quais organismos já foram identificados como hospedeiros dos coronavírus. A pesquisa foi realizada no banco de dados PubMed (https://pubmed.ncbi.nlm.nih.gov/), através de palavras-chave incluindo os nomes das espécies de coronavírus já identificadas pelo Comitê Internacional de Taxonomia dos Vírus (https://talk.ictvonline.org/). Nessa busca, foram identificados centenas de hospedeiros, no entanto, pensando nas análises posteriores, foi necessário identificar apenas aqueles organismos que apresentavam genoma de referência ou *RefSeq*.

De todos os hospedeiros identificados através de artigos científicos (distribuídos em 58 famílias e 1 ordem), cerca de 812 espécies apresentavam produtos de sequenciamento de diferentes projetos depositados em bancos de dados públicos. No entanto, desse grupo, apenas 68 apresentavam genoma de referência (*RefSeq*) confirmado através da plataforma/banco de dados Ensembl (https://www.ensembl.org/index.html). Dessa forma, foi realizada a busca dos produtos de sequenciamento de alto rendimento desses 68 organismos, utilizando-se palavras-chave, no banco de dados público *Sequence Read Archive (SRA)* (https://www.ncbi.nlm.nih.gov/sra).

As palavras-chave utilizadas estão representadas no exemplo abaixo, referente à pesquisa para as amostras de humanos:

WGS[Strategy] AND METAGENOMIC[Source] AND RANDOM[Selection] AND "illumina" AND ("**Homo sapiens**"[Organism]) AND “filetype fastq”[Properties] AND cluster_public[prop]

Para todos os organismos contemplados na pesquisa, foram obtidas 24367 amostras de sequenciamento. Todas as vezes que alguma pesquisa não apresentava resultados, eram feitas tentativas diversas eliminando uma palavra/critério de busca por vez, até obtermos resultados. As amostras obtidas foram pesquisadas no período de 25/09/2020 a 01/02/2021.

Com o objetivo de reduzir o número total de amostras de humanos obtidas, a pesquisa foi realizada novamente com filtros/palavras-chave adicionais, segundo o exemplo abaixo:

WGS[Strategy] AND METAGENOMIC[Source] AND RANDOM[Selection] AND "illumina" AND ("Homo sapiens"[Organism]) AND “filetype fastq”[Properties] AND cluster_public[prop] NOT "ancient" NOT "hominid" NOT "remains" NOT "mummies" NOT "skeletal" NOT "archaeological" NOT "colonial" NOT "urine" NOT "feces"

A partir dos resultados obtidos, alguns projetos foram escolhidos manualmente, no entanto, como o número de amostras permaneceu alto (n = 4675), utilizamos uma estratégia de sorteio simples de amostras através da linguagem de computação R e escolhemos 1169 sequências aleatórias.

Amostras de RNA

Amostras de RNA também foram utilizadas para comparar os resultados obtidos. Para tanto, foi realizada uma pesquisa no SRA no dia 18 de setembro de 200 com as palavras-chave "Homo sapiens"[Organism] AND (“Virome”), e os seguintes filtros: public, RNA, paired, Illumina, fastq. A partir dos resultados obtidos, 13 projetos foram escolhidos manualmente, resultando em 605 amostras.

1.1.2 CONTROLE DE QUALIDADE

Com relação ao controle de qualidade das amostras obtidas, todas foram submetidas ao software fastp (v 0.20.1) (CHEN et al., 2018) ou fastQC (v 0.11.4) (BABRAHAM BIOINFORMATICS, [s.d.]), com parâmetros padrões e Phred score ≥ 20.

1.1.3 ATRIBUIÇÃO TAXONÔMICA

Para a identificação das amostras positivas para coronavírus foram utilizadas três ferramentas. A primeira foi o software Kaiju 1.7.4 (MENZEL; NG; KROGH, 2016) com dois bancos de dados diferentes: um específico para proteínas de coronavírus dos gêneros *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus* e *Gammacoronavirus*, e outro de proteínas não redundantes (nr) geral, ambos disponíveis no *National Center for Biotechnology Information Search database (NCBI)* (https://www.ncbi.nlm.nih.gov/public/). As análises foram realizadas com parâmetros padrões e com o critério –m "Minimum Match Length" igual a 14.

A segunda foi o programa Burrows-Wheeler Aligner (BWA) (LI; DURBIN, 2009), utilizando como referência um banco com sequências de nucleotídeos de coronavírus extraídas do NCBI vírus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/), contendo 65 (819) genomas completos de *Alphacoronavirus* (TaxID:693996), *Betacoronavirus* (TaxID:694002), *Deltacoronavirus* (TaxID:1159901) e *Gammacoronavirus* (TaxID:694013).

E a terceira foi a ferramenta Genome Detective (VILSKER et al., 2019), com a aplicação Virus Tool (v 2.43), um aplicativo disponível online (web-based), utilizando parâmetros padrões.

1.1.4 MONTAGEM DOS GENOMAS VIRAIS

A montagem dos genomas foi realizada segundo a junção de *reads*, obtenção de *contigs*, *contig binning* e reconstrução de genomas utilizando a ferramenta SPAdes v3.15.3 (BANKEVICH et al., 2012). O programa CAP3 (HUANG; MADAN, 1999) também foi utilizado para comparação/confirmação dos resultados.

1.1.5 SCRIPTS

Um conjunto de scripts foi desenvolvido para otimizar o fluxo de trabalho, e estão disponíveis em: https://github.com/giovannabioinfo/CoVsSearch_supplements.git

1.1.6 GRÁFICOS

Os gráficos que ilustram os dados quantitativos foram obtidos com o programa Graph Pad Prism (v6.0) software (CA, USA).

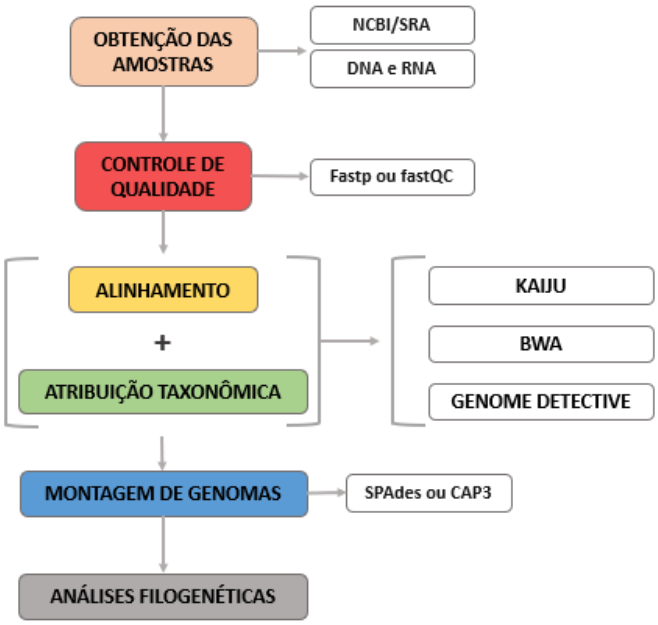


Figura 10: Fluxograma das Análises de Metagenomas e Metatranscriptomas

Fonte: Elaborado pela autora.

1.2 Análises Filogenéticas

1.2.1 AQUISIÇÃO DE SEQUÊNCIAS VIRAIS

Todas as sequências da proteína spike utilizadas neste trabalho foram obtidas do banco de dados de nucleotídeos do National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/nucleotide/) (“Home - Nucleotide - NCBI”, [s.d.]). A pesquisa foi realizada entre 17 e 19 de junho de 2023 e direcionada às sequências spike completas do

coronavírus humano NL63. Somente aquelas com 4068 a 4071 nucleotídeos e nenhuma base de nucleotídeo ambígua ou desconhecida foram consideradas.

1.2.2 ANÁLISES FILOGENÉTICAS

Todas as sequências foram alinhadas usando o muscle 3.8.31 (EDGAR, 2004) e os resultados foram verificados com o software MEGA 11 (TAMURA; STECHER; KUMAR, 2021). A árvore filogenética de máxima verossimilhança foi construída usando o método ModelFinder (KALYAANAMOORTHY et al., 2017) do IQ-TREE 2.1.3 (MINH et al., 2020) com 1.000 réplicas do UltrafastBootstraps (UFBoot) (HOANG et al., 2018) e editada com o programa iTOL (https://itol.embl.de/) (LETUNIC; BORK, 2021). Com o objetivo de confirmar os resultados, o alinhamento do muscle também foi utilizado para a obtenção de uma árvore filogenética pelo método Neighbour Joining utilizando a ferramenta RapidNJ (SIMONSEN; MAILUND; PEDERSEN, 2008) da plataforma Galaxy (https://galaxyproject.org/) (“Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update | Nucleic Acids Research | Oxford Academic”, [s.d.]). Com o auxílio do Microsoft Excel (“Software de planilha Microsoft Excel | Microsoft 365”, [s.d.]), as sequências também foram disponibilizadas segundo genótipo (resultado da árvore filogenética de máxima verossimilhança) e ano de coleta.

1.2.3 ANÁLISES EVOLUTIVAS E DE POLIMORFISMOS DE AMINOÁCIDO ÚNICO

Todas as sequências foram alinhadas usando o webPRANK do Instituto Europeu de Bioinformática (do inglês, *European Bioinformatics Institute* -EMBL-EBI- https://www.ebi.ac.uk/goldman-srv/webprank/) (LÖYTYNOJA; GOLDMAN, 2010) e os resultados verificados com o software MEGA 11 (TAMURA; STECHER; KUMAR, 2021). A análise de mutação foi realizada com o BioAider (V1.527) (ZHOU et al., 2020b) usando a amostra NC_005831.2 como referência. O número de substituições não sinônimas por sítio não sinônimo (dN), substituições sinônimas por sítio sinônimo (dS) e a razão dN/dS foram estimados usando os algoritmos MEME (MURRELL et al., 2012), SLAC (KOSAKOVSKY

POND; FROST, 2005) e FUBAR (MURRELL et al., 2013) no site Datamonkey (https://www.datamonkey.org) (WEAVER et al., 2018), com nível de significância de < 0,10 (MEME e SLAC) e > 0,90 (FUBAR). Os sítios sob seleção positiva foram confirmados por dois de três algoritmos (MEME, FUBAR e SLAC). A representação gráfica dos resultados de polimorfismos foi obtida com o auxílio dos softwares MEGA 11 (TAMURA; STECHER; KUMAR, 2021) e Microsoft Excel ("Software de planilha Microsoft Excel | Microsoft 365", [s.d.]).

1.2.4 ANÁLISES DE RECOMBINAÇÃO

Para detectar e caracterizar eventos recombinantes, todas as sequências foram alinhadas usando o webPRANK (do EMBL-EBI) (LÖYTYNOJA; GOLDMAN, 2010) e analisadas usando o RDP5 (MARTIN et al., 2021). Uma análise exploratória completa de sequências recombinantes foi realizada usando sete métodos diferentes disponíveis no RDP5 (RDP, GENECONV, BootScan, MaxChi, Chimaera, SiScan e 3Seq), e apenas os eventos positivos para quatro ou mais métodos foram considerados.

1.2.5 DATAÇÃO MOLECULAR E FILOGENÉTICA

Como a presença de eventos de recombinação pode afetar a datação molecular, todas as cepas recombinantes identificadas pelo RPD5 foram excluídas do nosso conjunto de dados. Um total de 120 sequências da spike consideradas não recombinantes foram alinhadas com o muscle 3.8.31, e a árvore temporal foi gerada usando o método IQ-TREE 2.1.3 least square dating (LSD) com as datas de coleta dos metadados (um modelo gaussiano relacionado ao relógio molecular Langley–Fitch) (TO et al., 2016). Para comparar, o mesmo foi realizado para todas as 173 sequências completas do gene da spike HCoV-NL63. Ambas as árvores foram editadas usando o FigTree 1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) ("FigTree", [s.d.]).

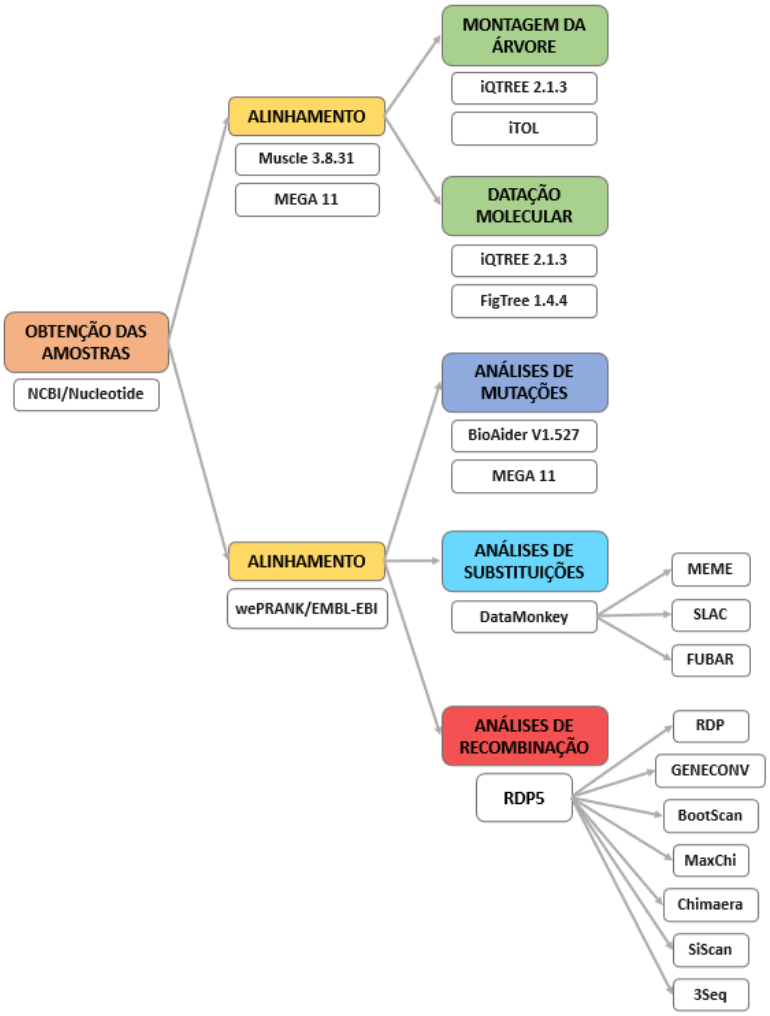


Figura 11: Fluxograma das Análises Filogenéticas

Fonte: Elaborado pela autora.

2 RESULTADOS

2.1 Análises de Metagenomas e Transcriptomas

2.1.1 OBTENÇÃO DE 25.031 SEQUÊNCIAS DE TRABALHO DE 68 ORGANISMOS

Após a realização da pesquisa de amostras, utilizando palavras-chave no banco de dados público SRA para todos os 68 organismos contemplados na pesquisa, foram obtidas 24367 amostras de sequenciamento, como apresentado na **Tabela 2**.

Tabela 2: Amostras de Hospedeiros dos Coronavírus

Hospedeiros de Coronavírus com Genoma de Referência					
Espécie		**Nº de Amostras**	**Espécie**		**Nº de Amostras**
Aves	*Anas platyrhynchos*	149	*Bovidae*	*Bos taurus*	3659
	Anser anser	136		*Bubalus bubalis*	158
	Anser brachyrhynchus	3		*Capra aegagrus*	5
	Anser cygnoides	30		*Capra hircus*	1354
	Athene cunicularia	137		*Ovis aries*	2279
	Aythya fuligula	1	*Camelidae*	*Camelus dromedarius*	63
	Cairina moschata	32	*Canidae*	*Canis lúpus*	1843
	Calidris pygmaea	23		*Lycaon pictus*	6
	Catharus ustulatus	11		*Vulpes vulpes*	45
	Chrysolophus pictus	12	*Cervidae*	*Cervus elaphus*	13
	Colinus virginianus	3		*Cervus hanglu yarkandensis*	4
	Columba livia	183		*Muntiacus reevesi*	1
	Coturnix coturnix	2	Cetacea	*Delphinapterus leucas*	45
	Coturnix japonica	35		*Tursiops aduncus*	9
	Cygnus olor	2	Chiroptera	*Chiroptera*	10
	Erithacus rubecula	3		*Myotis lucifugus*	13
	Erythrura gouldiae	24		*Myotis myotis*	2
	Gallus gallus	2756		*Pteropus vampyrus*	13
	Lonchura striata	5		*Rhinolophus ferrumequinum*	8
	Meleagris gallopavo	2	*Erinaceinae*	*Erinaceus europaeus*	3
	Melopsittacus undulatus	29	*Felidae*	*Acinonyx jubatus*	35
	Numida meleagris	124		*Felis catus*	273
	Passer domesticus	16		*Lynx canadensis*	1
	Pavo cristatus	2		*Panthera leo*	124
	Phasianus colchicus	16		*Panthera pardus*	12
	Pygoscelis adeliae	88		*Panthera tigris*	54
	Spheniscus humboldti	5		*Puma concolor*	41
	Spheniscus magellanicus	8	*Hominidae*	*Homo sapiens*	4011
	Strigops habroptila	1	*Moschidae*	*Moschus moschiferus*	7
	Taeniopygia guttata	6	*Muridae*	*Mus musculus*	3231
Bovidae	*Bison bison*	71		*Rattus norvegicus*	265
	Bos grunniens	278	*Mustelidae*	*Neovison vison*	20
	Bos indicus	395	*Soricinae*	*Sorex araneus*	2
	Bos mutus	6	*Suidae*	*Sus scrofa*	2164
Total: 68 espécies e 24367 amostras					

Fonte: Elaborada pela autora.

Como o número total de amostras de humanos obtidas ficou muito alto, uma nova pesquisa foi realizada para esse organismo utilizando-se filtros/palavras-chave adicionais. Após a segunda pesquisa realizada no SRA, foram selecionados 21 projetos apresentados na **Tabela 3**.

Tabela 3: Amostras de Humanos

Amostras de Humanos (*Homo sapiens*)			
Código de Identificação do Projeto	**Número de Amostras/ Experimentos**	**Tipo de Amostra**	**Lugar de Origem**
PRJEB38833	45	Fezes	China
PRJEB32767	66	Fezes	Israel
PRJNA707099	112	Tecido líquido: aspiração por agulha fina, peritoneal, lavagem broncoalveolar, pleural	Estados Unidos, Stanford
PRJNA682523	130	Saliva e biofilme subgengival	Hungria, Szeged
PRJEB41644	8	Swab da garganta e escarro induzido	Alemanha, Hanover
PRJNA678570	29	Líquido cerebrospinal	Estados Unidos, California, São Francisco
PRJEB37312	194	Nasofaríngea	África do Sul
PRJEB28058	164	Fezes	Alemanha
PRJEB27079	4	Escarro	Australia
PRJEB21872	66	Plasma	Alemanha
PRJEB30958	221	Plasma	Alemanha
PRJNA485882	1	Sangue	China
PRJNA385009	161	Plasma	Estados Unidos, Stanford
PRJNA385180	120	Plasma	Estados Unidos, Stanford
PRJEB9524	66	Fezes	Uganda
PRJEB8347	566	Fezes	Alemanha
PRJEB6070	1515	Fezes	Alemanha
PRJNA71831	38	Escarro, pulmões	USA, San Diego
PRJEB4530	8	Biópsias gástricas	Kuala Lumpur, Malásia
PRJEB1786	145	Fezes	Suécia, Gothenburg
PRJNA175224	11	Fezes	Estados Unidos, Harvard
Total:	3670 amostras/experimentos		

Fonte: Elaborada pela autora.

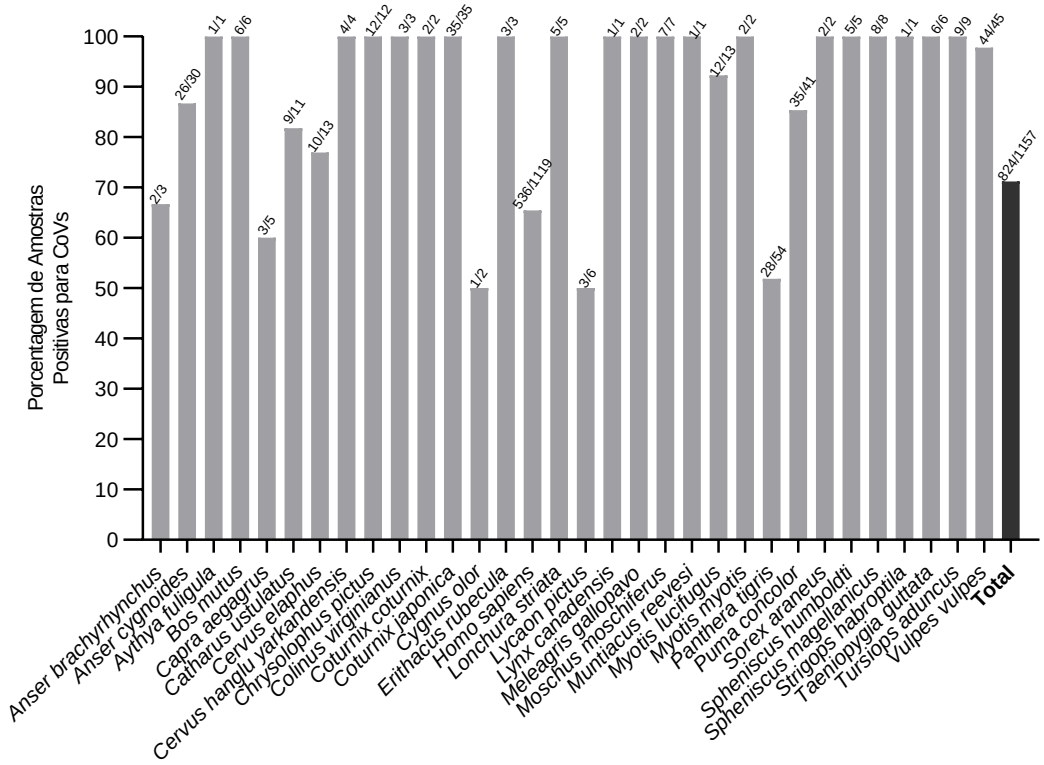
O número de amostras de humanos permaneceu alto (n = 4675), mesmo após a segunda pesquisa mais restritiva com palavras-chave adicionais, então utilizou-se uma estratégia de sorteio simples de amostras através da linguagem de computação R e 1169 sequências aleatórias foram escolhidas para as análises subsequentes.

2.1.2 TRIAGEM DE QUALIDADE E ATRIBUIÇÃO TAXONÔMICA POSITIVA COM O KAIJU PARA CORONAVÍRUS EM 836 AMOSTRAS DE 45 ORGANISMOS

Após a obtenção das amostras de trabalho, foi iniciou-se o processamento dos dados em relação ao controle de qualidade e a identificação da presença de sequências virais através da atribuição taxonômica. Todas as amostras passaram por correções e cortes de sequências de baixa qualidade pelo software fastp antes de serem submetidas ao software Kaiju para a atribuição taxonômica.

Num primeiro momento, o número de amostras positivas era bastante alto, bem próximo ao número total de amostras (em média 71,22%), como observado na Erro! Fonte de referência n ão encontrada.. Sendo assim, mais uma vez foi necessário mudar a estratégia de análise para tentar evitar resultados falso positivos.

Figura 12: Porcentagem de Amostras Positivas para Coronavírus Após Análise Inicial com o Kaiju



Acima de cada barra está representado o número de amostras positivas pelo número total para cada espécie. O gráfico foi obtido através de linha de comando utilizando a linguagem de programação R. Fonte: Elaborada pela autora.

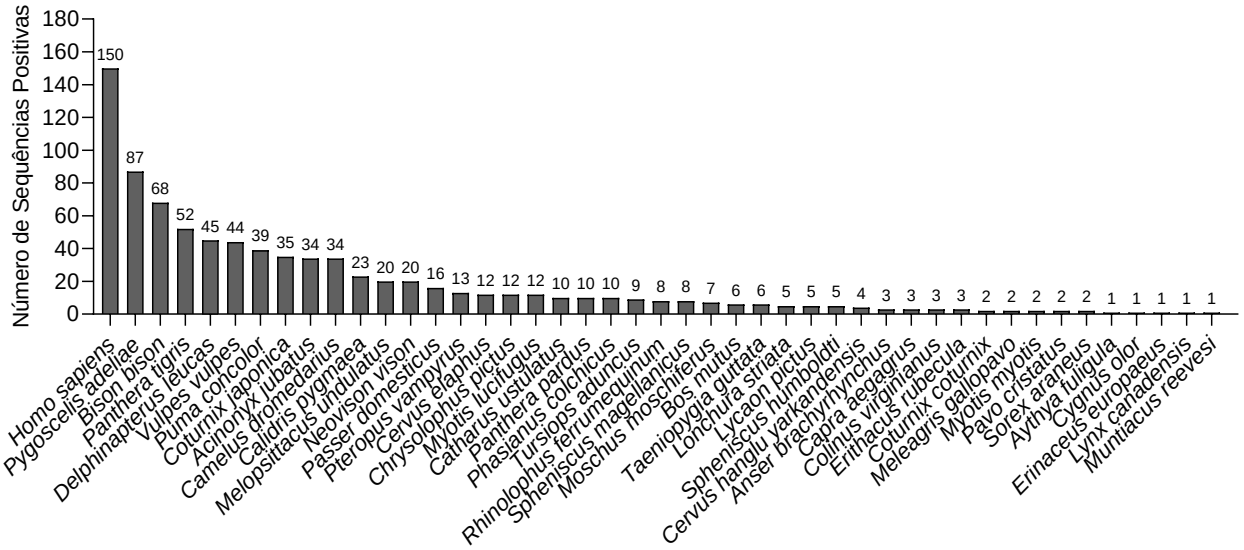
Com o objetivo de fazer uma dupla confirmação da atribuição taxonômica com o Kaiju, as sequências foram processadas em duas etapas: uma contra um banco de dados de proteínas dos vírus pertencentes à subfamília *Orthocoronavirinae*; e outra contra o banco de proteínas não redundante total do NCBI. Amostras de 45 diferentes organismos, incluindo os humanos, foram processadas. Os resultados dessa análise podem ser verificados na **Tabela 4** e na Erro! Fonte de referência não encontrada..

Tabela 4: Número de Amostras de Hospedeiros Positivas para os Coronavírus com o Kaiju

Número de Amostras de Hospedeiros Positivas para Coronavírus por Etapa de Classificação Taxonômica com o Kaiju		
	Amostras de Humanos	Amostras de Outras Espécies
Inicial	1169	752
Após 1ª Etapa	536	730
Após 2ª Etapa	150	686
TOTAL	836	

Fonte: Elaborada pela autora.

Figura 13: Distribuição de Sequências Positivas para Coronavírus Entre os Organismos Analisados com o Kaiju



Acima de cada barra está representado o número absoluto de amostras positivas após a análise em duas etapas. O gráfico foi obtido através de linha de comando utilizando a linguagem de programação R. Fonte: Elaborada pela autora.

Após a dupla confirmação, foram obtidas 836 amostras positivas para sequências de vírus da subfamília *Orthocoronavirinae*, das quais 96 são do tipo *single ended* e 740 *paired ended*. Após a triagem de amostras com a ferramenta Kaiju, foram realizadas várias tentativas de montagens dos genomas virais da subfamília *Orthocoronavirinae*. As amostras consideradas mais positivas, ou seja, com o maior número de sequências para coronavírus foram testadas, no entanto, a montagem dos genomas virais não foi possível e sempre resultava em sequências dos próprios organismos hospedeiros.

2.1.3 CONFIRMAÇÃO DA ATRIBUIÇÃO TAXONÔMICA E IDENTIFICAÇÃO DE FALSOS POSITIVOS EM TODAS AS AMOSTRAS INVESTIGADAS USANDO O BWA

Diante dos resultados obtidos, foi realizado um teste com uma ferramenta de alinhamento diferente, o BWA, que funciona promovendo o alinhamento de sequências de nucleotídeos dos metagenomas contra bases de referência obtidas do NCBI Virus. Para tanto, foi escolhida uma das amostras de humanos coletadas sabidamente positiva para um vírus conhecido (ID do projeto: PRJNA485882 – Virome of the blood sample from an HE patient China; amostra positiva para o GB vírus C – GBV-C). Isso foi necessário para eliminar a possibilidade de erro das ferramentas de montagem de genomas. Como resultado, as sequências do vírus GBV-C não apenas foram identificadas pelo BWA, como foi possível realizar a montagem do genoma viral utilizando o SPAdes.

Uma vez que o problema não estava na execução dos softwares de montagem, era necessário investigar melhor os resultados gerados pelo Kaiju. Como este software trabalha alinhando sequências de aminoácidos e todas as amostras coletadas para o projeto eram produtos de sequenciamento de DNA, foi utilizada a ferramenta BWA como estratégia alternativa, um alinhador de sequências de nucleotídeos. Foram analisadas 3642 amostras (do total de 3670) devido a alguns erros de download. Essas análises foram todas negativas para sequências de coronavírus, à exceção de uma. A amostra SRR606446 (BioProject PRJNA71831) apresentou duas sequências curtas positivas. Mesmo não sendo suficientes para montar genomas, é interessante destacar que essas sequências foram encontradas em amostras de escarro coletadas de indivíduos com fibrose cística. Elas foram identificadas através do BLAST (nt-nt) como "Human coronavirus NL63 strain Kilifi/IP/015/19-Nov-2012 spike protein S1 gene parcial cds" (cobertura de consulta 100% e identidade percentual 99,78%) e "Human coronavirus NL63 cepa ChinaGD01 genoma completo" (cobertura de consulta 99% e porcentagem de identidade 96,65%) (Erro! Fonte de referência não encontrada.). Uma das a mostras de humanos (ERR4181668) também foi positiva para *Streptococcus pneumoniae*. A Erro! Fonte de referência não encontrada. mostra o resultado dessa montagem.

Figura 14: Resultados dos alinhamentos para a amostra SRR606446 (BioProject PRJNA71831)

SRR606446.1819
TTTAGTGTTGTCAATGCTACCGTTACTGTTAATGTCACCACACTTAATGGCCGTATAGTTAATTACACTGTTTGTG
ATGATTGTAATGGTTACACTGATAACATATTTTCTGTTCAACAGGATGGCCGCATTCCTAATGGTTTTTCTTTTAA
TAATTGGTTTTTGTTAACTAATGGTTCTACACTAGTGGACGGTGTCTCTAGACTTTACCAACCACTTCGTTTAACT
TGTTTATGGCCTGTACCTGGTCTTAAATCTTCGACTGGTTTTGTTTATTTTAACGCCACTGGTTCTGATGTTAATTG
TAATGGTTATCAACATCATTCTGTTGCTGATGTTATGCGTTACAATCTTAACTTCAGTTCTAATTCTGTGGATAAT
CTCAAGAGTGGTGTTATAGTTTTTAAAACTCTACAGTACGATGTTTTGTTTACTGTAGTAATTCTTCTTC

SRR606446.14874
AGTCCATTGTATGTTGTAACATTTAGTAGTACTAAAGTAAACTGTTTTTGTTGTAACCTAAGGATGGTGGTTCAAT
TTTTNTTTCTGATGATTATCTTGGTATGTTGTAGATGACATTTATTATCCAGCTTCATGTAATGGTGTATTGCCAGT
TGCTTTTACAAAATTGGCAGGTGGTAAAATANCTTTTTCTGATGATGTTATAGTCCATGATGTTGAACCTACCCAT
AAAGTCAAGCTCATATTTGAGTTTGAAGATGATGTTGTTACCAGTCTTTGTAAGAAGAGTTTTGGTAAAGTCTAT
TATTTATACAGGTGATTGGGAAGGTTTACATGAAGTTCTTACATCTGCAATGAATGTCATTGGGCAACATATTAA
GTTGCCACAATTTTATATTTATGATGAAGAGGGTGGTTATGATGTTTCTAAACCAGTTATG

A figura apresenta duas sequências de DNA identificadas como sequências do coronavírus humano NL63 após análise com a ferramenta BWA. A primeira linha de cada sequência traz o código identificador da sequência e a posição do segmento separados por um ponto. Fonte: Elaborada pela autora.

Figura 15: Resultado da Montagem do Genoma do Organismo Streptococcus pneumoniae Encontrado em Uma das Amostras Analisadas

```
>Contig1
CCTTGTATCTAAAAGCGTCTTTCATCCGTGAGAATAGTTCCATGGAGGGAATATAGACCA
CCGAGCGCTTCAACTGGAACTTCTGGAAACTCTCGTTCAGTGTCTTAGCAATTCCAATTG
CTAGATGGCTCTTCCCAACTCCAGGCGGTCCGCTGATAATCGTATTCCCCTCATATCGCT
CCTTCACATAGTCAACCGTCACACGCTTGGCAAAATTGACCGCCTCAGCGTCTTGTTCTG
TGTGAATCTCAAAGTTACCGATA

>Contig2
GCACGTTCAATGGTTACAGAGTACGGTATGAGTGAAAAACTTGGCCCAGTACAATATGAA
GGAAACCATGCTATGTTTGGTGCACAGAATCCTCAAAAATCAATTTCAGAACAAACAGCT
TATGAAATTGATGAAGAGGTTCGTTCATTATTAAATGAGGCACGAAATAAAGCTGCTGAA
ATTATTCAGTCAAA

>Contig3
TGGTTAAGTAAAAGTCAACAGTTACTATAGTTTGCAAATTATTTAAAGAAAGAACAAAAA
TGCTCTCAAACACTGATTTCAAAGCATTTTTTGTTAGTTTAAAATTACTACCATTCTTCT
ATTCAAAACGTACAATATATCCAAAACCATTCAAAATAC
```

A figura apresenta três sequências de DNA em contigs formados e identificadas como sequências da espécie *Streptococcus pneumoniae*. As sequências foram obtidas através de alinhamento com a ferramenta BWA e as montagens em contigs realizadas com o auxílio das ferramentas SPAdes e CAP3. Fonte: Cortesia do Professor Dr. Jorge Estefano Santana de Souza.

Esses resultados são um indicativo de que provavelmente as amostras investigadas não apresentavam coronavírus (em sua maioria), apenas sequências conservadas e similares entre os organismos, o que foi superestimado pela ferramenta Kaiju.

2.1.4 ANÁLISES DE TRANSCRIPTOMAS UTILIZANDO A FERRAMENTA GENOME DETECTIVE

Com o objetivo de explorar as limitações do tipo de amostra utilizada e, com isso, também incluir mais uma ferramenta de classificação taxonômica para comparação, foi realizada uma nova pesquisa no banco do SRA voltada para amostras de RNA. A partir dessa pesquisa, 13 projetos foram escolhidos manualmente, resultando em um total de 605 amostras, como mostra a **Tabela 5**.

Tabela 5: Amostras de RNA de Humanos

Amostras de RNA de Humanos (*Homo sapiens*)			
Código de Identificação do Projeto	**Número de Amostras/ Experimentos**	**Tipo de Amostra**	**Lugar de Origem**
PRJEB30448	15	Placenta	Israel
PRJNA275568	96	Fezes	República Tcheca
PRJNA392272	47	Lavado broncoalveolar	Pensilvânia, EUA
PRJNA477357	17	Fezes	Pensilvânia, EUA
PRJNA526259	99	Sangue	São Francisco, EUA
PRJNA547643	16	Fezes	Sydney, Australia
PRJNA564995	82	Fezes	Pensilvânia, EUA
PRJNA588313	59	Fezes	Alemanha
PRJNA629087	21	Lavado broncoalveolar	Sydney, Australia
PRJNA641593	60	Fezes	Hohhot, Mongolia, China
PRJNA671738	58	Amostras clínicas de baixa biomassa	Nashville, Tennessee, EUA
PRJNA774620	7	Fezes	China
PRJNA794842	28	Sangue	Memphis, Tennessee, EUA
Total	605 amostras		

Fonte: Elaborada pela autora.

Quatro projetos foram analisados com a ferramenta Genome Detective Virus Tool: PRJNA794842, PRJNA774620, PRJNA671738 e PRJNA629087. Destes, apenas o projeto PRJNA671738 não foi concluído (apenas 35 processadas de 58) e apresentou 16 resultados positivos para coronavírus, conforme mostra a **Tabela 6**.

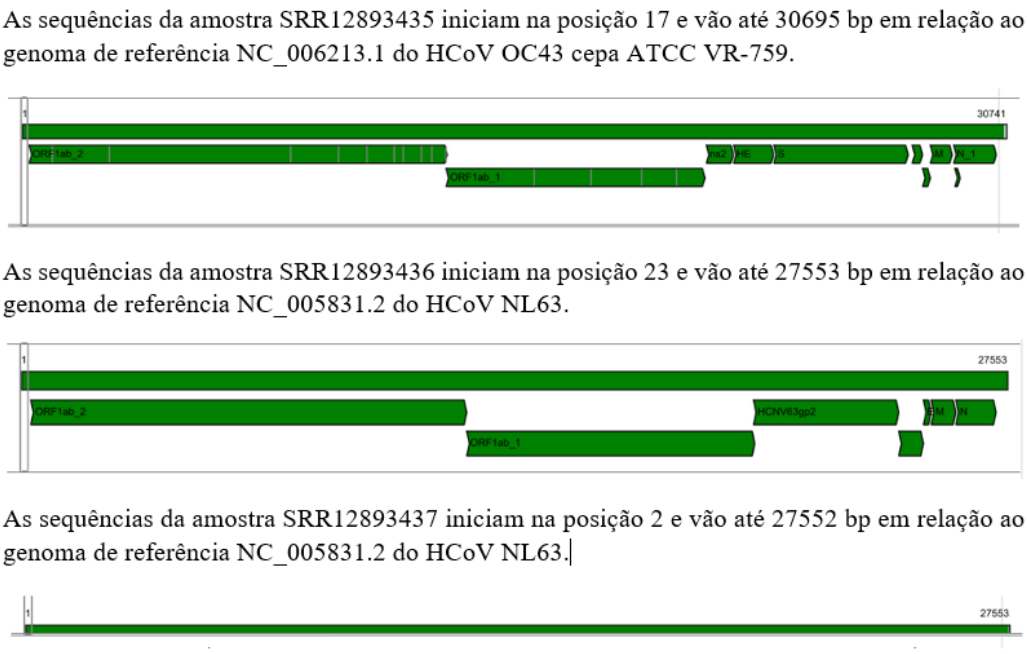
Tabela 6: Resultados Positivos para o Projeto PRJNA671738

Amostras do projeto PRJNA671738 positivas para CoVs							
Código da Amostra	**Espécie de CoV**	**Contigs**	**Reads**	**Cobertura (%)**	**Profundidade da Cobertura**	**Identidade (%)**	
						NT	**AA**
SRR12893435	BetaCoV OC43	1	~7000	99,79	~20	92,62	92,60
SRR12893436	HCoV NL63	1	7296	99,92	33,7	99,21	99,18
SRR12893437	HCoV NL63	1	~600000	99,99	~2000	99,01	99,00
SRR12904825	HCoV NL63	7	23	7,15	1,5	98,52	97,21
SRR12904830	HCoV NL63	6	10	3,48	1,3	98,33	96,52
SRR12904831	HCoV NL63	30	51	17,47	1,3	98,58	96,23
SRR12904844	HCoV NL63	37	74	22,41	1,5	98,62	98,12
SRR12904846	HCoV 229E	2	4	1,17	1,5	97,82	97,29
SRR12904847	BetaCoV OC43	2	2	0,87	0,9	99,62	98,88
SRR12904848	HCoV NL63	36	89	28,12	1,4	98,88	97,92
SRR12904858	HCoV NL63	29	189	37,61	2,2	98,71	98,23
SRR12904859	HCoV NL63	18	92	19,7	2,1	98,5	97,60
SRR12904862	HCoV NL63	17	29	10,1	1,3	98,1	97,20
SRR12904864	HCoV NL63	18	39	10,3	1,6	98,3	96,70
SRR12904865	HCoV NL63	4	10	3,3	1,4	98	97,60
SRR12904866	HCoV NL63	56	343	68,3	2,3	98,7	98

Fonte: Elaborada pela autora.

Desses 16 resultados, a maioria apresentou baixo número de reads e percentual de cobertura, o que impossibilitou a formação de um contig mais longo do genoma. As três primeiras amostras da **Tabela 6** (SRR12893435, SRR12893436 e SRR12893437) foram as exceções, pois essas permitiram a montagem quase completa do genoma viral correspondente Erro! Fonte de referência não encontrada..

Figura 16: Resultados dos Alinhamentos para as Amostras SRR12893435, SRR12893436 e SRR12893437



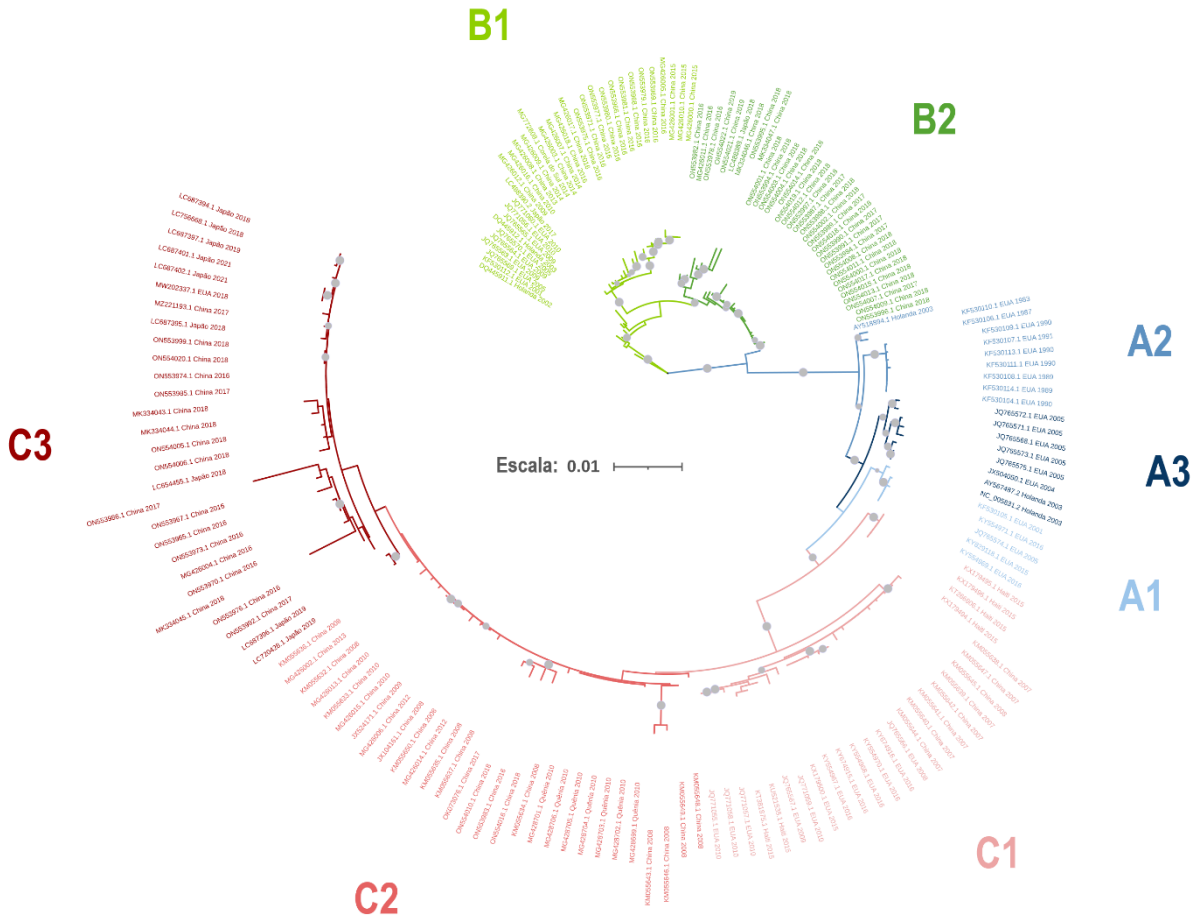
A figura apresenta os três melhores resultados de sequências de coronavírus para três das 16 amostras do projeto PRJNA671738 analisadas com a ferramenta Genome Detective. As três representações gráficas dos genomas do vírus HCoV NL63 foram retiradas dos relatórios finais obtidos através do pipeline de identificação de vírus da ferramenta Genome Detective. Fonte: Adaptado dos relatórios do Genome Detective.

2.2 Análises Filogenéticas

2.2.1 CLASSIFICAÇÃO FILOGENÉTICA E DISTRIBUIÇÃO GEOGRÁFICA DE 173 SEQUÊNCIAS DO GENE S DO HCOV-NL63

Considerando a importância do gene da proteína spike para a variabilidade genética do HCoV-NL63, 173 sequências do gene S presentes no banco de dados de nucleotídeos do NCBI (https://www.ncbi.nlm.nih.gov/nucleotide/) foram consideradas para as análises. O modelo escolhido para a árvore filogenética de máxima verossimilhança foi o TIM+F+I+G4 (**Figura 17**).

Figura 17: Árvore filogenética de máxima verossimilhança de 173 sequências do gene S do HCoV-NL63



A árvore filogenética representa a relação evolutiva entre 173 sequências Spike do HCoV NL63. Ela foi obtida através do método máxima verossimilhança utilizando-se o software IQ-TREE com a ferramenta ModelFinder. O modelo estatístico mais adequado após os testes foi o TIM+F+I+G4. As amostras apresentam código identificador, país de origem e ano de coleta. Os genótipos estão classificados por cor. Os valores de Ultrafast Booststrap ≥ 95 estão representados por círculos cinzas. Fonte: Elaborada pela autora.

Trabalhos anteriores mostraram que as cepas HCoV-NL63 são divididas em três genótipos (genótipos A, B e C) e oito subgenótipos (A1, A2, A3, B1, novo B, C1, C2 e C3) (CASTILLO et al., 2023; YE et al., 2023), dessa forma, as amostras foram distribuídas em A1 (2,9%), A2 (5,2%), A3 (4,6%), B1 (19,5%), novo B/B2 (20,1%), C1 (14,9%), C2 (16,7%) e C3 (16,1%). Os metadados das amostras consideradas também permitem classificá-las de acordo com país e data de coleta (**Tabela 7** e **Figura 18**). Enquanto os grupos B e C estão distribuídos pela Ásia, África e América do Norte, o grupo A, por outro lado, aparece exclusivamente nos EUA e na Holanda (**Figura 17** e **Tabela 7**). Além disso, as cepas mais recentes (datadas de 2018 a 2021) são principalmente dos grupos B2 e C3 (**Figura 17** e **Tabela 7**).

Tabela 7: Dados das 173 sequências do gene S do HCoV-NL63

ID da amostra	Ano de Coleta	País de Coleta	Genótipo
AY518894.1	2003	Holanda	A1
AY567487.2	2003	Holanda	A3
DQ445911.1	2002	Holanda	B1
DQ445912.1	2003	Holanda	B1
JQ765563.1	2009	EUA	B1
JQ765564.1	2009	EUA	B1
JQ765565.1	2009	EUA	B1
JQ765566.1*	2008	EUA	C1
JQ765567.1*	2009	EUA	C1
JQ765568.1	2005	EUA	A3
JQ765569.1	2005	EUA	B1
JQ765570.1	2005	EUA	B1
JQ765571.1	2005	EUA	A3
JQ765572.1	2005	EUA	A3
JQ765573.1	2005	EUA	A3
JQ765574.1	2005	EUA	A1
JQ765575.1	2005	EUA	A3
JQ771055.1*	2010	EUA	C1
JQ771056.1	2010	EUA	B1
JQ771057.1*	2010	EUA	C1
JQ771058.1*	2010	EUA	C1
JQ771059.1*	2010	EUA	C1
JQ771060.1	2010	EUA	B1
JX104161.1*	2008	China	C2
JX504050.1	2004	EUA	A3
JX524171.1*	2009	China	C2
KF530104.1	1990	EUA	A2
KF530105.1	2001	EUA	A1
KF530106.1	1987	EUA	A2
KF530107.1	1991	EUA	A2
KF530108.1	1989	EUA	A2
KF530109.1	1990	EUA	A2
KF530110.1	1983	EUA	A2
KF530111.1	1990	EUA	A2
KF530112.1	2001	EUA	B1
KF530113.1	1990	EUA	A2
KF530114.1	1989	EUA	A2
KM055632.1	2008	China	C2
KM055633.1*	2010	China	C2
KM055634.1*	2008	China	C2
KM055635.1*	2008	China	C2
KM055636.1*	2008	China	C2
KM055637.1*	2008	China	C2
KM055638.1*	2007	China	C1
KM055639.1*	2007	China	C1
KM055640.1*	2007	China	C1
KM055641.1	2007	China	C1
KM055642.1*	2007	China	C1
KM055643.1	2008	China	C2
KM055644.1*	2007	China	C1

KM055645.1*	2008	China	C1
KM055646.1	2008	China	C2
KM055647.1*	2007	China	C1
KM055648.1*	2008	China	C2
KM055649.1*	2008	China	C2
KM055650.1*	2008	China	C2
KT266906.1	2015	Haiti	C1
KT381875.1*	2015	Haiti	C1
KU521535.1*	2015	Haiti	C1
KX179494.1	2015	Haiti	C1
KX179495.1	2015	Haiti	C1
KX179496.1	2015	Haiti	C1
KX179500.1*	2015	EUA	C1
KY554967.1*	2016	EUA	C1
KY554968.1*	2016	EUA	C1
KY554969.1	2016	EUA	A1
KY554970.1*	2016	EUA	C1
KY554971.1	2016	EUA	A1
KY674915.1*	2016	EUA	C1
KY674916.1*	2016	EUA	C1
KY829118.1	2015	EUA	A1
LC488388.1	2018	Japão	B2
LC488390.2	2017	Japão	B1
LC654455.1	2018	Japão	C3
LC687394.1*	2018	Japão	C3
LC687395.1*	2018	Japão	C3
LC687396.1*	2019	Japão	C3
LC687397.1*	2019	Japão	C3
LC687401.1*	2021	Japão	C3
LC687402.1*	2021	Japão	C3
LC720428.1*	2019	Japão	C3
LC756668.1*	2018	Japão	C3
MG426000.1*	2015	China	B1
MG426001.1*	2015	China	B1
MG426002.1*	2013	China	C2
MG426003.1*	2014	China	B1
MG426004.1*	2016	China	C3
MG426005.1*	2016	China	B1
MG426006.1*	2012	China	C2
MG426007.1*	2014	China	B1
MG426008.1	2013	China	B1
MG426009.1*	2014	China	B1
MG426010.1*	2015	China	B1
MG426011.1	2016	China	B2
MG426012.1*	2009	China	B1
MG426013.1*	2010	China	C2
MG426014.1*	2012	China	C2
MG426015.1*	2010	China	C2
MG426016.1*	2010	China	B1
MG426017.1*	2016	China	B1
MG426018.1*	2016	China	B1
MG428699.1*	2010	Quênia	C2
MG428701.1*	2010	Quênia	C2
MG428702.1*	2010	Quênia	C2
MG428703.1*	2010	Quênia	C2

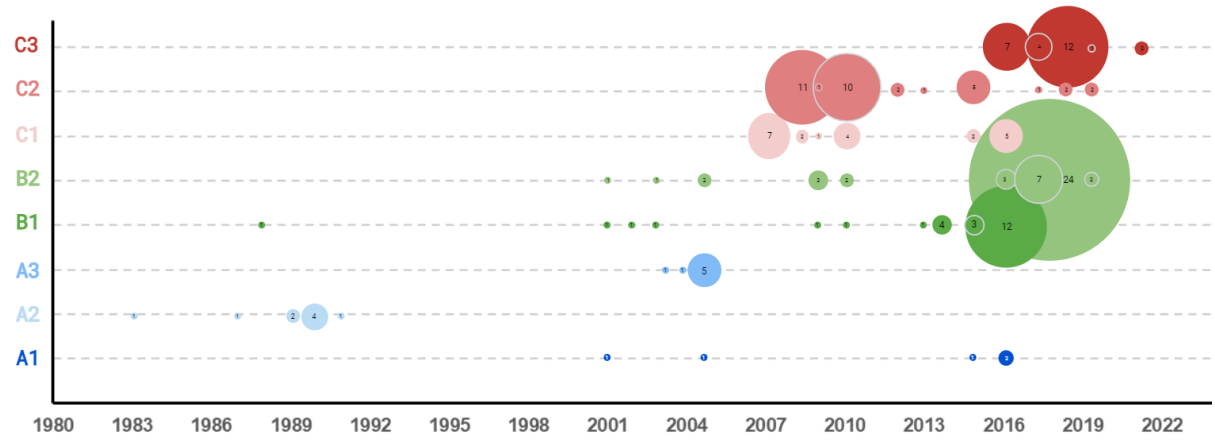
MG428704.1*	2010	Quênia	C2
MG428705.1*	2010	Quênia	C2
MG428706.1*	2010	Quênia	C2
MG772808.1*	2014	Coreia do Sul	B1
MK334043.1*	2018	China	C3
MK334044.1*	2018	China	C3
MK334045.1*	2018	China	C3
MK334046.1	2018	China	B2
MK334047.1	2018	China	B2
MW202337.1*	2018	EUA	C3
MZ221193.1	2017	China	C3
NC_005831.2	2003	Holanda	A3
OK073076.1*	2017	China	C2
ON553965.1*	2016	China	C3
ON553966.1*	2016	China	B1
ON553967.1*	2016	China	C3
ON553968.1*	2016	China	B1
ON553969.1*	2016	China	B1
ON553970.1*	2016	China	C3
ON553971.1*	2016	China	B1
ON553973.1*	2016	China	C3
ON553974.1*	2016	China	C3
ON553975.1*	2016	China	B1
ON553976.1*	2016	China	C3
ON553977.1*	2016	China	B1
ON553978.1	2016	China	B2
ON553979.1*	2016	China	B1
ON553980.1*	2016	China	B1
ON553981.1*	2016	China	B1
ON553982.1*	2016	China	B2
ON553983.1*	2016	China	C2
ON553984.1*	2017	China	B2
ON553985.1*	2017	China	C3
ON553986.1	2017	China	C3
ON553987.1*	2017	China	B2
ON553989.1*	2017	China	B2
ON553990.1*	2017	China	B2
ON553991.1*	2017	China	B2
ON553992.1*	2017	China	C3
ON553994.1*	2018	China	B2
ON553995.1	2018	China	B2
ON553996.1*	2018	China	B2
ON553997.1*	2018	China	B2
ON553998.1*	2018	China	B2
ON553999.1*	2018	China	C3
ON554000.1*	2018	China	B2
ON554001.1*	2018	China	B2
ON554002.1*	2018	China	B2
ON554003.1*	2018	China	B2
ON554004.1*	2018	China	B2
ON554005.1*	2018	China	C3
ON554006.1*	2018	China	C3
ON554007.1*	2017	China	B2
ON554008.1*	2018	China	B2
ON554009.1*	2018	China	B2

ON554010.1*	2018	China	C2
ON554011.1*	2018	China	B2
ON554012.1*	2018	China	B2
ON554013.1*	2018	China	B2
ON554014.1*	2018	China	B2
ON554015.1*	2018	China	B2
ON554016.1*	2018	China	C2
ON554017.1*	2018	China	B2
ON554018.1*	2018	China	B2
ON554019.1*	2018	China	B2
ON554020.1*	2018	China	C3
ON554021.1	2019	China	B2
ON554022.1*	2019	China	B2

*Amostras consideradas para a datação filogenética

Fonte: Elaborada pela autora.

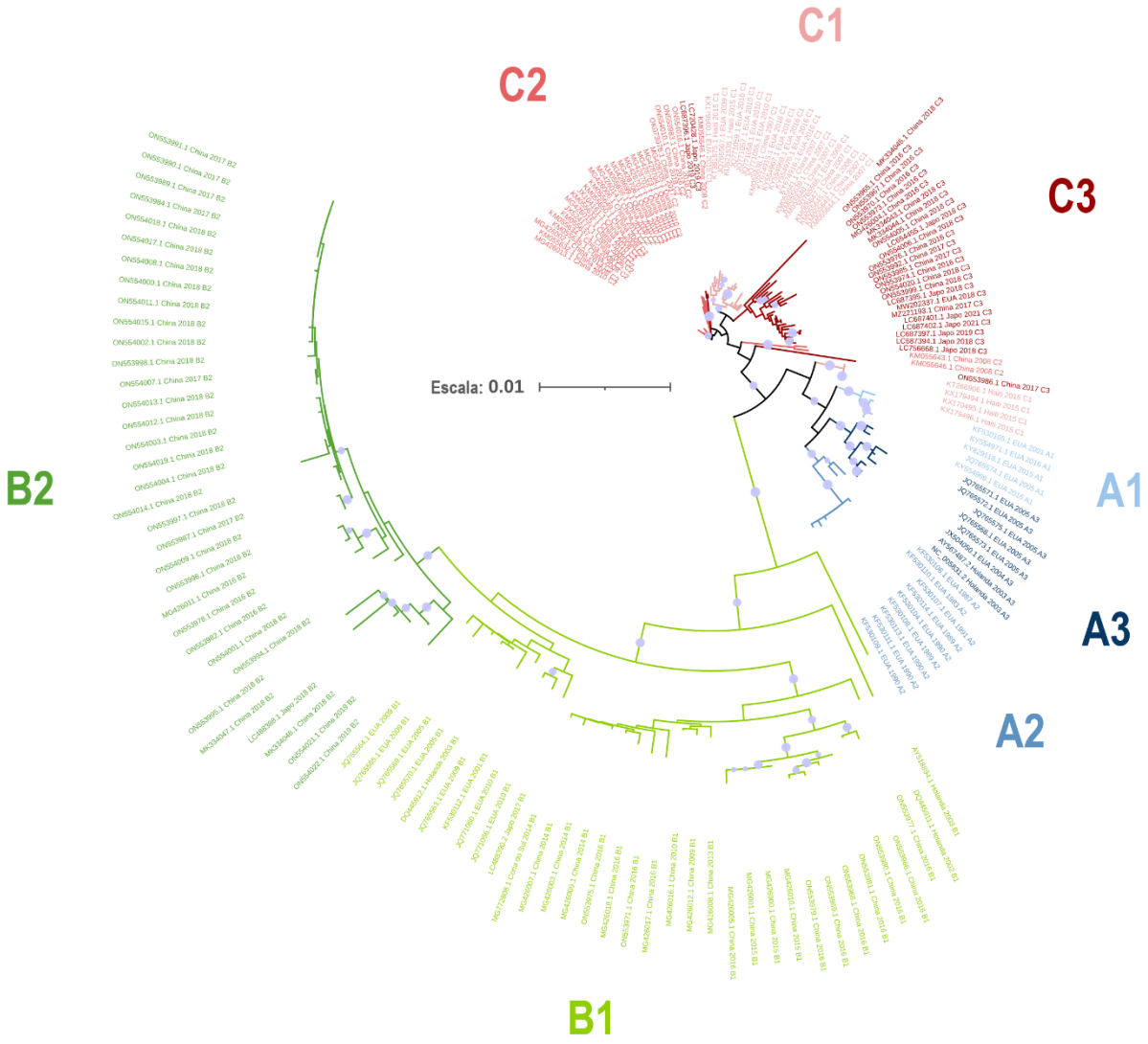
Figura 18: Distribuição temporal das 173 sequências do gene S do HCoV-NL63



O gráfico foi obtido com o auxílio do Microsoft Excel e representa a distribuição das 173 sequências do gene S do HCoV NL63 segundo genótipo (resultado da árvore filogenética na **Figura 17**) e data de coleta. Os genótipos estão classificados por cor. Fonte: Elaborada pela autora.

Com o objetivo de validar os resultados obtidos através do IQ-TREE para a árvore filogenética de máxima verossimilhança das 173 sequências do gene S do vírus HCoV-NL63, também foi obtida a árvore filogenética pelo método Neighbour Joining utilizando o modelo evolutivo Kimura. Os resultados apresentados foram semelhantes à distribuição dos genótipos obtidos pela árvore de máxima verossimilhança (**Figura 19**).

Figura 19: Árvore filogenética Neighbour Joining de 173 sequências do gene S do HCoV-NL63



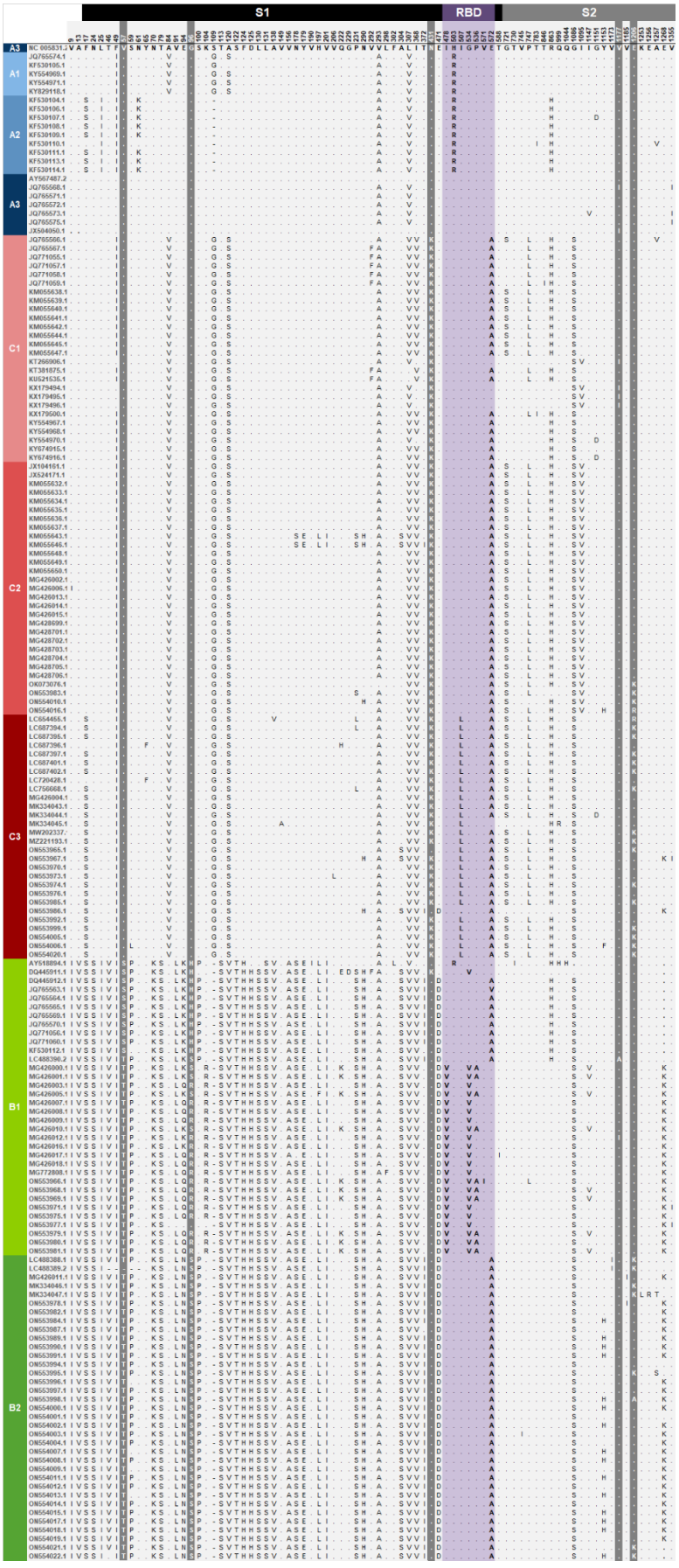
A árvore filogenética representa uma das versões de relação evolutiva entre 173 sequências Spike do HCoV NL63. Ela foi obtida através do método Neighbour Joining utilizando-se o software RapidNJ da plataforma Galaxy. O modelo estatístico utilizado foi o Kimura. As amostras apresentam código identificador, país de origem, ano de coleta e genótipo. Os genótipos estão classificados por cor, segundo a classificação da árvore de máxima verossimilhança. Os valores de booststrap ≥ 80 estão representados por círculos cinzas. Fonte: Elaborada pela autora.

2.2.2 ANÁLISES DE SUBSTITUIÇÃO DE AMINOÁCIDOS PARECEM SER MAIS FREQUENTES NO DOMÍNIO S1 E NO GENÓTIPO B

Para melhor caracterizar os genótipos e subgenótipos do HCoV-NL63 foram realizadas análises de substituição de aminoácidos da proteína spike. 223 locais de substituição foram detectados com uma frequência ≥ 10% nas 173 amostras (**Figura 20**). Entre essas mutações,

141 (63,2%) eram sinônimas, 80 (35,9%) eram não sinônimas e 2 (0,9%) eram do tipo indel. Foram encontradas 61 (76,2%) e 19 (23,8%) substituições não sinônimas nas regiões S1 e S2, respectivamente. O genótipo B apresentou o maior número de substituições não sinônimas (46), seguido pelos genótipos C (15) e A (11). Três resíduos na região S1 (V57, G96, N431) e dois na região S2 (V1177, E1206) foram identificados sob seleção positiva. O genótipo B possui substituições V57 e G96. Na posição V57, as substituições mudaram as propriedades de resíduos não polares para polares (V57S/T para B1 e V57S para B2). Na posição G96, não há alteração nas propriedades do resíduo (G96H/S/R em B1 e G96S em subgenótipo B2). O genótipo C mudou de não polar para polar na posição N431K. Outra substituição, embora menos frequente, no genótipo C foi a E1206K. O genótipo A não apresentou aminoácidos positivamente selecionados. Um total de 6 mutações foram observadas no domínio de ligação ao receptor DLR (I478V, H503R, I507L, G534V, P536A, E572A). H503R é encontrado apenas nos subgenótipos A1 e A2. O subgenótipo B1 apresentou quatro alterações: I478V, G534V, P536A e E572A. Entre essas substituições no DLR, apenas E572A é encontrada no subgenótipo B2 e em todo o genótipo C. Finalmente, C3 é caracterizado por uma substituição, I507L.

Figura 20: Análise de polimorfismo de aminoácido único da proteína spike do HCoV-NL63



173 proteínas Spike do HCoV-NL63 foram alinhadas com o webPRANK e utilizadas para a análise de polimorfismo de aminoácido único. As análises de mutação foram realizadas com o auxílio da ferramenta BioAider, e as análises de seleção positiva/negativa com as ferramentas DataMonkey, MEME, SLAC e FUBAR. A representação gráfica foi obtida com o auxílio do software MEGA 11 e do Microsoft Excel. Cinco aminoácidos selecionados positivamente estão destacados em cinza. As substituições de aminoácidos na região RDB estão destacadas em roxo. A classificação dos genótipos está codificada por cores à esquerda. Fonte: Elaborada pela autora.

2.2.3 ANÁLISES DE RECOMBINAÇÃO INDICAM UMA REGIÃO "*HOTSPOT*" DENTRO DO DOMÍNIO S1 DA SPIKE

46 potenciais pontos de cruzamento de recombinação foram detectados dentro dos genes da spike envolvendo diferentes relações parentais (**Tabela 8**). O maior número de eventos recombinantes foi 26, envolvendo a sequência MK334046.1 (B2) como parental principal, seguido por 10 eventos com sequências parentais ON553978.1 (maior/B2) e KM055641.1 (menor/C1) e 5 com a JQ765564. 1 (B1) como sequência parental principal (**Tabela 8**). As sequências envolvidas nos eventos recombinantes (n = 53) abrangeram todas as sequências do grupo A, além de alguns representantes de outros grupos (**Tabela 8**).

As análises também identificaram quatro regiões da spike envolvidas na recombinação (**Tabela 8**). O primeiro e mais frequente, com 28 eventos, ocorre da posição do nucleotídeo 36-37 a 920-923, correspondendo aos aminoácidos 12-13 a 306-307. Com apenas 2 eventos, o segundo está localizado entre as posições 518 a 1166-1180, correspondendo aos aminoácidos 172 a 388-393. A terceira região entre os nucleotídeos 742 a 2266 corresponde aos aminoácidos 247 a 755 e apareceram em apenas 1 evento. A quarta região de recombinação foi identificada em 15 eventos, dos nucleotídeos 2264 a 3822-3964 correspondendo aos aminoácidos 754 a 1274-1321.

Tabela 8: Resultados das análises de recombinação com o RDP5

N° de eventos	Início	Fim	Sequências Recombinantes	Parental Maior	Parental Menor	RDP	GENECONV	BootScan	MaxChi	Chimaera	SiScan	3Seq
1	36	923	DQ445911.1	JQ765574.1	Desconhecido	2.086×10^{-10}	1.510×10^{-04}	2.018×10^{-09}	2.301×10^{-10}	-	6.389×10^{-18}	1.018×10^{-07}
1	36	920	AY518894.1	JQ765574.1	Desconhecido	1.648×10^{-07}	6.615×10^{-06}	4.542×10^{-07}	5.087×10^{-10}	-	4.817×10^{-13}	5.875×10^{-08}
26	37	922	KY554969.1	MK334046.1	Desconhecido	3.197×10^{-04}	1.398×10^{-03}	3.301×10^{-02}	2.258×10^{-07}	2.936×10^{-06}	5.745×10^{-03}	1.977×10^{-04}
1	518	1166	KM055643.1	KM055632.1	MG426008.1	-	3.901×10^{-10}	5.840×10^{-12}	1.798×10^{-05}	1.783×10^{-05}	5.728×10^{-06}	2.305×10^{-17}
1	518	1180	KM055646.1	KM055632.1	MG426011.1	-	2.508×10^{-09}	4.886×10^{-11}	3.111×10^{-05}	6.066×10^{-05}	5.919×10^{-08}	9.415×10^{-11}
1	742	2266	ON553986.1	LC654455.1	ON553995.1	4.483×10^{-13}	4.805×10^{-11}	4.511×10^{-13}	6.175×10^{-08}	4.188×10^{-08}	1.584×10^{-07}	4.424×10^{-20}
5	2264	3964	ON553995.1	JQ765564.1	Desconhecido	2.547×10^{-08}	2.758×10^{-06}	2.752×10^{-08}	2.554×10^{-05}	3.552×10^{-05}	3.025×10^{-19}	2.064×10^{-10}
10	2264	3822	JQ765564.1	ON553978.1	KM055641.1	7.635×10^{-05}	1.375×10^{-03}	1.236×10^{-04}	4.334×10^{-03}	-	1.743×10^{-14}	2.728×10^{-05}

Os resultados da tabela foram retirados dos relatórios obtidos através do RDP5 com as análises de recombinação considerando todas as 173 sequências Spike do HCoV NL63. Os genótipos das amostras estão classificados por cores: azul claro (A1), verde claro (B1), verde escuro (B2), vermelho claro (C1), vermelho médio (C2) e vermelho escuro (C3). Fonte: Elaborada pela autora.

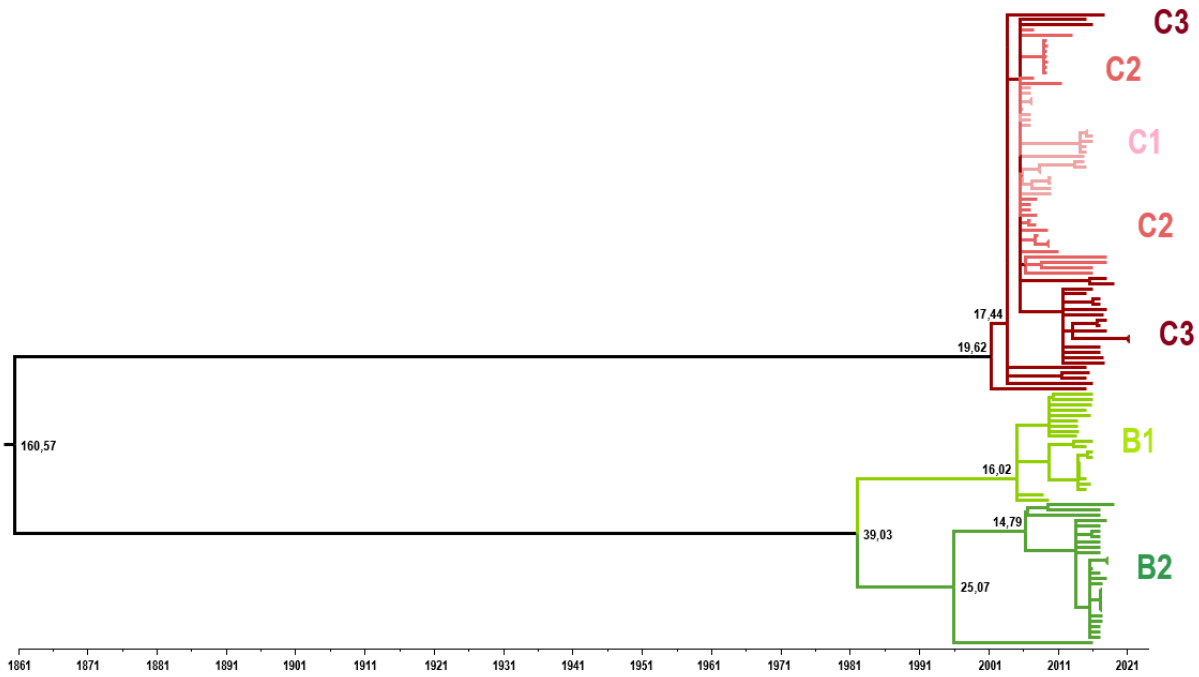
2.2.4 A DATAÇÃO MOLECULAR INDICA UMA COMPLEXA RELAÇÃO ENTRE OS GENÓTIPOS ANALISADOS

Para entender melhor a história evolutiva do HCoV-NL63, foram realizadas análises de datação molecular. Como a presença de eventos de recombinação pode afetar a datação molecular, todas as cepas recombinantes identificadas pelo RPD5 foram excluídas do nosso conjunto de dados (n = 53) e, como resultado, 120 sequências spike não recombinantes (**Tabela 7**) foram consideradas. Todas as sequências do grupo A foram excluídas da nossa análise devido ao provável envolvimento em múltiplos eventos de recombinação.

Os grupos B e C parecem surgiram há cerca de 121 e 140 anos, respectivamente, partindo do ano de 1861 (o tempo estimado para o ancestral comum mais recente - tMRCA - é 07/06/1861) (**Figura 21**). O Grupo B foi o primeiro a surgir. Os subgrupos B1 e B2 surgiram cerca de 23 e 13 anos depois. O subgrupo C3 emergiu 2 anos depois, dando origem ao C2 e ao C1, quase ao mesmo tempo, cerca de 2 anos depois.

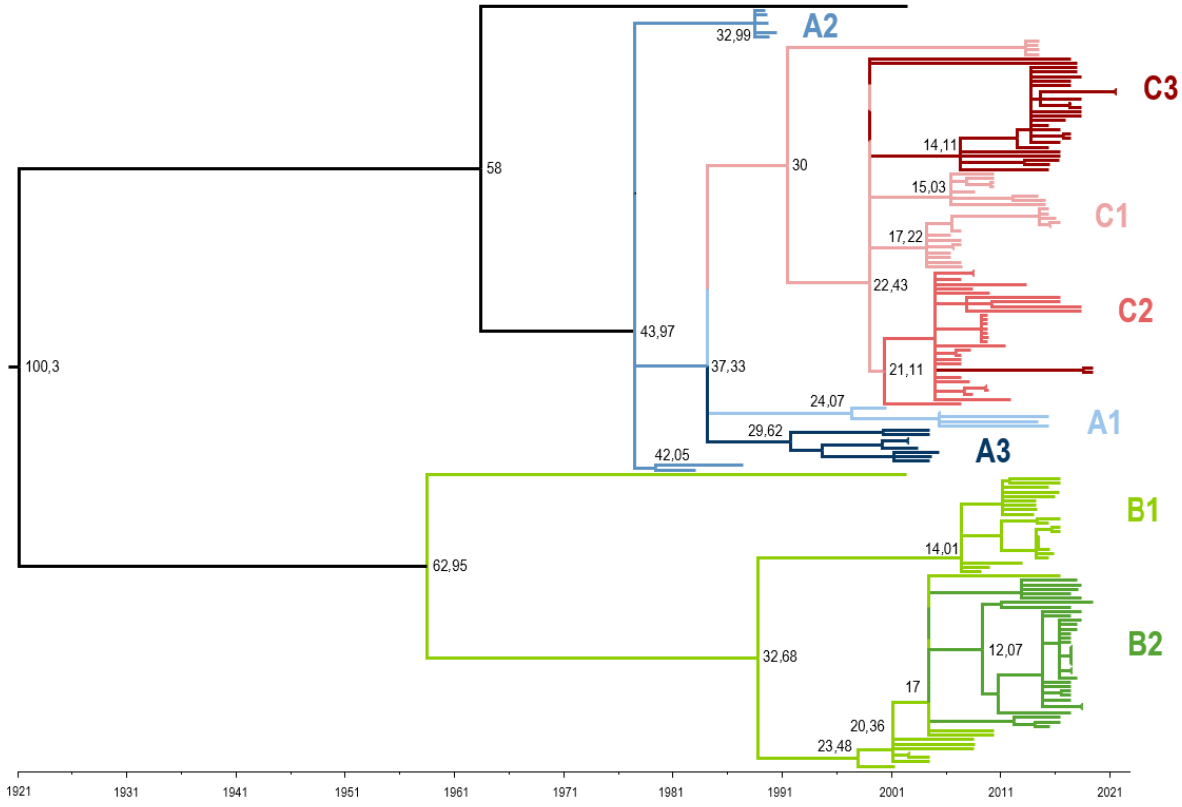
A critério de comparação também foi realizado o relógio molecular de todas as 173 amostras (**Figura 22**). O tMRCA foi 12/09/1921. A adição das amostras do grupo A indicam que este genótipo parece dar origem ao grupo C.

Figura 21: Datação filogenética de 120 sequências do gene S do HCoV-NL63



Relógio molecular considerando 120 sequências Spike do vírus HCoV NL63. A análise de datação filogenética foi realizada utilizando o software IQ-TREE com a ferramenta Least Square Dating (LSD) e as datas das coletas das amostras (metadados). Os genótipos estão classificados por cores. As idades dos nós são exibidas nos nós centrais. O modelo escolhido foi o TIM+F+R2 e o tempo estimado para o tMRCA foi 07/06/1861. Fonte: Elaborada pela autora.

Figura 22: Datação filogenética de 173 sequências do gene S do HCoV-NL63



Relógio molecular considerando todas as 173 sequências Spike do vírus HCoV NL63 consideradas no trabalho. A análise de datação filogenética foi realizada utilizando o software IQ-TREE com a ferramenta Least Square Dating (LSD) e as datas das coletas das amostras (metadados). Os genótipos estão classificados por cores. As idades dos nós são exibidas nos nós centrais. O modelo escolhido foi o TIM+F+I+G4 e o tMRCA foi 12/09/1921. Fonte: Elaborada pela autora.

3 DISCUSSÃO

3.1 Análises de Metagenomas e Transcriptomas

Os CoVs são vírus RNA zoonóticos que podem sofrer mutação e recombinação em taxas alarmantes, existem muitos fatores moleculares e virais intrínsecos que contribuem para isso (DHAMA et al., 2020; SU et al., 2016; WOO et al., 2009; YE et al., 2020) e principalmente sua capacidade de infectar uma ampla gama de hospedeiros é preocupante. É por isso que a proximidade entre humanos e possíveis hospedeiros de CoVs pode ser problemática. Com tudo isso, fica claro como os estudos epidemiológicos são essenciais para prevenir e responder adequadamente a futuros surgimentos de CoVs. Levando tudo isso em consideração, e na tentativa de valorizar parte dos dados de sequenciamento e metadados associados disponíveis no NCBI/SRA, foi investigada a potencial presença de diferentes CoVs em diversos dados de sequenciamento humano disponíveis em bancos de dados públicos. Em nosso estudo, usamos duas ferramentas (Kaiju e BWA) com estratégias diferentes (baseada em sequências de proteínas e de nucleotídeos, respectivamente) e uma outra ferramenta (Genome Detective) com ambas as estratégias associadas para tentar estabelecer a distribuição de CoVs em amostras de sequenciamento humano de DNA e de RNA.

Usando a ferramenta Kaiju baseada em proteínas (gene-based), de todas as amostras de humanos analisadas (n = 1169) contra um banco de dados de proteínas *Orthocoronavirinae*, obtivemos um número maior inesperado de sequências positivas para CoVs (n = 536) (na **Tabela 4** e na Erro! Fonte de referência não encontrada.). Para confirmar esses resultados, decidimos implementar uma segunda rodada de análise com um banco de dados de referência diferente (proteínas nr). Depois disso, apenas 150 sequências foram positivas (na **Tabela 4** e na Erro! Fonte de referência não encontrada.), das quais nenhuma veio a ser realmente de CoVs e t ambém não foi possível montar nenhum genoma viral. Supomos que a alta sensibilidade da ferramenta de classificação taxonômica e a estratégia baseada em aminoácidos de alguma forma resultaram em resultados falsos positivos, e as sequências detectadas positivamente eram na verdade sequências de proteínas humanas homólogas. Curiosamente, um trabalho anterior de Smits e colaboradores (SMITS et al., 2021) mostrou que o SARS-CoV-2 é aparentemente incapaz de integrar suas sequências no genoma humano e, portanto, concluímos que esses resultados não foram por causa disso. Nossos achados sugerem que as amostras investigadas

não apresentavam coronavírus, apenas sequências conservadas e semelhantes entre os organismos, o que foi superestimado pela ferramenta Kaiju, que pode não ser a melhor opção para busca de viromas de RNA em metagenomas como é para classificação taxonômica de positivos conhecidos sequências.

Com a ferramenta BWA baseada em nucleotídeos (reference based), de todas as amostras analisadas (n = 3642), apenas uma foi positiva: duas sequências curtas da amostra SRR606446 (BioProject PRJNA71831) (Erro! Fonte de referência não encontrada.). Mesmo n ão sendo suficiente para montar um genoma, foi interessante que essas sequências foram encontradas em amostras de escarro expectorado coletadas de indivíduos com fibrose cística, e foram identificadas através do BLAST (nt-nt) como “Human coronavirus NL63 strain Kilifi/IP/015/19-Nov-2012 spike protein S1 gene partial cds" (cobertura 100% e identidade 99,78%) e "Human coronavirus NL63 strain ChinaGD01 complete genome" (cobertura 99% e identidade 96,65%). Especialmente porque o HCoV-NL63 (*Alphacoronavírus*) é conhecido por causar doença respiratória moderada em crianças pequenas, idosos e pacientes imunocomprometidos (SU et al., 2016; YE et al., 2020), exceto por um relato recente de infecção grave do trato respiratório inferior na China (Y et al., 2020; YE et al., 2020). Não é possível determinar se essas sequências foram resultado da amostra humana real, indicando uma infecção do sujeito, ou se era um contaminante. No entanto, os resultados confirmam que a estratégia utilizada pode realmente identificar sequências virais se presentes. Talvez esse pipeline associado a outros metadados das amostras possa ter um significado mais relevante. Além disso, os resultados corroboram com o observado por Wahba et al. (2021) (WAHBA et al., 2020) com 99,8% de seus dados sem correspondência com as sequências de SARS-CoV-2. Uma outra possibilidade seria a incapacidade dos coronavírus de integrar suas sequências ao genoma humano, o que tornaria mais viável a pesquisa de sequências virais em amostras de transcriptoma, já que o genoma desses vírus é de RNA.

A abordagem do Genome Detective, de dupla estratégia (aminoácidos e nucleotídeos), foi realizada para quatro projetos de amostras de RNA: PRJNA794842, PRJNA774620, PRJNA671738 e PRJNA629087. Destes, apenas o projeto PRJNA671738 não foi concluído (35 amostras processadas de 58). Sendo assim, das 91 amostras analisadas, 16 resultados foram positivos para coronavírus, todos pertencentes ao projeto PRJNA671738 (**Tabela 6**). Embora essa ferramenta realize a busca ampla por vírus e apresente, na maioria dos casos, múltiplos resultados com diferentes níveis de cobertura, identidade e tamanhos de sequências, ela é uma das mais rápidas e intuitivas já que não necessita de execução manual para cada etapa ou da

dependência de um terminal de computador. Desses 16 resultados, a maioria apresentou baixo número de reads e percentual de cobertura, o que impossibilitou a formação de um contig mais longo do genoma. As três primeiras amostras da **Tabela 6** (SRR12893435, SRR12893436 e SRR12893437) foram as exceções (Erro! Fonte de referência não encontrada.), com a obtenção d os genomas quase completos dos vírus HCoV-OC43 (*Betacoronavírus*) e HCoV-NL63 (*Alphacoronavírus*). Mais uma vez, no entanto, não é possível determinar se essas sequências foram resultado de uma infecção real do sujeito ao qual pertence a amostra ou se seria o caso de contaminantes, mas os resultados confirmam que a estratégia utilizada funciona sensibilidade o bastante para identificar sequências virais, se presentes. Também é interessante ressaltar que juntamente com o HCoV-229E, o HCoV-OC43 causa cerca de 15 a 29% dos quadros gripais comuns e são muito bem caracterizados por esse motivo (MONTO, 1974; SU et al., 2016). Além disso, o HCoV-229E, o HCoV-OC43 e o HCoV-NL63 são distribuídos globalmente e apresentam padrão de transmissão aumentada durante o inverno em países de clima temperado, com o HCoV-NL63 também podendo apresentar maior transmissão no período de primavera-verão (CHIU et al., 2005, p. 63; HENDLEY; FISHBURNE; GWALTNEY, 1972; SU et al., 2016). Talvez essas características sejam indícios do porquê eles apareceram nas amostras investigadas, e em ambos os tipos de amostras (DNA e RNA), apesar de não ter sido possível realizar estudos de caracterização molecular e genômica.

Os resultados obtidos evidenciam algumas das dificuldades ao se trabalhar com a vigilância de patógenos virais em amostras de metagenomas. As limitações das ferramentas Kaiju e BWA ilustram uma problemática recorrente nesse tipo de análise, representada pelos falsos positivos identificados devido à alta identidade de nucleotídeos em segmentos curtos de sequências. Esse fator de dificuldade, embora comum, ainda não apresenta uma solução definitiva, mas pode ser superado com uma melhor compreensão das sequências homólogas identificadas. Uma opção seria a criação de bancos de dados dessas sequências para constituir uma das etapas de controle de qualidade antes das análises de identificação taxonômica.

3.2 Análises Filogenéticas

As cepas do vírus HCoV-NL63 já haviam sido identificadas e divididas em três genótipos (genótipos A, B e C) e oito subgenótipos (A1, A2, A3, B1, novo B, C1, C2 e C3) por outros trabalhos (CASTILLO et al., 2023; YE et al., 2023), a única diferença para o presente

trabalho foi que o genótipo emergente “novo B” descrito por Ye et al. (2023) foi considerado como B2 (YE et al., 2023). Os dados mostraram consistência na divisão filogenética de subgenótipos com estudos anteriores, e as amostras foram distribuídas em A1 (2,9%), A2 (5,2%), A3 (4,6%), B1 (19,5%), B2 (20,1%), C1 (14,9%), C2 (16,7%) e C3 (16,1%), como pode ser visto na **Figura 17**. Quanto à distribuição geográfica das amostras, enquanto os grupos B e C estão distribuídos pela Ásia, África e América do Norte, o grupo A, por outro lado, aparece exclusivamente nos EUA e na Holanda (**Figura 17** e **Tabela 7**). Essas informações podem ser coerentes com uma possível tendência de origem e/ou prevalência dessas cepas nas respectivas regiões. Outra característica notável é que as cepas mais recentes (datadas de 2018 a 2021) são principalmente dos grupos B2 e C3, o que pode indicar uma possível prevalência com tendência evolutiva (**Figura 17, Figura 18** e **Tabela 7**).

Quanto às análises de substituição de aminoácidos da proteína spike das amostras, de forma geral, 223 locais de substituição foram detectados com uma frequência ≥ 10% nas 173 amostras (**Figura 20**). Entre essas mutações, 141 (63,2%) eram sinônimas, 80 (35,9%) eram não sinônimas e 2 (0,9%) eram do tipo indel. O número de substituições não sinônimas na região S1 foi maior do que na região S2, como observado anteriormente em outros trabalhos (FORNI et al., 2022a). Foram encontradas 61 (76,2%) e 19 (23,8%) substituições não sinônimas nas regiões S1 e S2, respectivamente. Em especial, do ponto de vista evolutivo, essa característica pode ser algo extremamente vantajoso para a espécie HCoV-NL63, já que o DLR encontra-se na região S1. O genótipo B apresentou o maior número de substituições não sinônimas (46), seguido pelos genótipos C (15) e A (11). Essas informações podem também corroborar com a prevalência dos grupos B e C na atualidade.

Três resíduos na região S1 (V57, G96, N431) e dois na região S2 (V1177, E1206) foram identificados sob seleção positiva. O genótipo B possui as substituições V57 e G96. Na posição V57, as substituições mudaram as propriedades de resíduos não polares para polares (V57S/T para B1 e V57T para B2). Na posição G96, não há alteração nas propriedades do resíduo (G96H/S/R em B1 e G96S em B2). O genótipo C mudou de não polar para polar na posição N431K. Outra substituição menos frequente no genótipo C foi a E1206K. O genótipo A não apresentou aminoácidos positivamente selecionados. Um total de 6 mutações foram observadas no domínio de ligação ao receptor DLR (I478V, H503R, I507L, G534V, P536A, E572A). H503R é encontrado apenas nos subgenótipos A1 e A2. O subgenótipo B1 apresentou quatro alterações: I478V, G534V, P536A e E572A. Entre essas substituições no DLR, apenas E572A é encontrada no subgenótipo B2 e em todo o genótipo C. Finalmente, C3 é caracterizado por

uma substituição, I507L. Entre essas mutações, os resíduos G536 e P534 são aminoácidos bem conhecidos, críticos para a ligação entre a proteína Spike e a ACE2. A mutação I507L foi previamente identificada como facilitadora da entrada do vírus nas células hospedeiras (WANG et al., 2020). O significado biológico das outras substituições ainda é desconhecido, no entanto, todas as alterações identificadas podem, pela primeira vez, detalhar e descrever as principais características moleculares de cada subgenótipo. Além disso, considerando o papel da proteína Spike para a entrada do vírus na célula hospedeira e a pressão evolutiva que os coronavírus sofrem para adaptar-se a cada vez mais aos seus hospedeiros, seria coerente entender que as alterações pontuais positivas identificadas podem contribuir para um maior poder de infecção e/ou fuga do sistema imune. Além disso, essa caracterização também é um ponto de partida e grande aliada ao monitoramento de novas variantes.

46 potenciais pontos de recombinação cruzada foram detectados dentro dos genes da spike envolvendo diferentes relações parentais (**Tabela 8**). O maior número de eventos recombinantes foi 26, envolvendo a sequência MK334046.1 (B2) como parental principal, seguido por 10 eventos com as sequências parentais ON553978.1 (maior/B2) e KM055641.1 (menor/C1) e 5 com a JQ765564. 1 (B1) como sequência parental principal (**Tabela 8**). Curiosamente, as sequências envolvidas em eventos recombinantes cruzados (n = 53) abrangeram todas as sequências do grupo A, além de alguns representantes de outros grupos (**Tabela 8**).

As análises de recombinação também identificaram quatro principais regiões da spike envolvidas no processo (**Tabela 8**). A primeira da posição de nucleotídeos 36-37 a 920-923, correspondendo aos aminoácidos 12-13 a 306-307 (com 28 eventos); o segundo entre as posições 518 a 1166-1180 nts e 172 a 388-393 aa (2 eventos); a terceira região entre 742 a 2266 nts e 247 a 755 aa (1 evento); e a quarta região entre 2264 a 3822-3964 nts e 754 a 1274-1321 aa (15 eventos). A primeira e a terceira regiões de recombinação identificadas correspondem a duas regiões dentro da parte 5' do gene da spike do HCoV-NL63 previamente identificadas por Pyrc e colaboradores (PYRC et al., 2006). Este grupo identificou uma região do nucleotídeo 21072 ao 21161, correspondente aos aminoácidos 200 ao 230 (30 aa), e outra do nucleotídeo 21662 ao 21884, correspondendo aos aminoácidos 397 as 471 (74 aa), ambas no subdomínio S1. Como a terceira região possui apenas um evento e a primeira contém a maioria deles (28 eventos), é plausível concluir que esta última é uma região hipervariável de grande importância biológica, um possível "*hotspot*" dentro do subdomínio S1 da Spike desses vírus. Um importante ponto de interesse aos estudos evolutivos dos CoVs seria verificar se essas 4 regiões

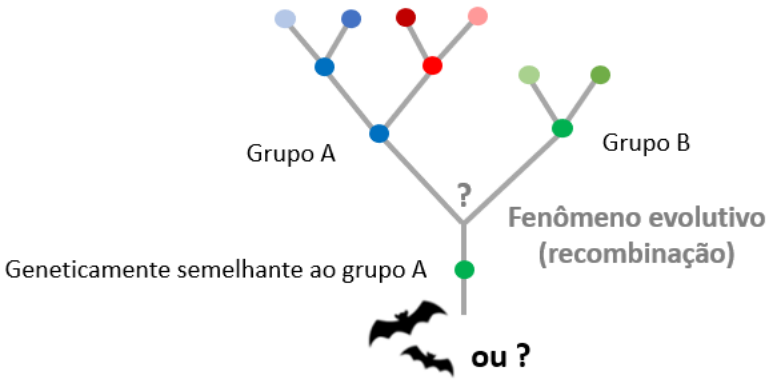
de recombinação (considerando as posições de quebra) seriam comuns a outros CoVs, especialmente o HCoV-229E, espécie mais próxima evolutivamente do HCoV-NL63.

Para entender melhor a história evolutiva do HCoV-NL63, foram realizadas análises de datação molecular. Como a presença de eventos de recombinação pode afetar a datação molecular, todas as cepas recombinantes identificadas pelo RPD5 foram excluídas do nosso conjunto de dados (n = 53) e, como resultado, 120 sequências spike não recombinantes (**Tabela 7**) foram consideradas. Todas as cepas do grupo A foram excluídas da nossa análise devido ao provável envolvimento em múltiplos eventos de recombinação. De acordo com os resultados, os grupos B e C parecem ter surgido há cerca de 121 e 140 anos, respectivamente, partindo de uma época inicial no ano de 1861 (o tempo estimado para o ancestral comum mais recente - tMRCA - é 07/06/1861) (**Figura 21**). O Grupo B foi o primeiro a surgir, com uma taxa de diversidade mais lenta que o C. Os subgrupos B1 e B2 surgiram cerca de 23 e 13 anos depois, respectivamente, e estão restritos à região da China (**Figura 21** e **Tabela 7**). O grupo C apresenta um processo evolutivo mais rápido, com o subgrupo C3 emergindo 2 anos depois, se espalhando principalmente pela região da Ásia (China e Japão), e dando origem ao C2 (China e Quênia) e ao C1 (China, EUA e Haiti) quase ao mesmo tempo, cerca de 2 anos depois (**Figura 21** e **Tabela 7**). Os resultados da datação molecular sem amostras do grupo A apresentaram baixa resolução, com muitos ramos parecidos, e o fenômeno de “long branch attraction”, com processo evolutivo temporal não confiável. No entanto, a caracterização da transmissão contínua e sazonal de infecções por HCoV-NL63 na China já foi descrita e acompanhada em outros trabalhos (CHIU et al., 2005; REN et al., 2011; SHAO et al., 2022; WANG et al., 2020; ZHANG et al., 2020). E além disso, estudos anteriores também exploraram os relógios moleculares para melhor compreender a evolução do HCoV-NL63 (AL-KHANNAQ et al., 2016; FORNI et al., 2022b; ROCHARS et al., 2017; WANG et al., 2020). No entanto, ao contrário de Forni e colaboradores (FORNI et al., 2022b), nenhum deles excluiu possíveis sequências recombinantes. Todos os relógios moleculares anteriores apresentaram semelhantes padrões de divergência entre os genótipos A, B e C, as diferenças mais significativas ocorrem principalmente para os tempos estimados, o que pode ser explicado pelos diferentes conjuntos de dados de trabalho considerados.

A critério de comparação, foi considerado também o relógio molecular com todas as 173 amostras de trabalho (**Figura 22**). Ambos os modelos de relógio molecular parecem manter a mesma estrutura de divergência entre os grupos e, sem dúvida, os grupos B e C são os mais dispersos na atualidade, especialmente B2, C2 e C3, concordando com uma possível tendência

de dispersão, conforme Ye e colaboradores (YE et al., 2023) propuseram. Por outro lado, a adição das amostras do grupo A leva ao questionamento de que este genótipo parece apenas dar origem às cepas do grupo C, perdendo a competição de coexistência imediatamente em seguida. Embora esse grupo apresente o menor número de amostras, ele foi o único grupo que não apresentou aminoácidos positivamente selecionados, teve todas as suas amostras envolvidas em eventos de recombinação, e ainda não apresentou amostras detectadas nos últimos anos. Essas características podem ser explicadas como efeitos de uma seleção purificadora ou negativa, um importante passo no processo evolutivo do HCoV-NL63. Levando-se em conta os resultados obtidos com as análises filogenéticas e de datação molecular, além das características apresentadas pelo grupo A, é possível supor que um ancestral comum entre os genótipos B e A tenha perdido ou ganhado a maior parte das substituições de aminoácidos que a porção S1 do gene S do grupo B apresenta. Essa hipótese evolutiva pode contribuir para melhor esclarecer a história evolutiva do HCoV-NL63 (**Figura 23**).

Figura 23: Hipótese evolutiva proposta para o genótipo A do HCoV-NL63



Embora as análises tenham sido realizadas com ferramentas e metodologias atuais e já utilizadas por outros grupos, ainda se deve considerar as diversas limitações do trabalho. Os resultados das análises de relógio molecular não podem ser tomados como absolutos, visto que apresentam diversos fatores de complicação. Mesmo tentando identificar as sequências recombinantes das análises, não podemos afirmar que as sequências restantes não passaram por processos de recombinação. Especialmente por tratar-se de sequências do gene S, região já sabidamente mais variável em relação ao genoma viral. Esse problema é compartilhado com

outros autores, o que evidencia uma deficiência dos métodos existentes na existência de um modelo matemático de datação molecular que evidencie o processo evolutivo do vírus HCoV-NL63 levando-se em conta a recombinação. Além disso, falsos eventos de recombinação podem ser produto de erros ou vieses na montagem dos genomas virais, o que configura mais uma dificuldade técnica a esse tipo de análise. Outro ponto importante é que as datas de coleta das amostras consideradas não necessariamente representam datas coerentes com o período de surgimentos dos genótipos ou variantes, podendo resultar em um viés na datação do processo evolutivo. No entanto, as referidas análises, mesmo com limitações e indefinições, contribuíram para a formulação da hipótese evolutiva relacionada ao genótipo A. Também é relevante ressaltar que os dados de distribuição geográfica das amostras estão limitados pelo grupo amostral considerado, o que pode não corresponder às tendências de dispersão reais do vírus no mundo. Tal limitação relaciona-se ao fato de que o vírus HCoV-NL63, por apresentar baixa patogenicidade, acaba sendo subestimado por não ter o seu diagnóstico clínico confirmado laboratorialmente com frequência. Esse problema, inclusive, dificulta o trabalho epidemiológico e de acompanhamento do patógeno por autoridades governamentais.

Diante do que foi exposto, o presente trabalho evidencia uma descrição genética detalhada e atualizada dos diferentes genótipos e subgenótipos da espécie HCoV-NL63, além de reunir análises com o maior número de sequências Spike para essa espécie. Como o representativo de amostras presentes nos bancos de dados públicos pode não corresponder exatamente aos padrões biológicos reais, embora sejam um referencial, melhores estratégias de monitoramento dessa espécie envolveriam iniciativas de contínuo sequenciamento de amostras clínicas, especialmente do gene S. O presente trabalho pode ser um primeiro passo na direção do estabelecimento de testes clínicos e de vigilância em saúde para o HCoV-NL63, já que a identificação de genótipos e subgenótipos através da sequência do gene S pode servir de referencial para a padronização de testes moleculares através de marcadores específicos.

4 CONCLUSÕES

As amostras de DNA/metagenomas aqui investigadas provavelmente não apresentavam nenhum coronavírus, apenas sequências de proteínas conservadas e semelhantes entre os organismos, o que foi superestimado pela ferramenta Kaiju. No entanto, é necessário entender que, embora o Kaiju não seja a melhor opção para pesquisar viromas de RNA em metagenomas humanos diversos, ainda é uma ótima ferramenta de classificação taxonômica para sequências positivas conhecidas. Além disso, os resultados confirmam que o BWA pode identificar viromas em vários produtos de sequenciamento. No caso dos transcriptomas/amostras de RNA, o Genome Detective se destaca como ferramenta de fácil uso e acesso, com alta sensibilidade, para detectar genomas virais.

Quanto às análises das sequências do coronavírus HCoV-NL63, os resultados apresentados destacam a complexidade da evolução dessa espécie. Sua classificação pode agora ser estabelecida em oito subgenótipos (A1, A2, A3, B1, B2, C1, C2 e C3). Comparado aos outros genótipos, o genótipo B tem uma quantidade notável de substituições de aminoácidos, incluindo aquelas positivamente selecionadas e na região do DLR. As mutações encontradas neste grupo são claramente vantajosas, pois B1 e B2 têm sido os subgenótipos mais prevalentes na China e no Japão desde 2016. Apesar de serem estreitamente relacionados, os genótipos A e C seguiram trajetórias distintas. O grupo C, particularmente o subgenótipo C3, é também o segundo genótipo mais comum no mundo. Por outro lado, o genótipo A, que tem o maior número de eventos de recombinação, mostra um período restrito de emergência tanto no tempo quanto no espaço, exibindo um padrão de seleção purificadora.

Como perspectivas futuras, os resultados apresentados podem auxiliar estratégias de vigilância epidemiológica e filogenética contínua de patógenos para prever/acompanhar a evolução de futuras variantes. Expandir a busca de CoVs para diferentes espécies virais e organismos hospedeiros seria interessante para ajudar a monitorar e conter o surgimento de novos vírus.

REFERÊNCIAS

A, K. et al. Detection of human coronavirus NL63, human metapneumovirus and respiratory syncytial virus in children with respiratory tract infections in south-west Sweden. **Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases**, v. 12, n. 11, nov. 2006.

AL-KHANNAQ, M. N. et al. Diversity and Evolutionary Histories of Human Coronaviruses NL63 and 229E Associated with Acute Upper Respiratory Tract Symptoms in Kuala Lumpur, Malaysia. **The American Journal of Tropical Medicine and Hygiene**, v. 94, n. 5, p. 1058, 5 maio 2016.

ARDEN, K. E. et al. New human coronavirus, HCoV-NL63, associated with severe lower respiratory tract disease in Australia. **Journal of Medical Virology**, v. 75, n. 3, p. 455–462, 1 mar. 2005.

BABRAHAM BIOINFORMATICS, B. B. **Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data**. Disponível em: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Acesso em: 4 ago. 2022.

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

BASTIEN, N. et al. Human Coronavirus NL63 Infection in Canada. **The Journal of Infectious Diseases**, v. 191, n. 4, p. 503–506, 15 fev. 2005.

BLACKSHIELDS, G. et al. Sequence embedding for fast construction of guide trees for multiple sequence alignment. **Algorithms for Molecular Biology : AMB**, v. 5, p. 21, 14 maio 2010.

C, M. et al. Phylogenetic analysis of human coronavirus NL63 circulating in Italy. **Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology**, v. 43, n. 1, set. 2008.

CARBO, E. C. et al. Performance of Five Metagenomic Classifiers for Virus Pathogen Detection Using Respiratory Samples from a Clinical Cohort. **Pathogens**, v. 11, n. 3, p. 340, 11 mar. 2022.

CASTILLO, G. et al. Molecular mechanisms of human coronavirus NL63 infection and replication. **Virus Research**, v. 327, p. 199078, 4 abr. 2023.

CC, H. et al. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. **Journal of virology**, v. 82, n. 4, fev. 2008.

CHAKRABORTY, I.; MAITY, P. COVID-19 outbreak: Migration, effects on society, global environment and prevention. **Science of The Total Environment**, v. 728, p. 138882, ago. 2020.

CHEN, S. et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. **Bioinformatics**, v. 34, n. 17, p. i884–i890, 1 set. 2018.

CHEN, Y.; LIU, Q.; GUO, D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. **Journal of Medical Virology**, v. 92, n. 4, p. 418–423, abr. 2020.

CHENG, V. C. C. et al. Severe Acute Respiratory Syndrome Coronavirus as an Agent of Emerging and Reemerging Infection. **Clinical Microbiology Reviews**, v. 20, n. 4, p. 660–694, out. 2007.

CHIU, S. S. et al. Human coronavirus NL63 infection and other coronavirus infections in children hospitalized with acute respiratory disease in Hong Kong, China. **Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America**, v. 40, n. 12, p. 1721–1729, 15 jun. 2005.

CORMAN, V. M. et al. Evidence for an Ancestral Association of Human Coronavirus 229E with Bats. **Journal of Virology**, v. 89, n. 23, p. 11858, 12 dez. 2015.

CROSSLEY, B. M. et al. Identification and Characterization of a Novel Alpaca Respiratory Coronavirus Most Closely Related to the Human Coronavirus 229E. **Viruses**, v. 4, n. 12, p. 3689, dez. 2012.

CUI, J.; LI, F.; SHI, Z.-L. Origin and evolution of pathogenic coronaviruses. **Nature Reviews Microbiology**, v. 17, n. 3, p. 181–192, mar. 2019.

DHAMA, K. et al. Coronavirus Disease 2019 –COVID-19. **Clinical Microbiology Reviews**, v. 33, n. 4, p. 48, 2020.

DOMINGUEZ, S. R. et al. Genomic analysis of 16 Colorado human NL63 coronaviruses identifies a new genotype, high sequence diversity in the N-terminal domain of the spike gene and evidence of recombination. **The Journal of General Virology**, v. 93, n. Pt 11, p. 2387, nov. 2012.

DRIOUICH, J.-S. et al. Reverse Genetics of RNA Viruses: ISA-Based Approach to Control Viral Population Diversity without Modifying Virus Phenotype. **Viruses**, v. 11, n. 7, p. 666, 20 jul. 2019.

EDGAR, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. **BMC Bioinformatics**, v. 5, p. 113, 2004.

FAYE, M. N. et al. Epidemiology of Non-SARS-CoV2 Human Coronaviruses (HCoVs) in People Presenting with Influenza-like Illness (ILI) or Severe Acute Respiratory Infections (SARI) in Senegal from 2012 to 2020. **Viruses**, v. 15, n. 1, p. 20, jan. 2023.

FigTree. Disponível em: <http://tree.bio.ed.ac.uk/software/figtree/>. Acesso em: 17 jul. 2023.

FORNI, D. et al. The substitution spectra of coronavirus genomes. **Briefings in Bioinformatics**, v. 23, n. 1, 17 jan. 2022a.

FORNI, D. et al. Dating the Emergence of Human Endemic Coronaviruses. **Viruses**, v. 14, n. 5, maio 2022b.

Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update | Nucleic Acids Research | Oxford Academic. Disponível em: <https://academic.oup.com/nar/article/50/W1/W345/6572001>. Acesso em: 3 dez. 2023.

GRIMALDI, A. et al. Improved SARS-CoV-2 sequencing surveillance allows the identification of new variants and signatures in infected patients. **Genome Medicine**, v. 14, n. 1, p. 90, 12 ago. 2022.

GUINDON, S. et al. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. **Nucleic Acids Research**, v. 33, n. Web Server issue, p. W557–W559, 1 jul. 2005.

HALL, B. G. **Phylogenetic Trees Made Easy: A How-To Manual**. Fifth Edition, New to this Edition:, Fifth Edition, New to this Edition: ed. Oxford, New York: Oxford University Press, 2017.

HARTENIAN, E. et al. The molecular virology of coronaviruses. **The Journal of Biological Chemistry**, v. 295, n. 37, p. 12910, 9 set. 2020.

HENDLEY, J. O.; FISHBURNE, H. B.; GWALTNEY, J. M. Coronavirus infections in working adults. Eight-year study with 229 E and OC 43. **The American Review of Respiratory Disease**, v. 105, n. 5, p. 805–811, maio 1972.

HOANG, D. T. et al. UFBoot2: Improving the Ultrafast Bootstrap Approximation. **Molecular Biology and Evolution**, v. 35, n. 2, p. 518–522, 1 fev. 2018.

HOFMANN, H. et al. Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 22, p. 7988, 5 maio 2005.

Home - Nucleotide - NCBI. Disponível em: <https://www.ncbi.nlm.nih.gov/nucleotide/>. Acesso em: 27 jun. 2023.

HUANG, X.; MADAN, A. CAP3: A DNA Sequence Assembly Program. **Genome Research**, v. 9, n. 9, p. 868–877, 1 set. 1999.

HUYNH, J. et al. Evidence Supporting a Zoonotic Origin of Human Coronavirus Strain NL63. **Journal of Virology**, v. 86, n. 23, p. 12816, dez. 2012.

JF, C. et al. Middle East respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease. **Clinical microbiology reviews**, v. 28, n. 2, abr. 2015.

JIANG, S. et al. A novel coronavirus (2019-nCoV) causing pneumonia-associated respiratory syndrome. **Cellular & Molecular Immunology**, v. 17, n. 5, p. 554–554, maio 2020.

KALYAANAMOORTHY, S. et al. ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. **Nature methods**, v. 14, n. 6, p. 587, jun. 2017.

KARTHIKEYAN, S. et al. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. **Nature**, v. 609, n. 7925, p. 101–108, 1 set. 2022.

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

KAWASAKI, J. et al. Hidden Viral Sequences in Public Sequencing Data and Warning for Future Emerging Diseases. **mBio**, v. 12, n. 4, p. e01638-21, 31 ago. 2021.

KESHEH, M. M. et al. An overview on the seven pathogenic human coronaviruses. **Reviews in Medical Virology**, v. 32, n. 2, p. e2282, mar. 2022.

KIYUKA, P. K. et al. Human Coronavirus NL63 Molecular Epidemiology and Evolutionary Patterns in Rural Coastal Kenya. **The Journal of Infectious Diseases**, v. 217, n. 11, p. 1728, 6 jun. 2018.

KOSAKOVSKY POND, S. L.; FROST, S. D. W. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. **Molecular Biology and Evolution**, v. 22, n. 5, p. 1208–1222, 1 maio 2005.

LAU, S. K. P. et al. Molecular Epidemiology of Human Coronavirus OC43 Reveals Evolution of Different Genotypes over Time and Recent Emergence of a Novel Genotype due to Natural Recombination. **Journal of Virology**, v. 85, n. 21, p. 11325, nov. 2011.

LEMEY, P.; SALEMI, M.; VANDAMME, A.-M. **The Phylogenetic Handbook | Genomics, bioinformatics and systems biology**. 2. ed. [s.l.] Anne-Mieke Vandamme, Guy Bottu, Marc Van Ranst, Philippe Lemey, Des Higgins, Korbinian Strimmer, Arndt von Haeseler, Marco Salemi, Yves Van de Peer, Heiko A. Schmidt, Fredrik Ronquist, Paul van der Mark, John P. Huelsenbeck, David L. Swofford, Jack Sullivan, Fred R. Opperdoes, David Posada, Oliver G. Pybus, Beth Shapiro, Sergei L. Kosakovsky Pond, Art F. Y. Poon, Simon D. W. Frost, Mika Salminen, Darren Martin, Allen Rodrigo, Alexei Drummond, Andrew Rambaut, Mary K. Kuhner, Xuhua Xia, Vincent Moulton, Katharina T. Huber, 2009.

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

LEVY, S. E.; BOONE, B. E. Next-Generation Sequencing Strategies. **Cold Spring Harbor Perspectives in Medicine**, v. 9, n. 7, p. a025791, jul. 2019.

LI, F. Structure, Function, and Evolution of Coronavirus Spike Proteins. **Annual review of virology**, v. 3, n. 1, p. 237, 9 set. 2016.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. **Bioinformatics**, v. 25, n. 14, p. 1754–1760, 15 jul. 2009.

LÖYTYNOJA, A.; GOLDMAN, N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. **BMC Bioinformatics**, v. 11, p. 579, 2010.

MARTIN, D. P. et al. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. **Virus Evolution**, v. 7, n. 1, jan. 2021.

MELNICK, M. et al. Application of a bioinformatic pipeline to RNA-seq data identifies novel virus-like sequence in human blood. **G3 Genes|Genomes|Genetics**, v. 11, n. 9, p. jkab141, 6 set. 2021.

MENZEL, P.; NG, K. L.; KROGH, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. **Nature Communications**, v. 7, n. 1, p. 11257, set. 2016.

MINH, B. Q. et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. **Molecular Biology and Evolution**, v. 37, n. 5, p. 1530–1534, 1 maio 2020.

MONTO, A. S. Medical reviews. Coronaviruses. **The Yale Journal of Biology and Medicine**, v. 47, n. 4, p. 234–251, dez. 1974.

MORAIS, D. A. A. et al. MEDUSA: A Pipeline for Sensitive Taxonomic Classification and Flexible Functional Annotation of Metagenomic Shotgun Sequences. **Frontiers in Genetics**, v. 0, 2022.

MULABBI, E. N.; TWEYONGYERE, R.; BYARUGABA, D. K. The history of the emergence and transmission of human coronaviruses. **The Onderstepoort Journal of Veterinary Research**, v. 88, n. 1, 2021.

MURRELL, B. et al. Detecting Individual Sites Subject to Episodic Diversifying Selection. **PLOS Genetics**, v. 8, n. 7, p. e1002764, 12 jul. 2012.

MURRELL, B. et al. FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. **Molecular Biology and Evolution**, v. 30, n. 5, p. 1196–1205, 1 maio 2013.

NG, O.-W.; TAN, Y.-J. Understanding bat SARS-like coronaviruses for the preparation of future coronavirus outbreaks — Implications for coronavirus vaccine development. **Human Vaccines & Immunotherapeutics**, v. 13, n. 1, p. 186–189, 2 jan. 2017.

PAYNE, S. Family Coronaviridae. Em: **Viruses**. [s.l.] Elsevier, 2017. p. 149–158.

PYRC, K. et al. Mosaic Structure of Human Coronavirus NL63, One Thousand Years of Evolution. **Journal of Molecular Biology**, v. 364, n. 5, p. 964, 12 dez. 2006.

RA, F. et al. A previously undescribed coronavirus associated with respiratory disease in humans. **Proceedings of the National Academy of Sciences of the United States of America**, v. 101, n. 16, 20 abr. 2004.

REN, L. et al. Prevalence of human coronaviruses in adults with acute respiratory tract infections in Beijing, China. **Journal of Medical Virology**, v. 83, n. 2, p. 291–297, 1 fev. 2011.

ROCHARS, V. M. B. D. et al. Isolation of Coronavirus NL63 from Blood from Children in Rural Haiti: Phylogenetic Similarities with Recent Isolates from Malaysia. **The American Journal of Tropical Medicine and Hygiene**, v. 96, n. 1, p. 144, 1 jan. 2017.

RONQUIST, F. et al. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. **Systematic Biology**, v. 61, n. 3, p. 539–542, maio 2012.

S, VAN B. et al. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. **mBio**, v. 3, n. 6, 20 nov. 2012.

SH, H. et al. Epidemiology of human coronavirus NL63 infection among hospitalized patients with pneumonia in Taiwan. **Journal of microbiology, immunology, and infection = Wei mian yu gan ran za zhi**, v. 50, n. 6, dez. 2017.

SHAO, N. et al. Molecular evolution of human coronavirus-NL63, -229E, -HKU1 and -OC43 in hospitalized children in China. **Frontiers in Microbiology**, v. 13, 2022.

SIEVERS, F.; HIGGINS, D. G. Clustal Omega. **Current Protocols in Bioinformatics**, v. 48, n. 1, p. 3.13.1-3.13.16, 2014.

SIMONSEN, M.; MAILUND, T.; PEDERSEN, C. N. S. **Rapid Neighbour-Joining**. Algorithms in Bioinformatics. **Anais**... Em: INTERNATIONAL WORKSHOP ON ALGORITHMS IN BIOINFORMATICS. Springer, Berlin, Heidelberg, 2008. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-540-87361-7_10>. Acesso em: 2 dez. 2023

SK, L. et al. Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related Rhinolophus bat coronavirus in China reveal bats as a reservoir for acute, self-limiting infection that allows recombination events. **Journal of virology**, v. 84, n. 6, mar. 2010.

SMITS, N. et al. No evidence of human genome integration of SARS-CoV-2 found by long-read DNA sequencing. **Cell Reports**, v. 36, n. 7, 17 ago. 2021.

Software de planilha Microsoft Excel | Microsoft 365. Disponível em: <https://www.microsoft.com/pt-br/microsoft-365/excel?ef_id=_k_60538491c7e91ccce30fe11f0afead0b_k_&OCID=AIDcmmq9ldqz5w_SEM__k_60538491c7e91ccce30fe11f0afead0b_k_&msclkid=60538491c7e91ccce30fe11f0afead0b>. Acesso em: 15 dez. 2023.

SONG, S. et al. Rapid screening and identification of viral pathogens in metagenomic data. **BMC Medical Genomics**, v. 14, n. S6, p. 289, dez. 2021.

SOONNARONG, R. et al. Molecular epidemiology and characterization of human coronavirus in Thailand, 2012–2013. **SpringerPlus**, v. 5, n. 1, p. 1–8, dez. 2016.

SU, S. et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. **Trends in Microbiology**, v. 24, n. 6, p. 490–502, jun. 2016.

SUCHARD, M. A. et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. **Virus Evolution**, v. 4, n. 1, p. vey016, 8 jun. 2018.

TAMURA, K.; STECHER, G.; KUMAR, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. **Molecular Biology and Evolution**, v. 38, n. 7, p. 3022, jul. 2021.

TANG, G.; LIU, Z.; CHEN, D. Human coronaviruses: Origin, host and receptor. **Journal of Clinical Virology**, v. 155, p. 105246, out. 2022.

TANG, X. et al. Adaptive Evolution of the Spike Protein in Coronaviruses. **Molecular Biology and Evolution**, v. 40, n. 4, 4 abr. 2023.

TAO, Y. et al. Surveillance of Bat Coronaviruses in Kenya Identifies Relatives of Human Coronaviruses NL63 and 229E and Their Recombination History. **Journal of Virology**, v. 91, n. 5, p. e01953-16, 14 fev. 2017.

TEMMAM, S. et al. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. **Nature**, v. 604, n. 7905, p. 330–336, abr. 2022.

TERRÓN-CAMERO, L. C. et al. Comparison of Metagenomics and Metatranscriptomics Tools: A Guide to Making the Right Choice. **Genes**, v. 13, n. 12, dez. 2022.

TK, C.; N, B. Human coronavirus NL-63 infection in a Brazilian patient suspected of H1N1 2009 influenza infection: description of a fatal case. **Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology**, v. 53, n. 1, jan. 2012.

TO, T.-H. et al. Fast Dating Using Least-Squares Criteria and Algorithms. **Systematic Biology**, v. 65, n. 1, p. 82–97, 1 jan. 2016.

TORII, S. et al. Establishment of a reverse genetics system for SARS-CoV-2 using circular polymerase extension reaction. **Cell Reports**, v. 35, n. 3, p. 109014, abr. 2021.

TORTORICI, M. A. et al. Structure, receptor recognition, and antigenicity of the human coronavirus CCoV-HuPn-2018 spike glycoprotein. **Cell**, v. 185, n. 13, p. 2279- 2291.e17, 23 jun. 2022.

VABRET, A. et al. Human Coronavirus NL63, France. **Emerging Infectious Diseases**, v. 11, n. 8, p. 1225, ago. 2005.

VABRET, A. et al. Detection of the New Human Coronavirus HKU1: A Report of 6 Cases. **Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America**, v. 42, n. 5, p. 634, 3 mar. 2006.

VAN DER HOEK, L. et al. Identification of a new human coronavirus. **Nature Medicine**, v. 10, n. 4, p. 368–373, abr. 2004.

VIJGEN, L. et al. Complete Genomic Sequence of Human Coronavirus OC43: Molecular Clock Analysis Suggests a Relatively Recent Zoonotic Coronavirus Transmission Event. **Journal of Virology**, v. 79, n. 3, p. 1595, fev. 2005.

VILSKER, M. et al. Genome Detective: an automated system for virus identification from high-throughput sequencing data. **Bioinformatics**, v. 35, n. 5, p. 871–873, 1 mar. 2019.

WAHBA, L. et al. An Extensive Meta-Metagenomic Search Identifies SARS-CoV-2-Homologous Sequences in Pangolin Lung Viromes. **mSphere**, v. 5, n. 3, p. e00160-20, 24 jun. 2020.

WANG, Y. et al. Discovery of a subgenotype of human coronavirus NL63 associated with severe lower respiratory tract infection in China, 2018. **Emerging Microbes & Infections**, v. 9, n. 1, p. 246, 2020.

WEAVER, S. et al. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. **Molecular Biology and Evolution**, v. 35, n. 3, p. 773, mar. 2018.

Which multiple alignment algorithm should I use? Disponível em: <https://help.geneious.com/hc/en-us/articles/360044627712-Which-multiple-alignment-algorithm-should-I-use->. Acesso em: 15 dez. 2023.

WILSON, M. R. et al. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. **New England Journal of Medicine**, v. 370, n. 25, p. 2408–2417, 19 jun. 2014.

WOO, P. C. Y. et al. Coronavirus Diversity, Phylogeny and Interspecies Jumping. **Experimental Biology and Medicine**, v. 234, n. 10, p. 1117–1127, out. 2009.

WU, F. et al. A new coronavirus associated with human respiratory disease in China. **Nature**, v. 579, n. 7798, p. 265–269, 12 mar. 2020.

Y, W. et al. Discovery of a subgenotype of human coronavirus NL63 associated with severe lower respiratory tract infection in China, 2018. **Emerging microbes & infections**, v. 9, n. 1, 29 jan. 2020.

YANG, P.; WANG, X. COVID-19: a new challenge for human beings. **Cellular & Molecular Immunology**, v. 17, n. 5, p. 555–557, maio 2020.

YE, R.-Z. et al. Continuous evolution and emerging lineage of seasonal human coronaviruses: A multicenter surveillance study. **Journal of Medical Virology**, v. 95, n. 6, p. e28861, 1 jun. 2023.

YE, S. H. et al. Benchmarking Metagenomics Tools for Taxonomic Classification. **Cell**, v. 178, n. 4, p. 779–794, ago. 2019.

YE, Z.-W. et al. Zoonotic origins of human coronaviruses. **International Journal of Biological Sciences**, v. 16, n. 10, p. 1686–1697, 2020.

ZHANG, L. et al. Complete Genome Sequences of Five Human Coronavirus NL63 Strains Causing Respiratory Illness in Hospitalized Children in China. **Microbiology Resource Announcements**, 20 fev. 2020.

ZHANG, Y. et al. Etiology and clinical characteristics of SARS-CoV-2 and other human coronaviruses among children in Zhejiang Province, China 2017–2019. **Virology Journal**, v. 18, n. 1, p. 1–11, dez. 2021.

ZHOU, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. **Nature**, v. 579, n. 7798, p. 270–273, 2020a.

ZHOU, Z.-J. et al. BioAider: An efficient tool for viral genome analysis and its application in tracing SARS-CoV-2 transmission. **Sustainable Cities and Society**, v. 63, p. 102466, dez. 2020b.