UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE BIOINFORMÁTICA

Élisson Nogueira Lopes

AVALIAÇÃO DO USO DE CÓDONS EM INFECÇÕES VIRAIS USANDO COMO MODELO O VÍRUS DA SÍNDROME RESPIRATÓRIA AGUDA GRAVE (SARS-COV-2)

Belo Horizonte

2022

ÉLISSON NOGUEIRA LOPES

AVALIAÇÃO DO USO DE CÓDONS EM INFECÇÕES VIRAIS USANDO COMO MODELO O VÍRUS DA SÍNDROME RESPIRATÓRIA AGUDA GRAVE (SARS-CoV-2)

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

Orientadora: Dr.ª Marta Giovanetti
Coorientador: Dr. Luiz Carlos Júnior Alcântara

Belo Horizonte
2022

043 Lopes, Élisson Nogueira.
Avaliação do uso de códons em infecções virais usando como modelo o vírus da Síndrome Respiratória Aguda Grave 2 (SARS-CoV-2) [manuscrito] / Élisson Nogueira Lopes. – 2022.
82 f. : il. ; 29,5 cm.

Orientadora: Dra. Marta Giovanetti. Coorientador: Dr. Luiz Carlos Júnior Alcântara.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Síndrome respiratória aguda grave. 3. Vírus. 4. Códon. I. Giovanetti, Marta. II. Alcântara, Luiz Carlos Júnior. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

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

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

ATA DA DEFESA DE TESE

ÉLISSON NOGUEIRA LOPES

Às nove horas do dia **29 de junho de 2022**, reuniu-se, através de videoconferência, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Avaliação do uso de códons em infecções virais usando como modelo o vírus da Síndrome Respiratória Aguda Grave 2 (SARS-CoV-2)"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, a Presidente da Comissão, **Dra. Marta Giovanetti**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Professor(a)/Pesquisador(a)	Instituição	Indicação
Dra. Marta Giovanetti - Orientadora	Fundação Oswaldo Cruz	**Aprovado**
Dr. Luiz Carlos Júnior Alcântara - Coorientador	Fundação Oswaldo Cruz	**Aprovado**
Dra. Glória Regina Franco	Universidade Federal de Minas Gerais	**Aprovado**
Dr. Gabriel da Rocha Fernandes	Fundação Oswaldo Cruz	**Aprovado**
Dra. Fernanda Khouri Barreto	Universidade Federal da Bahia	**Aprovado**
Dr. Rodrigo Juliani Siqueira Dalmolin	Universidade Federal do Rio Grande do Norte	**Aprovado**

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

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

Belo Horizonte, 29 de junho de 2022.

Documento assinado eletronicamente por **Fernanda Khouri Barreto, Usuária Externa**, em 30/06/2022, às 10:22, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rodrigo Juliani Siqueira Dalmolin, Usuário Externo**, em 30/06/2022, às 11:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gabriel da Rocha Fernandes, Usuário Externo**, em

30/06/2022, às 15:02, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gloria Regina Franco, Professora do Magistério Superior**, em 30/06/2022, às 19:09, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marta Giovanetti, Usuária Externa**, em 01/07/2022, às 12:00, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Luiz Carlos Junior Alcantara, Usuário Externo**, em 08/07/2022, às 16:55, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1485897** e o código CRC **46A5635B**.

Referência: Processo nº 23072.230757/2022-00 SEI nº 1485897

AGRADECIMENTOS

A minha mãe, meu pai, minha irmã e esposa, este trabalho dedico a vocês. Vocês são a base de quem sou, que lutaram comigo e por mim, que estão ao meu lado nas tempestades, e, hoje, desfrutaremos juntos deste dia alegre de sol.

Agradeço em especial, a minha mentora, a Professora Marta, pela paciência e sabedoria nos ensinamentos, que, com certeza, foram a base de tudo que aprendi e que levarei sempre na minha carreira como pesquisador. Agradeço muito pela honra de ter trabalhado com você. Aos membros da banca, pelas sugestões que tornaram este rascunho, uma tese.

Agradeço, de modo particular, a todos da família Bioinformática, Vagner por me ajudar sempre que tive um problema com programação; Stephane, Hegger e Xavier por me ajudarem a escrever melhor; Rodrigo Profeta e Paulo Soares pela amizade e companheirismo do dia a dia; membros da secretaria que, sempre que precisei, estenderam-me a mão. E a todos os demais não citados aqui, mas que contribuíram no meu cotidiano, fica aqui o meu mais sincero obrigado!

A Deus, meu guia, meu maior agradecimento, quando tudo parecia impossível, eu nunca estive sozinho.

"*É necessário sempre acreditar que o sonho é possível*
Que o céu é o limite e você, truta, é imbatível
Que o tempo ruim vai passar, é só uma fase
Que o sofrimento alimenta mais a sua coragem…"
Racionais MC's - A vida é desafio

RESUMO

O surto de (*Severe Acute Respiratory Syndrome Coronavirus 2*) SARS-CoV-2 que começou no ano de 2019, em Wuhan, China, rapidamente tornou-se uma preocupação global. Após alguns meses, o vírus foi identificado em todos os continentes do mundo e, em 11 de março de 2020, a Organização Mundial da Saúde (OMS) declarou a doença causada pelo SARS-CoV-2 (*Corona Virus Disease*: Covid-19) como pandemia. A partir disso, a rápida disseminação viral e o alarmante número de indivíduos infectados contribuíram para o surgimento de diferentes variantes ao longo do tempo. No presente trabalho, propôs-se uma metodologia de bioinformática associada à vigilância genômica como um meio de ajudar a elucidar o processo infeccioso causado por esse patógeno emergente. Para isso, foram analisadas todas as sequências referências disponíveis de Betacoronavírus de forma a correlacioná-las com a disponibilidade de tRNAs dos hospedeiros. Essas análises permitiram identificar um mecanismo de competição entre os Betacoronavírus e os hospedeiros, através da seleção de códons, além de identificar oito hospedeiros suscetíveis como potenciais reservatórios para o SARS-CoV-2. Adicionalmente, foi estimado o impacto de Variantes de Preocupação Internacionais (VOCs) introduzidas no Brasil, avaliando como a substituição de diferentes nucleotídeos, ao longo do tempo, possibilitou o aumento dos números de casos e de mortes no país. Os resultados reforçam a necessidade da implementação da vigilância genômica em larga escala, para compreender as medidas de transmissão de variantes e para auxiliar a implementação de rápidas medidas de contenção. A metodologia aplicada neste estudo pode ser replicada para a caracterização de outros possíveis patógenos emergentes e reemergentes em âmbitos nacional e mundial.

Palavras-chave: Vírus, epidemia, códons, bioinformática, vigilância.

ABSTRACT

The SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) outbreak that started in the year 2019 in Wuhan, China has quickly become a global concern. After a few months, the virus was identified all over the world. March 11, 2020, the World Health Organization (WHO) declared the disease caused by SARS-CoV-2 (Corona Virus Disease: Covid-19) a pandemic. The fast viral spread causes the emergence of different variants over time. In the present work, we use a bioinformatics methodology associated with genomic surveillance as a way to elucidate the infectious process caused by this emerging pathogen. For this purpose, all available reference sequences of Betacoronavirus were analyzed and correlated with the availability of tRNAs in the hosts. Our results allowed us to identify a competition mechanism between Betacoronaviruses and their hosts, through codon selection, in addition suggesting eight susceptible hosts as potential reservoirs for SARS-CoV-2. Additionally, we estimated the impact of International Variants of Concern (VOCs) introduced in Brazil, evaluating how the substitution of circulating variants, over time, allowed the increase in the number of cases and deaths. Our results reinforce the need for large-scale implementation of genomic surveillance, to understand variant transmission and to assist in the implementation of rapid containment measures. This methodology applied in this study can be replicated for the characterization of other possible emerging and reemerging pathogens at national and global levels.

Keywords: Virus, epidemic, codons, bioinformatic, surveillance.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ACE2 – Enzima conversora de angiotensina 2

Aids - Síndrome da imunodeficiência humana

CDS – Sequência codificadora

CHIKV – Chikungunya Vírus

Covid-19 – Corona Virus Disease (Doença do Coronavírus), "19" se refere ao ano 2019

CoVs – Família dos Coronavírus

CUB – *Codon Usage Bias*

DENV – Dengue Vírus

EBOV – Vírus Ebola

ED – Distância Euclidiana

eEFs – Fatores de alongamento eucariótico

eIFs – Fatores de tradução eucariótico

eRFs – Fatores de liberação eucariótico

FMV – *Formerly monitored variant*

GISAID – Global Initiative on Sharing All Influenza Data

HCOV-229e – Coronavírus que pode infectar humano 229e

HCOV-HKU1 – Coronavírus que pode infectar humano HKU1

HCOV-MERS – Coronavírus que pode infectar humano MERS

HCOV-NL63 – Coronavírus que pode infectar humano NL63

HCOV-OC43 – Coronavírus que pode infectar humano OC43

HCOV-SARS – Coronavírus que pode infectar humano SARS

HIV – Vírus da Imunodeficiência Humana

H1N1 - Influenza virus

H5N1 - Influenza Aviaria

ICTV – International Committee on Taxonomy of Viruses

Kb - Kilobase

kDa - KiloDalton

MAFFT - *Multiple Sequence Alignment Program*

MERS - Síndrome Respiratória do Oriente Médio

mRNA - RNA mensageiro

NCBI – National Center for Biotechnology Information

NSPs – Proteínas não Estruturais

OMS – Organização Mundial da Saúde

ORF – *Open Read Frame*

PAHO – Organização Pan-americana da Saúde

RNA – Ácido ribonucleico

RSCU – *Relative Synonymous Codon Usage*

RTC – Complexo de Replicação e Tradução

RTS – Sítio de Tradução do Ribossomo

SARS – Síndrome Respiratória Aguda Grave

SARS-CoV-2 – Coronavírus 2 da Síndrome Respiratória Aguda Grave

(+)ssRNA – RNA fita simples senso positivo

tRNA – RNA transportador

TTC – Tempo de Tradução por Códon

VIPR – Virus Pathogen Resource

VOC – Variante de preocupação internacional

VUM – Variante sob monitoramento

YFV – Vírus da Febre Amarela

ZIKV – Zika Vírus

SUMÁRIO

1. Introdução

1.1. Emergência e reemergência viral

A economia e a saúde da população mundial têm sido afetadas por doenças infecciosas emergentes e reemergentes cujos principais registros, desde 1990, incluem doenças como "Gripe Espanhola", Febre Amarela, Dengue, Febre do Nilo Ocidental, Febre Hemorrágica Ebola, HIV/Aids, entre outras (HUANG; HIGGS; VANLANDINGHAM, 2019). Nesse contexto, a reemergência refere-se à detecção de um vírus já conhecido em uma nova população hospedeira (HUANG; HIGGS; VANLANDINGHAM, 2019). Esses eventos são ocasionados por dois fatores: ecológicos, como invasão de áreas de *habitat* natural dos vetores, e genéticos, mediante mutações adquiridas pelos vírus (GEOGHEGAN; HOLMES, 2017).

As mutações são um processo natural que ocorrem durante a replicação do material genético. Todos os seres apresentam uma taxa de mutação. Os genomas virais, porém, revelam uma elevada taxa de mutação, variando de 10^{-3} a 10^{-6} por nucleotídeo, para cada cópia do genoma, ou seja, uma das mais altas conhecidas, principalmente devido à ausência do mecanismo de correção da polimerase e à necessidade de uso do maquinário de transcrição de diferentes hospedeiros (DENNEHY, 2017).

Uma alta taxa de mutação provoca instabilidade ao processo de replicação viral. Apesar de algumas vezes gerar uma vantagem evolutiva, o número alto de mutações geralmente causa efeitos deletérios. Como resultado, apenas uma pequena fração da população viral realmente adquire características de sobrevivência eficazes ao hospedeiro e, para que esse novo genótipo sobreviva, a taxa de transmissão precisa ser elevada (GEOGHEGAN; HOLMES, 2017).

Nos últimos 40 anos, ocorreram vários eventos de emergências virais em humanos (Figura 1), tendo como destaque os vírus: Influenza A H1N1/09, Vírus da Imunodeficiência Humana (HIV), SARS Coronavírus, MERS Coronavírus e Vírus Ebola (EBOV). Como resultado, há um crescente número de dados na literatura buscando compreender e predizer os seus eventos de transmissão (IBRAHIM *et al.*, 2018).

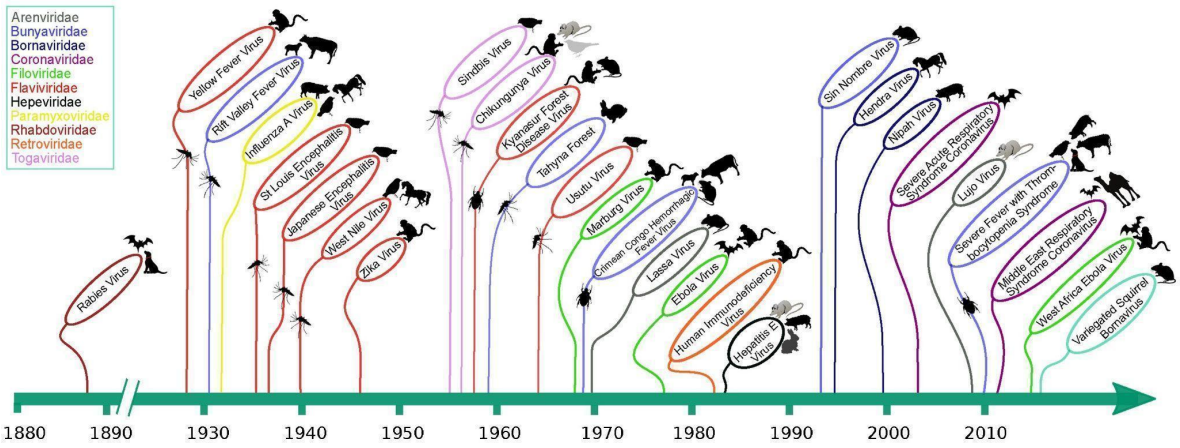

Figura 1. Patógenos virais emergentes e reemergentes. Os hospedeiros conhecidos estão ilustrados no ramo, e famílias estão representadas em cores. Fonte: Ibrahim *et al*., (2018).

A adaptação a uma nova espécie de hospedeiro representa também um grande desafio biológico, e não é sempre bem-sucedida pelo vírus, como, por exemplo, o H5N1, o vírus da Influenza Aviária, que apresenta registro de infecção em humanos, porém sem sucesso de continuidade e estabelecimento de um novo ciclo (LAM *et al.*, 2015). Por sua vez, alguns vírus tiveram a adaptação tão bem-sucedida a ponto de estabelecerem uma cadeia de transmissão somente em humanos, sem necessidade de outros hospedeiros, como o do vírus HIV (GEOGHEGAN; HOLMES, 2017).

Teoricamente, os vírus tendem a evoluir de modo que consigam coexistir com seus hospedeiros, dessa forma, garantindo a sua sobrevivência (DENNEHY, 2017). A virulência é um aspecto que não porta vantagens para a adaptação, como, por exemplo, o vírus Ebola (EBOV), sua transmissão se dá por contato direto com um indivíduo infectado ou suas secreções. EBOV tem uma taxa de mortalidade alta para humanos, em torno de 95%, como resultado seu potencial de transmissão é sustentado. Por esse motivo, as ondas de infecções já registradas por Ebola são regionais (DIEHL *et al.*, 2016; DUDAS *et al.*, 2017; URBANOWICZ *et al.*, 2016).

Existe um padrão relativo à tendência de maior emergência de vírus de RNA. Esse aspecto tem uma relação com a alta taxa de mutação. Vírus que cruzaram a barreira das espécies e foram bem-sucedidos têm características em comum: pequenos genomas, alta taxa de mutação, alta diversidade genética e alta pressão seletiva (DENNEHY, 2017).

Um grande desafio no que se refere à predição desses eventos é que muitos hospedeiros

secundários ainda são desconhecidos (GEOGHEGAN; HOLMES, 2017).

1.2. A Vigilância Genômica no Brasil

A Vigilância Genômica tem por definição o objetivo de rastrear patógenos virais emergentes e reemergentes por meio do sequenciamento completo do seu material genético. Essa técnica se tornou fundamental na luta contra patógenos emergentes e reemergentes nos últimos anos, possibilitando o monitoramento de diferentes vírus zoonóticos e endêmicos (MAGALHAES *et al.*, 2020).

Os Arbovírus são um grupo de vírus transmitidos por artrópodes e os principais causadores de epidemias no Brasil. Eles são divididos em cinco famílias: *Flaviviridae*, *Togaviridae*, *Bunyaviridae*, *Reoviridae* e *Rhabdoviridae*. Dentre os arbovírus, os principais causadores de infecção no Brasil são: o vírus da Febre Amarela (YFV), o vírus Chikungunya (CHIKV), o vírus Zika (ZIKV) e o vírus da Dengue (DENV) (ZANOTTO; LEITE, 2018).

A distribuição global desses vírus está associada a áreas com presença de seus vetores, comumente sendo tropicais ou subtropicais. Grandes centros urbanos próximos a florestas, em países como o Brasil, viabilizam a fácil transição de ciclos virais entre silvestre e urbano, como consequência ocasionando várias epidemias (HUANG; HIGGS; VANLANDINGHAM, 2019).

A primeira introdução do vírus da Febre Amarela do Brasil foi em 1685. Nesse primeiro registro de infecção, houve principalmente casos em cidades portuárias da região Nordeste (COSTA *et al.*, 2011). Após quase um século sem notificações da doença, em 1850, mais um surto de YFV foi registrado na região Sudeste do Brasil, no estado do Rio de Janeiro, onde causou mais de quatro mil mortes (COSTA *et al.*, 2011). Depois de anos de registros esporádicos e pesquisa sobre YFV, em 1937, uma vacina foi finalmente desenvolvida e foram feitas campanhas de imunização em massa (DE OLIVEIRA FIGUEIREDO *et al.*, 2020).

O primeiro caso suspeito de Dengue no Brasil foi registrado em 1846. A primeira onda de infecções foi confirmada em 1981, causada pela circulação de dois sorotipos, o DENV-1 e DENV-4, no estado de Roraima, ao norte do Brasil (WERMELINGER *et al.*, 2016). Após quatro anos sem casos confirmados de dengue, o DENV-1 foi identificado no estado do Rio de Janeiro, na região Sudeste do Brasil, onde também foram registrados quatro casos fatais. Em 1990, o DENV-2 teve seu primeiro registro no estado do Rio de Janeiro, onde houve o primeiro caso de

febre hemorrágica registrado (NOGUEIRA *et al.*, 1990). Nos anos seguintes até 1999, vários casos de Dengue foram reportados ao longo de todo o país. Em 2000, o sorotipo DENV-3 foi identificado no Brasil, pela primeira vez, no estado do Rio de Janeiro. Esse sorotipo se espalhou rapidamente pelo país, causando, em 2002, uma das piores epidemias já registradas, com mais de 150 mortes confirmadas (DE ARAÚJO *et al.*, 2009). Em 2008, 17 anos após a introdução de DENV-2, houve uma reemergência desse sorotipo, com registros de mais de 500 mortes causadas por febre hemorrágica. Nos anos seguintes, em 2009 e 2010, foram registradas as reemergências de DENV-1 e DENV-4, respectivamente, com mais de 600 mortes registradas ao norte e ao sudeste do Brasil. Em 2015 foi registrada a reemergência e circulação dos quatro sorotipos ao longo do país, com uma epidemia de mais de um milhão de casos e mais de 900 mortes (NUNES, P. C. G. *et al.*, 2019). O DENV, desde seu primeiro registro, causou mais de dez milhões de infectados no período de 30 anos. Em 2021-2022, no Brasil, o DENV teve um total de 90.335 casos suspeitos, com aumento de 43,2% em relação ao ano anterior (PAHO, 2022).

O CHIKV teve duas introduções registradas no Brasil, entre os anos 2013-2014 quando a cocirculação de dois genótipos ou linhagens (asiático e africano) foi detectada em duas distintas regiões brasileiras, Norte (Oiapoque, Amapá) e Nordeste (Feira de Santana, Bahia), respectivamente (NUNES, M. R. T. *et al.*, 2015). Desde então mais de 6,002 casos prováveis foram registrados no país, com taxa de incidência de 2,8 por cem mil habitantes (PAHO, 2022).

O Zika Vírus foi identificado no Brasil, pela primeira vez, em maio de 2015, na região Nordeste, e, no ano seguinte, o vírus foi detectado em todo o território nacional (MUSSO; GUBLER, 2016). A epidemia de ZIKV se espalhou ao longo do país em poucos meses, atingindo um número alarmante de indivíduos infectados: 205.578 casos em 2016; e 13.353 em 2017 (LOWE *et al.*, 2018). De acordo com o último boletim epidemiológico publicado, entre 2021 e 2022, houve 323 casos prováveis, com taxa de incidência de 0,15 por cem mil habitantes (PAHO, 2022).

As várias emergências virais despertaram preocupação nacional para a vigilância genômica como meio de elucidar o ciclo de infecção utilizado pelo arbovírus (FIGUEIREDO, 2019; NUNES, P. C. G. *et al.*, 2019; MAGALHAES *et al.*, 2020).

Durante a pandemia de Covid-19, a vigilância genômica alcançou proporções nunca atingidas, possibilitando o monitoramento da evolução desse patógeno e a identificação de milhares de variantes/linhagens ao longo do tempo (FLORES-VEGA *et al.*, 2022). As diferentes

variantes foram classificadas pela Organização Mundial da Saúde (OMS), em diferentes grupos: VOI (do inglês, *Variant of Interest*), VUM (do inglês, *Variant Under Monitoring*), FMV (do inglês, *Formerly Monitored Variant*), e as VOCs (do inglês, *Variants of Concern*). As VOCs, atualmente, incluem um grupo de cinco variantes que apresentam uma constelação de mutações linhagem-específica, muitas das quais são localizadas no domínio RDB (do inglês, *Receptor-Binding Domain*) da proteína viral *Spike* (S), que desempenha um papel importante no reconhecimento do receptor celular ACE2. Esse conjunto de mutações confere a essas variantes uma elevada capacidade de transmissão, uma elevada infectividade e um mecanismo de escape da resposta imune do hospedeiro (DUMACHE *et al.*, 2022).

A VOC *Alpha* (B.1.1.7), foi a variante identificada no Reino Unido, em setembro de 2020. Após sua identificação, espalhou-se globalmente. Suas mutações foram relacionadas ao aumento da virulência e transmissibilidade (MENG *et al.*, 2021).

A variante *Beta* (B.1.315) foi identificada na África do Sul, em maio de 2020, e teve suas mutações associadas à neutralização de anticorpos monoclonais em humanos, levando a uma redução na eficácia da vacina em comparação aos resultados obtidos na variante *Alpha* (DUMACHE *et al.*, 2022).

A *Gamma* (P.1) é a VOC brasileira, identificada na cidade de Manaus, em novembro de 2020. *Gamma* apresenta mais de 17 mutações não sinônimas, associadas à evasão imune, ao aumento da afinidade com o receptor ACE2 e ao aumento de transmissibilidade (FARIA *et al.*, 2021). A *Gamma* teve um grande impacto após sua introdução no Brasil, sendo a responsável pela segunda onda de infecções registradas, de dezembro de 2020 a dezembro de 2021, período com o maior número de infectados e mortos registrados no país (BANHO *et al.*, 2022). Dois meses após sua primeira identificação, em janeiro de 2021, já se tinha tornado predominante entre os registros ao longo do país, com mais de 85% de dominância dos registros de SARS-CoV-2 (DEJNIRATTISAI *et al.*, 2021).

A variante *Delta* (B.1.617.2) foi identificada na Índia, em outubro de 2020, e suas mutações são concentradas, em maioria, na região *spike,* como resultado, *Delta* apresenta maior patogenicidade e transmissibilidade (MITTAL; KHATTRI; VERMA, 2022).

A variante *Omicron* (B.1.1.529) teve seu primeiro registro em Botsuana e na África do Sul, em novembro de 2021, e suas mutações estão associadas à maior transmissibilidade, afinidade com receptor hospedeiro e infectabilidade (VIANA *et al.*, 2022).

1.3. Coronavírus

Os Coronavírus (CoVs) pertencem à subfamília *Coronavirinae*, família *Coronaviridae*. São vírus com única fita de RNA e um nucleocapsídeo (estrutura composta pelo ácido nucleico do vírus e seu invólucro proteico, o capsídeo) helicoidal (ALANAGREH; ALZOUGHOOL; ATOUM, 2020). Seu nome se deve à presença de espículas (estruturas proeminentes) presentes na superfície do nucleocapsídeo vírus, o que lhe dá a aparência de uma coroa solar (*corona* em latim)(MALIK, 2020).

Os Coronavírus são vírus relacionados principalmente às infecções do trato respiratório e gastrointestinal, sendo taxonomicamente classificados em quatro gêneros: Alphacoronavírus, Betacoronavírus, Gammacoronavírus e Deltacoronavírus (CHAN *et al.*, 2020a). Até o presente momento, foram descritas sete diferentes espécies de CoVs causadores de infecção em humanos: HCOV-SARS, HCOV-OC43, HCOV-NL63, HCOV-MERS, HCOV-229e, HCOV-HKU1 e o recém-descoberto SARS-CoV-2 (KHAILANY; SAFDAR; OZASLAN, 2020). O grupo pode também ser dividido em vírus endêmicos e epidêmicos de acordo com o registro de sua descoberta. O primeiro registro de infecção de Coronavírus em humanos foi reportado na década de 1960, quando HCOV-OC43 e HCOV-229e foram descritos como causadores de “resfriados” até então sem nenhuma complicação grave (HAMRE; PROCKNOW, 1966; MCINTOSH *et al*., 1967). O segundo registro endêmico de coronavírus que infectam humanos ocorreu em 2004-2005, quando foram descobertas as espécies HCOV-NL63 e HCOV-HKU1. Até esse ponto, os Coronavírus eram classificados como uma família de vírus endêmicos e associados a curtos resfriados, sendo equiparados ao vírus Influenza (VAN DER HOEK *et al*., 2004; WOO *et al*., 2005).

O primeiro caso de Síndrome Respiratória Aguda Grave (SARS) foi identificado em Foshan, na China, em novembro de 2002. Alguns meses depois, em julho de 2003, o vírus já se tinha espalhado em mais de 30 países, causando cerca de oito mil infecções e aproximadamente 800 mortes. Nove anos depois, uma nova espécie de Coronavírus foi identificada em uma amostra de pulmão de um paciente de 60 anos de idade, que morrera de insuficiência respiratória na Arábia Saudita, sendo essa a primeira notificação do MERS (Middle East Respiratory Syndrome, HCoV-MERS) Coronavírus. Na epidemia de 2012, na Arábia Saudita, o HCoV-MERS causou mais de 2.500 casos com um total de 861 mortes, com taxa de mortalidade

de 35% (CHAFEKAR; FIELDING, 2018; DE GROOT *et al.*, 2013; PERLMAN; NETLAND, 2009).

Os CoVs são vírus envelopados e com material genético de RNA de fita simples de polaridade positiva (+ssRNA), com genoma de 26 a 32 kilobases (kb) de tamanho. O RNA dos coronavírus contém múltiplas ORFs (*open read frames*). A maior ORF localizada na extremidade 5’ ocupa cerca de ⅔ do genoma, é composta por duas ORFs, que se sobrepõem denominadas ORF1a e ORF1b, responsáveis por codificarem sua poliproteína (CHEN; ZHONG, 2020). Outra principal ORF está localizada na extremidade 3’ e codifica quatro proteínas estruturais, representadas na Figura 2 e sumarizadas a seguir:

- *Spike* (S), com cerca de 150 kDa, é responsável pela fase de adsorção do vírus no receptor celular, o que resulta na fusão e entrada viral (CHEN; ZHONG, 2020).
- *Membrane* (M), com cerca de 30 kDA, é a proteína mais abundante do vírus e que dá forma ao envelope viral (CHEN; ZHONG, 2020).
- *Envelope* (E), com cerca de 12 kDA, é a menor proteína estrutural do vírus durante a replicação e é encontrada em grande quantidade na célula infectada. Porém, apenas, parte dela é incorporada ao envelope viral, sua função principal é ajudar na montagem e liberação de novas partículas virais (MALIK, 2020).
- *Nucleocapsid* (N), a única a conectar-se ao genoma viral, tem sua função associada à montagem e à liberação viral da célula infectada (CHEN; ZHONG, 2020).

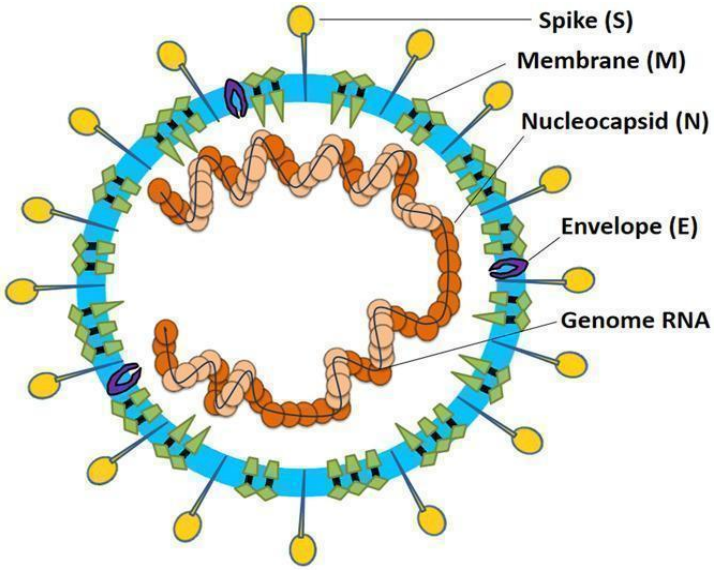

Figura 2. Estrutura genômica dos Coronavírus. Fonte: LI *et al*., (2020a).

Além das proteínas estruturais que são codificadas pela ORF 3', estão presentes regiões codificadores para proteínas acessórias como a proteína HE, a proteína 3a/b e a proteína 4a/b. Após a infecção viral e a fase de desnudamento, o RNA viral é traduzido em duas poliproteínas 1a e 1b, que, em seguida, são processadas em 16 proteínas não estruturais (NSPs), para a formação do complexo de replicação e tradução (RTC) (CHEN *et al*., 2020).

1.3.1 Covid-19

Em dezembro de 2019, a OMS recebeu a informação de casos de pneumonia na cidade Wuhan, província de Hubei na China, causados por um agente etiológico desconhecido (WHO, 2020). No dia 7 de janeiro de 2020, testes sorológicos identificaram um novo Coronavírus nesses pacientes, nomeado como 2019 *novel coronavirus* (2019-nCoV). Em 16 de janeiro de 2020, 43 pacientes foram diagnosticados com o 2019-nCoV, incluindo, entre estes, dois pacientes na Tailândia e no Japão. Logo após, em 23 de janeiro de 2020, 835 casos foram confirmados em mais de 32 províncias, municípios e regiões próximas a Wuhan, incluindo Hong Kong, Macau e Taiwan (WHO, 2020). Nessa mesma data, infecções em profissionais da saúde que tratavam pacientes foram reportadas e estudos relatam a transmissão direta humano-humano (CHAN *et al*., 2020b). No início do mês seguinte, em 11 de fevereiro de 2020, WHO nomeia a pneumonia

induzida pelo novo coronavírus como *Coronavirus Disease 2019* (Covid-19). Concomitantemente, a Comissão Internacional de Classificação Viral anuncia o novo nome para o vírus, antes nomeado de 2019-nCoV para *Severe Acute Respiratory Syndrome Coronavirus 2* (SARS-CoV-2).

A partir desses eventos citados, o vírus se tem disseminado e propagado progressivamente e, atualmente, encontra-se presente em mais de 200 países. O número de indivíduos infectados mundialmente ultrapassa 500 milhões de casos registrados e mais de seis milhões de mortes foram notificadas, até junho de 2022, pela Covid-19, de acordo com WHO (2022).

No dia 26 de fevereiro de 2020, o Ministério de Saúde do Brasil anunciou o primeiro caso confirmado de SARS-CoV-2 em solo nacional, identificado em um paciente de 61 anos, em São Paulo, no Hospital Albert Einstein. Desde o primeiro caso, o número de infectados no Brasil aumentou diariamente, contabilizando, em 20 de abril de 2022, 30.261.088 casos confirmados da doença. Com um total de 662.026 mil óbitos associados à doença, além disso o país já registrou uma das médias de morte diárias mais altas do mundo, com aproximadamente duas mil mortes (WHO, 2022).

Os primeiros registros de sintomas de Coronavírus eram similares aos dos vírus Influenza em humanos, sendo, muitas vezes, associados à gripe (FUNG; LIU, 2019). A partir da descoberta dos Coronavírus epidêmicos, em 2002 (SARS-CoV e HCoV-MERS), os sintomas descritos têm sido mais severos. Os principais sintomas da Covid-19, de acordo com a OMS, são: febre, tosse seca e cansaço. Os sintomas mais graves são: complicações respiratórias e pneumonia, podendo levar a óbito (WHO, 2022). Os sintomas mais severos representam maior risco em grupos mais suscetíveis, como pacientes imunodeprimidos, idosos e portadores de doenças crônicas, em que possa ocorrer o desenvolvimento de complicações respiratórias e em outros órgãos (ALANAGREH; ALZOUGHOOL; ATOUM, 2020). Cerca de 81% dos pacientes infectados desenvolvem os sintomas considerados principais, como uma leve pneumonia ou nenhuma. Porém de 5 a 14% desenvolvem sintomas mais críticos e graves, próximos à pneumonia e bronquite. Além disso, a quantidade de quadros assintomáticos ou com sintomas clínicos leves permite que a infecção se dissemine de forma silenciosa, contagiando e expondo uma grande parcela da população (LI *et al*., 2020).

Diversos estudos buscaram a caracterização molecular do vírus assim que foi descoberto, e, a partir de abordagens de sequenciamento genômico, foi notada a alta similaridade genômica

(79.6%) com SARS-CoV, sendo, portanto, nomeado, posteriormente, como SARS-CoV-2. Estudos apontam também, em análises de genômica comparativa, que SARS-CoV-2 de isolado humano tem uma identidade de 96% a uma sequência do coronavírus isolado de morcego. E a análise proteica, considerando o alinhamento de sete domínios de proteínas não estruturais, revelou que o vírus pertence à espécie SARS-CoV (LU *et al*., 2020; WU *et al.*, 2020; ZHOU *et al*., 2020; ZHU *et al.,* 2020).

Análises transcricionais do genoma do vírus permitiram identificar 380 substituições de aminoácidos que podem ter contribuído para modificações funcionais e diferenças na patogenia observadas no SARS-CoV-2 (WU *et al.*, 2020). A partir de estudos genéticos comparativos também foi possível confirmar relações evolutivas com outras linhagens virais encontradas na natureza, descartando teorias conspiratórias envolvendo manipulação artificial e sugerindo a possível origem relacionada a morcegos e pangolins (ZHOU *et al.*, 2020).

Trabalhos recentes sugerem que o SARS-CoV-2 seja mais contagioso que as linhagens relacionadas a surtos de SARS e MERS, podendo ocorrer transmissão na fase assintomática e pré-sintomática da infecção (WÖLFEL *et al*., 2020; ZHANG; HOLMES, 2020). Embora evidências indiquem que a taxa de casos fatais para SARS-CoV-2 seja inferior à observada em infecções por Influenza, essa taxa pode ser alterada à medida que novos estudos demonstrem a real proporção de casos.

A taxa de mutação estimada para os Coronavírus é considerada entre moderada à alta para vírus (+)ssRNA (ALANAGREH; ALZOUGHOOL; ATOUM, 2020). Dois *loci* são considerados sítios com maior variação para SARS-CoV. O primeiro é identificado na sequência da proteína *spike* (S) e o segundo sítio de variação é localizado na ORF8. Na proteína S, há três pequenas inserções no domínio N-terminal, e quatro/cinco alterações nos resíduos do *motif* da proteína receptora (ZHOU *et al*., 2020). A organização de outras regiões do genoma e expressão de genes é bem próxima a outros Coronavírus, conforme Figura 3.

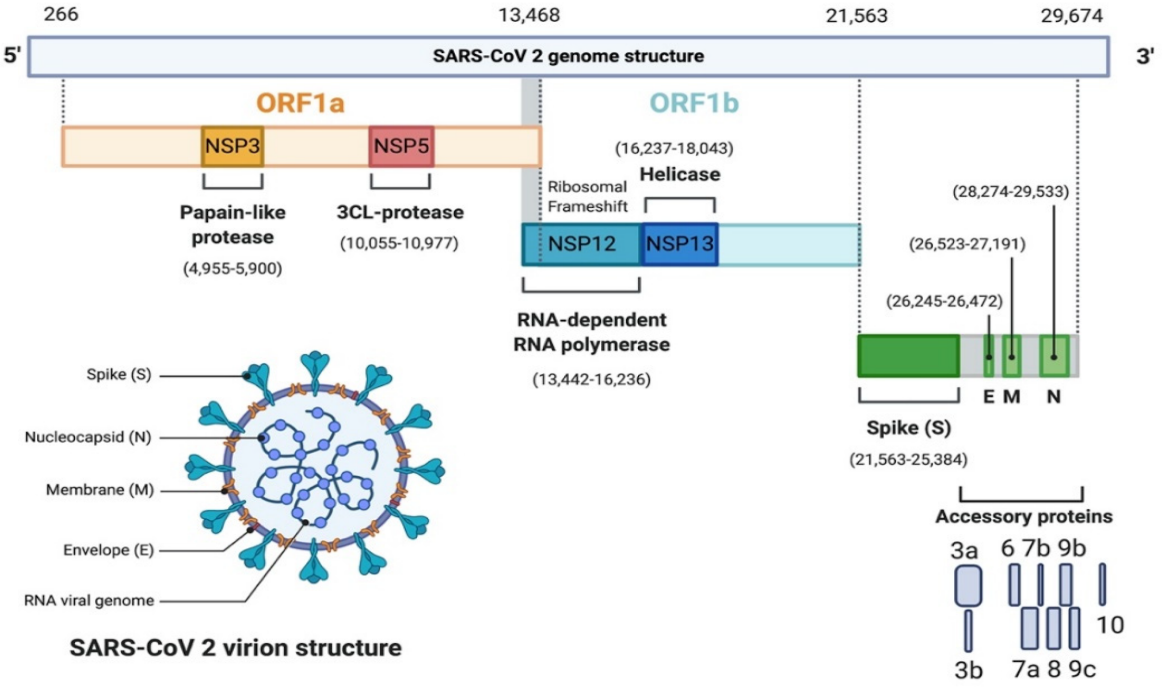

Figura 3. Organização do genoma do vírus SARS-CoV-2. Fonte: Alanagreh *et al.,* (2020)

Os Coronavírus são considerados patógenos zoonóticos porque possuem uma origem animal e podem ser transmitidos através de um contato direto para humanos. Todos os CoVs, incluindo SARS-CoV-2, têm como origem mais provável morcegos (CORMAN *et al*., 2018). No entanto é um problema comum um Coronavírus cruzar a barreira de espécies e causar surtos em humanos, como ocorrido em 2003 e 2012, após HCOV-SARS e HCOV-MERS infectarem hospedeiros mamíferos intermediários, *Paguma larvata* e *Camelus dromedarius*, respectivamente, e, em seguida, humanos (CHAN *et al*., 2020a). Dessa forma, o vírus adquiriu mutações que permitiram melhor adaptação ao hospedeiro humano.

É possível acenar à hipótese de que o SARS-CoV-2 tenha adquirido mutações em um hospedeiro intermediário, ainda não identificado, que desencadeou vantagens adaptativas, com isso, culminando na atual pandemia de Covid-19 (FUNG; LIU, 2019).

A partir da análise da composição nucleotídica do vírus é possível investigar o processo de adaptação do genoma viral para outros hospedeiros, em especial, o humano. Uma estratégia promissora é conhecida como *Codon Usage Bias* (CUB), que será apresentada na próxima seção.

1.4. A Bioinformática no estudo de patógenos virais emergentes e reemergentes

O estudo da emergência viral e como ela pode ser predita se tornou um importante tópico na área da ciência. As tecnologias de sequenciamento de segunda geração permitiram a geração de dados em grande quantidade e o depósito em repositórios públicos de dados brutos do sequenciamento, como o *Sequence Reads Archive* (SRA) (https://www.ncbi.nlm.nih.gov/sra) e de dados processados, como genes e genomas no *GenBank* (https://www.ncbi.nlm.nih.gov/genbank/). Pela aplicação da bioinformática a esses dados foi possível o desenvolvimento de um sistema digital de vigilância genômica global. A análise rápida e informativa é possível devido ao tamanho relativamente pequeno do genoma viral, o que permite o sequenciamento no contexto clínico de pacientes.

A bioinformática atua na vigilância genômica por meio da disponibilidade de ferramentas que elucidem questões a respeito de identificação, rastreamento de patógeno, relações filogenéticas, resistência a fármacos, mutações e virulência (IBRAHIM *et al.*, 2018).

Os *pipelines* e programas de bioinformática consistem geralmente em *scripts* de programação de uma linguagem que se adapte ao problema biológico em questão. Portanto, esses *pipelines* podem ser executáveis prontos ou desenvolvidos pelo pesquisador, normalmente, divididos em duas etapas: processamento dos dados e análise (GIOVANETTI *et al.*, 2020). Na sequência estão listados alguns exemplos de bioinformática aplicados à vigilância genômica:

- Banco de dados: ViPR (https://www.viprbrc.org/), HIV database (https://hivdb.stanford.edu/) e GISAID (https://www.gisaid.org/).
- Montagem de genomas: SOAP (LI *et al.*, 2008) e SPAdes (BANKEVICH *et al.*, 2012).
- Genômica comparativa: EDGAR (DIECKMANN *et al.*, 2021), GET_HOMOLOGUES (CONTRERAS-MOREIRA; VINUESA, 2013).
- Filogenia: Adapatch (TUSCHE; STEINBRÜCK; MCHARDY, 2012) e AntigenicTree (SAITOU; NEI, 1987).

Os algoritmos desenvolvidos trazem à luz as possibilidades de interpretação de dados baseados em ferramentas da informática. No entanto novas implementações são carentes de desenvolvimento, abordando, por exemplo: RNAi, elementos de inserção, epigenética, pseudogenes e elementos de estruturas conservadas (GIOVANETTI *et al.*, 2020). O número alto de mutações nos genomas virais é um desafio particular para a bioinformática e extração dessas

informações biológicas diferenciando erros no sequenciamento de mutações reais. Como resultado, tem-se a necessidade de desenvolvimento de melhores algoritmos e programas para análise. A análise completa da emergência viral requer novos mecanismos de integração de dados genéticos e ecológicos. Na próxima seção deste trabalho é apresentada uma abordagem de bioinformática que visa contribuir no contexto genético da emergência viral.

1.5. *Codon Usage*

Em todos os organismos vivos, cada célula obedece ao código genético, onde se estabelece que cada códon codifica para seu aminoácido correspondente. Existe um total de 20 aminoácidos que são codificados por 61 códons, indicando a presença de um código genético degenerado. Os códons que codificam para um mesmo aminoácido são chamados códons sinônimos. Em um cenário ideal, cada códon sinônimo seria usado de forma igual. Todavia é observado um desbalanceamento no uso dos códons sinônimos no processo de tradução.

O desbalanceamento no uso dos códons por organismo é chamado de C*odon Usage Bias (CUB)*, é uma observação relativa associada a uma tendência na seleção de códons (SHARP; TUOHY; MOSURSKI, 1986). *Relative Synonymous Codon Usage* (RSCU) é a medida quantitativa do uso desigual desses códons sinônimos em um determinado organismo. Cada vírus exibe um diferente valor de RSCU, dependendo do tipo de hospedeiro infectado (JENKINS; HOLMES, 2003; SHARP; TUOHY; MOSURSKI, 1986).

$$R = \{RSCU_{i,j} | \; i = 1, 2,..., 20, \; j = 1, 2,..., n_i\}$$

é uma matriz de *Relative Synonymous Codon Usage*. Considerando um aminoácido *i* codificado por n_i códons sinônimos, uma distribuição de códons sinônimos uniforme implicaria em uma mesma proporção, ou seja $1/n_i$ na sequência. Contudo o observado são códons com maiores valores, classificados como códons preferidos. A premissa desta escolha é que seria alcançada uma vantagem traducional em relação ao hospedeiro, por esse motivo os códons preferidos são chamados *códons ótimos* (SHARP *et al.,* 1986). Para mensurar a uniformidade de distribuição dos códons usa-se a seguinte fórmula (SHARP *et al.,* 1986):

$$RSCU_{i,j} = n_i \frac{x_{i,j}}{\sum_{k=1}^{n_i} x_{i,k}}, \quad i = 1, 2, \ldots, 20$$

onde i indica o aminoácido, $x_{i,j}$ representa o número de ocorrências do jth códons sinônimos para o aminoácido ith codificado por n_i códons sinônimos. Para obter n_i para cada um dos 20 aminoácidos (aa) é utilizado o código genético: 2 aa, Metionina e Triptofano são codificados por único códon; 9 aa por 2 códons; Isoleucina por 3 códons; 5 aa por 4 códons e 3 aa por 6 códons sinônimos.

No contexto de vigilância genômica, o CUB pode ser utilizado para sugerir potenciais novos hospedeiros. Li *et al.* (2022) encontraram evidências de potenciais hospedeiros comparando o *codon usage* de 17 animais com o do vírus Hepatite E (HEV-1). Malhotra e Kumar (2021), a partir do estudo dos padrões de uso de códons dos vírus Chapare e Sabia, avaliaram o potencial de adaptação ao hospedeiro humano. He *et al.* (2022) compararam uso de códons na adaptação dos *Narcisus degeneration virus* (NDV), *narcisus late season yellows virus* (NLSYV), *narcissus yellow stripe virus* (NYSV), como assinaturas filogenéticas.

Existem vários fatores que influenciam o RSCU apresentado por determinado organismo. Entre eles a pressão seletiva, a composição genômica, as forças relativas à evolução dos genes e outros (FRIAS *et al*., 2013).

1.5.1 *Codon usage* e a relação com disponibilidade de RNAs transportadores (tRNAs)

O CUB é uma metodologia que leva em consideração a coevolução do uso dos códons, considerando que os códons sinônimos são adaptados à disponibilidade do seu meio (SHARP; TUOHY; MOSURSKI, 1986), enquanto a quantidade de cópias de genes de tRNAS no genoma é considerada impulsionada pela frequência relativa de códons sinônimos (SMITH, 1978).

No entanto, o cálculo de RSCU não leva em consideração o tamanho e a composição do *pool de* tRNAs disponível para a tradução do mRNA da célula hospedeira. A tradução do mRNA tem quatro principais etapas:

1) **Iniciação**: uma molécula livre de mRNA se liga a um ribossomo livre. Logo após, mais de 12 fatores de tradução eucarióticos - eIFs, somam-se ao complexo. Assim, a tradução

começa a partir do primeiro códon que codifica para o aminoácido Metionina (HINNEBUSCH; LORSCH, 2012).

2) **Alongamento:** cada códon exposto no sítio de tradução do ribossomo (RTS) espera por seu tRNA complementar. Após a chegada, é realizada a ligação e o próximo códon ocupa o RTS. Esse processo continua até o último códon, o códon de terminação, presente no mRNA. Fatores de alongamento eucarióticos (eEFs) participam desse processo (DEVER; GREEN, 2012).
3) **Terminação:** o complexo formado no ribossomo reconhece um *stop-codon*, que é auxiliado por fatores de liberação eucarióticos (eRFs). Então, a molécula de mRNA e o polipeptídio são liberados (NÜRENBERG; TAMPÉ, 2013).
4) **Reciclagem do tRNA:** a liberação do tRNA pode ser reaproveitada para participação em outro processo de tradução (NÜRENBERG; TAMPÉ, 2013).

No contexto traducional do hospedeiro, existem sete principais recursos pelos quais os mRNAs do hospedeiro e viral devem competir: ribossomos, fatores de iniciação, fatores de alongamento, fatores de liberação, fatores de aminoacilação, maturação do tRNAs e o aminoácido correspondente (FRIAS *et al*., 2013).

As etapas da tradução dependem dos recursos disponíveis. Se houver uma limitação na disponibilidade de um tRNA, um ribossomo ativo pode parar seu processo de alongamento, passando para um estado indisponível e terminando precocemente a proteína. Nesse contexto, o tempo de tradução por códon (TTC) é aumentado.

Existem duas opções para o vírus acessar os recursos disponíveis na célula hospedeira:

1. Competir pelos recursos mais abundantes, mas com grande demanda.
2. Disputar recursos menos abundantes, porém com menor demanda.

Em ambos os casos, há uma competição, pois o mRNA viral também utilizará os mesmos recursos da célula hospedeira. Para definir qual estratégia o vírus usa, é necessário mapear a disponibilidade de tRNA no hospedeiro para cada códon.

1.6. Justificativa

A Covid-19 é a terceira e maior epidemia/pandemia causada por Coronavírus nos últimos 20 anos. O número de casos no mundo ultrapassou 500 milhões e o número de mortes, seis milhões (WHO, 2022). O Brasil tem, em Fevereiro de 2022, mais de 600 mil mortos devido à Covid-19 e fica atrás apenas dos Estados Unidos no número total de infectados. O efeito econômico de uma epidemia em um país em desenvolvimento se revelou devastador, e a falta de recursos e controle fez do Brasil um dos maiores epicentros da doença. Todos esses acontecimentos revelaram a necessidade de mais estudos sobre emergência viral.

Dois aspectos principais estão ligados à emergência viral: a ecologia e a genética viral. A ecologia tem uma relação no estabelecimento do contato entre hospedeiros primários e secundários, o que pode viabilizar a evolução de um vírus multi-hospedeiro. A teoria mais aceita sobre o início da Covid-19 é que o SARS-CoV-2, agente etiológico dessa doença, sofreu uma mutação em um hospedeiro secundário, ainda não identificado, para assim adquirir o fitness ao hospedeiro humano (FLORES-VEGA et al., 2022). Alguns fatores ecológicos podem ter contribuído para isso: alterações na demografia, agricultura, mudança climática, ocupação de áreas e desmatamento. O aumento do contato entre o vírus e seus diferentes hospedeiros pode gerar o novo fitness para um deles ou ambos.

Uma vez estabelecido seu ciclo em humanos, um importante meio de estudo do comportamento viral é através de seu genoma (DE SOUZA et al., 2022). O SARS-CoV-2 é um vírus de RNA, com alta taxa de mutação (DUMACHE et al., 2022). O número de variantes que continua diariamente a ser identificado, desde o início da pandemia em 2019, é o principal desafio científico para o controle da Covid-19 no mundo. O comportamento de um patógeno zoonótico é complexo porque é preciso se adaptar a vários hospedeiros, com recursos biológicos diferentes. Por consequência, o genoma do SARS-CoV-2 tem sofrido mutações desde sua descoberta, e a cada nova variante emergente aumenta o risco de evasão da resposta imune gerada pela vacina.

Nesse sentido, durante a pandemia o mundo científico passou a trabalhar com vigilância genômica de forma intensa e colaborativa, gerando um alto volume de dados moleculares. Mediante a vigilância genômica é possível identificar cada variante em tempo real, refazer seu rastro e identificar sua origem. A análise desses dados só é possível devido à bioinformática.

Nesse contexto, a genômica comparativa é o campo que tem obtido os melhores resultados para o entendimento de como o vírus se adapta ao hospedeiro (OUDE MUNNINK et al., 2021), através por exemplo do estudo de códons. Esse tipo de análise pode contribuir muito para o estabelecimento de precisas informações acerca de: mutações e suas regiões mais suscetíveis no genoma viral, novos hospedeiros baseados na filogenia, modelo transcricional sugerido e reconstrução filogenética que ilustra, ao longo do tempo, o modelo evolutivo viral perante diferentes hospedeiros.

Dada essa situação e a variabilidade genética dos vírus emergentes e reemergentes, o presente trabalho buscou estudar o comportamento viral e sua relação com os hospedeiros, usando como modelo o SARS-CoV-2 e o hospedeiro humano durante uma infecção. A partir desta análise, espera-se esclarecer os mecanismos de adaptação ao novo hospedeiro e possíveis novas espécies em risco de infecção; além de avaliar o efeito da introdução de cada genótipo viral diferente, ao longo do tempo de pandemia no Brasil, e sua correlação com o número de mortes.

2. Objetivos

2.1 Geral

Analisar *in silico* o comportamento do genoma do SARS-CoV-2 e a sua relação com o hospedeiro humano.

2.2 Específicos

- Correlacionar a medida do uso de códons e tRNAs hospedeiros, com especial foco na taxa relativa ao vírus SARS-CoV-2 e suas implicações;

- Comparar o uso de códons entre os Betacoronavírus e o hospedeiro humano com base na distância euclidiana;

- Estimar o impacto das Variantes de Preocupação em relação ao número de casos de infectados e óbitos no Brasil.

3. Metodologia

As sequências utilizadas neste trabalho foram selecionadas e divididas em dois bancos de dados. Em seguida, foram utilizadas diversas métricas de cálculo de composição de sequências e correlações estatísticas, todas as análises foram realizadas utilizando *scripts* desenvolvidos na linguagem perl e R.

3.1. Montagem dos banco de dados

Inicialmente, foi feita a coleta de todas as sequências de referência para os coronavirus que tem registro em humanos, 2 alphacoronavirus e 5 betacoronavírus, Estas sequências foram usadas para estudos comparativos de composição genômica. Em seguida, foram coletadas todas as 14 sequências de referência juntamente com a anotação de regiões genômicas, para os Betacoronavirus disponíveis no NCBI (*The national center for biotechnology information*). Para a análise da seleção de códons foi feita a separação de cada *Open reading frame* (ORF). Todas as sequências usadas neste trabalho estão identificadas na tabela 1.

Para o cálculo do *Codon Usage Bias (*CUB) dos hospedeiros, houve a limitação de uso à apenas aqueles com os valores de codons disponíveis publicamente, para este fim foi usado o banco de dados Kazusa (http://www.kazusa.or.jp), como resultado foram usadas 8 organismos disponíveis; são eles: *Homo sapiens, Bostaurus, Canis familiaris, Equus caballus, Felis catus, Mus musculus, Mustela putorius furo, Rattus sp*.

Coronavírus	NCBI-ID
Bat Hp-betacoronavirus/Zhejiang2013	NC_025217.1.
Betacoronavirus HKU24	NC_026011.1
Bovine coronavirus	NC_003045.1
Betacoronavirus Erinaceus/VMC/DEU/2012	NC_039207.1
Human coronavirus HKU1	NC_006577.2
Rabbit coronavirus HKU14	NC_017083.1
Tylonycteris bat coronavirus HKU4	NC_009019.1
Rousettus bat coronavirus HKU9	NC_009021.1
Human coronavirus MERS	NC_019843.3
Human coronavirus OC43	NC_006213.1
Pipistrellus bat coronavirus HKU5	NC_009020.1
Rat coronavirus Parker	NC_012936.1
Human coronavirus SARS	NC_004718.3
Human coronavirus SARS-CoV-2	NC_045512.2
Coronavirus 229e	NC_002645.1
Coronavirus NL63	NC_005831.2

Tabela 1 - Identificação das sequências de referência utilizadas.

A análise de disponibilidade tRNA do hospedeiro por codon viral foi realizada através dos valores obtidos do banco de dados (http://gtrnadb.ucsc.edu/), tabela 2. Estes dados foram curados e deram origem ao banco de dados para a análise dos codons e tRNA's, utilizados na investigação genômica dos Betacoronavírus.

CODON	TRNA	Homo sapiens	Rattus sp.	Mustela putorius	Mus musculus	Felis catus	Equus caballus	Canis familiaris	Bos taurus
TTT	AAA	0	46	25	21	16	0	9	403
CTT	AAG	13	673	102	26	84	8	77	59
GTT	AAC	12	1337	12	214	56	15	12	146
ATT	AAT	19	134	18	78	30	25	20	504
TCT	AGA	11	1989	19	180	29	12	60	253
CCT	AGG	11	72282	100	76	91	10	459	39
GCT	AGC	38	68445	20	1628	2835	28	61	106
ACT	AGT	10	2183	19	261	41	10	27	163
TGT	ACA	1	77	130	29	530	3	111	22569
CGT	ACG	8	113	275	10	4918	11	534	451
GGT	ACC	0	863	17	117	127	0	26	1529
AGT	ACT	1	2412	52	961	98	4	42	345
TAT	ATA	1	25	209	12	57	1	121	1707
CAT	ATG	0	119	10712	5	1030	1	3687	309
GAT	ATC	1	415	57	63	70	8	28	240
AAT	ATT	2	276	195	192	342	2	241	80
TTC	GAA	18	33	22	22	43	13	33	3181
CTC	GAG	0	21	53	12	90	1	146	204

GAC	GAC	0	77	7	58	50	3	17	438
ATC	GAT	6	276	13	327	25	1	14	118
TCC	GGA	1	104	18	28	42	2	25	2218
CCC	GGG	2	372	54	20	71	2	589	64
GCC	GGC	2	627	10	257	199	1	135	346
ACC	GGT	0	266	40	178	41	0	37	56
TGC	GCA	36	420	251	226	751	33	507	125998
CGC	GCG	0	61	124	37	554	1	4543	2052
GGC	GCC	15	701	45	369	177	12	243	20442
AGC	GCT	9	37631	58	24388	107	29	99	2347
TAC	GTA	16	39	63	32	79	18	296	7178
CAC	GTG	10	78	1176	47	557	15	13778	1635
GAC	GTC	19	99	38	70	71	49	114	1850
AAC	GTT	41	2807	296	2509	866	19	376	244
TTG	CAA	8	701	294	1911	393	8	288	1430
CTG	CAG	10	122	4532	30	1809	11	2144	132
GTG	CAC	22	99	80	95	145	32	73	269
ATG	CAT	23	52	482	61	508	43	309	110
TCG	CGA	4	68	128	38	793	10	117	488
CCG	CGG	4	487	451	8	381	6	248	20
GCG	CGC	5	161	22	41	35	10	11	147

ACG	CGT	6	39	126	28	337	13	100	22
TGG	CCA	8	54	426	33	987	14	437	11504
CGG	CCG	4	9	523	3	1289	12	346	315
GGG	CCC	13	44	79	39	224	36	89	11002
AGG	CCT	7	246	1208	194	2297	14	1157	1028
TAG	CTA	0	22	229	11	554	15	374	906
CAG	CTG	23	36	1056	28	2169	37	4789	326
GAG	CTC	13	139	828	115	2418	891	801	1930
AAG	CTT	26	283	15264	243	45221	98	19381	266
TTA	TAA	6	20	198	27	481	5	176	187
CTA	TAG	4	106	1426	22	2393	8	1163	33
GTA	TAC	7	49	41	43	54	7	33	312
ATA	TAT	5	39	47	25	103	6	51	65
TCA	TGA	5	77	262	60	1064	5	348	427
CCA	TGG	9	197	1662	23	4707	13	2541	39
GCA	TGC	11	432	49	225	124	12	60	1165
ACA	TGT	6	79	93	75	197	9	131	131
TGA	TCA	2	20	38872	14	164230	18	20021	4298
CGA	TCG	6	13	64416	7	90971	9	12120	163
GGA	TCC	11	94	2472	102	3697	19	1651	9374
AGA	TCT	6	414	2842	408	3414	15	1738	1208

TAA	TTA	2	10	1616	8	3184	0	2112	407
CAA	TTG	10	21	25522	19	48137	18	46856	253
GAA	TTC	17	78	468	74	639	75	722	6132
AAA	TTT	20	72	1241	90	2474	25	2101	653

Tabela 2 - Contagem de tRNA's por hospedeiro.

Para a análise das variantes de impacto no Brasil, foram utilizadas sequências de SARS-CoV-2 coletadas do GISAID (gisaid.org) disponíveis em 19 de Fevereiro de 2022. Como controle de qualidade, apenas genomas com mais de 29.000 bp e menos de 1% de ambiguidade foram mantidos, tendo como resultado um banco de dados com o número final de 111.626 sequências. Agregado as sequências, os metadados associados também foram avaliados para a seleção final. Além disso, as informações de casos diários de SARS-CoV-2 no Brasil foram obtidas através do repositório covid.saude.gov.br, na mesma data citada.

3.2. Análise dos Códons

Uma série de métricas relacionadas aos códons foram usadas neste trabalho, e são descritas abaixo:

- $N = \{n_1, n_2, ..., n_{64}\}$ um vetor que conta cada códon começando pela primeira base da *coding sequence* (CDS). Para isso, uma matriz numérica atribuindo um valor para cada códon foi criada, o número total foi calculado da seguinte maneira:

$$n_{all} = \sum_{i=1}^{64} n_i$$

- $F = \{f_1, f_2, ..., f_{64}\}$ um vetor de códons com a frequência relativa, onde:

$$f_i = \frac{n_i}{n_{all}}$$

- $D = \{d_{1,1}, d_{1,2}, ..., d_{4,4}\}$ uma matriz *4x4* da frequência de dinucleotídeos onde os nucleotídeos são enumerados: *A=1, G=2, C=3* e *T=4*. Essa medida foi calculada de acordo com (KARLIN; BURGE, 1995):

$$d_{i,j} = \frac{n_{i,j}}{n_i n_j}$$

onde $n_i n_j$ é a frequência esperada de dinucleotídeos ij, e $n_{i,j}$ é o número de ocorrências de dinucleotídeos *ij* na CDS. Se $d_{i,j}$ é < 0.78 ou $d_{i,j}$ > 1.23, é considerado que o dinucleotídeo

ij está sub ou super-representado. Esta conclusão é correlacionada com a associação aleatória de nucleotídeos combinada com sua abundância relativa (BUTT et al., 2016).

- $R = \{RSCU_{i,j} | \ i = 1, 2,..., 20, \ j = 1, 2,..., n_i\}$ uma matriz com *Relative Synonymous Codon Usage*, que é uma medida não uniforme de uso de códons sinônimos. Considerando um aminoácido *i* traduzido por n_i códons sinônimos, um códon usage uniforme seria onde os códons sinônimos são usados na mesma proporção na sequência $1/n_i$. Porém, como o códon usage é desigual, usa-se a seguinte fórmula para a quantificação (SHARP; TUOHY; MOSURSKI, 1986):

$$RSCU_{i,j} = n_i \frac{x_{i,j}}{\sum_{k=1}^{n_i} x_{i,k}}, \quad i = 1, 2,..., 20$$

onde *i* indica o aminoácido, $x_{i,j}$ representa o número de ocorrências do *jth* códon sinônimo para o *ith* aminoácido que é codificado por n_i códons sinônimos. Para se obter n_i para cada um dos 20 aminoácidos (aa) foi considerado o uso do código genético: 2 aa, Metionina e Triptofano, traduzidos por um único códon, 9 aa por 2 códons, Isoleucina por 3 códons, 5 aa por 4 codons and 3 aa por 6 códons sinônimos. O valor máximo encontrado no cálculo de RSCU é relativo ao número de codon sinônimos, na ausência de *codon bias* o RSCU é igual a 1. RSCU menor do que 1 representa codons usados menos frequentes, e maior que 1 usados mais frequentes.

- $r_i = \frac{f_{i,SARS-2}}{f_{i,host}}, \ i = 1, 2,...., 61$ adaptando algumas variáveis da fórmula do item acima, foi calculada a correlação dos codons menos usados pelo hospedeiro humanos e mais utilizados para o vírus SARS-CoV-2.
- $ED = \{ed_{species\ i,\ species\ j} | \ i \neq j, \ i, j = human, SARS1, HKU1,..., SARS2\}$ foi calculado uma matriz com a distância euclidiana entre RSCU e os sete modelos virais de estudo e o hospedeiro humano. A distância Euclidiana é um algoritmo que pode ser usado para facilitar a identificação de relações entre as sequências, como já descrito em estudos prévios (JI et al., 2020; VAN HEMERT et al., 2018).

- O valor de RSCU humano foi calculado usando um script em R, a partir da frequência de códons baixado do Kazusa. Indicado por $R_{species\ i}$ a matriz de RSCU de *ith* espécies (descrito acima), a distância euclidiana foi calculada por:

$$ed_{species\ i,\ species\ j} = \left\| R_{species\ i} - R_{species\ j} \right\|_{L_2}$$

onde $\left\| - Y \right\|_{L_2}$ indicam a raiz quadrada da soma das diferenças quadradas dos elementos semelhantes dos vetores (matrizes) X e Y em caso geral. A matriz de distância ED foi usada para construir uma árvore baseada no viés do uso de códons.

- $\{GC, GC_1, GC_2, GC_3\}$ Número de ocorrências geral dos nucleotídeos G+C, chamados de conteúdo *GC* onde G+C são estimados nas 3 posições dos códons.

3.3. Análise do uso de códons através da disponibilidade de tRNAs

3.3.1. Modelo de *fitness* traducional

A disponibilidade de tRNA foi calculada a partir do genoma humano, contando o número de genes que codificam cada tRNA e considerando o mecanismo de compartilhamento de tRNAs entre códons sinônimos que terminam em pirimidinas (FRIAS et al., 2013).

A distância relativa para o hospedeiro humano foi calculada usando:

$$D_{host} = 100 \frac{F_{rich,\ host} - F_{rich,virus}}{F_{rich,host}}$$

E o Modelo de *fitness* traducional variando entre 0%-100% dado por:

$$TFI = 100 - D_{host}$$

3.4. Disponibilidade de códons

A partir do cálculo de disponibilidade de tRNA por códons, os códons foram divididos em duas classes: códons mais abundantes, que são traduzidos em média 2,9% de todos os tRNAs, e códons menos abundantes que têm uma média de tradução de 1,14% (FRIAS et al., 2013).

3.4.1. Códons alvo para terapia antiviral baseados na inibição dos tRNAs

A metodologia de *Transfer RNA Inhibition Therapy Index* (TRIT) introduzida por Frias (2013), determina os códons mais indicados para inibição levando em consideração a frequência dos códons no hospedeiro e no genoma viral e frequência de tRNA's. A fórmula original utilizada para este cálculo foi:

$$TRIT_i = \frac{f_{i,virus} - f_{i,host}}{f_{i,host} F_{i,tRNA}}$$

onde $f_{i,host}$ é a frequência do códon *ith* no hospedeiro, $f_{i,host}$ é a frequência do códon *ith* no vírus e $F_{i,tRNA}$ é a frequência da espécie de tRNA cognato para o códon *ith*.

No presente trabalho, foi usada a seguinte fórmula modificada:

$$TRIT_i = max(0, \frac{f_{i,virus} - f_{i,host}}{F_{i,tRNA}}(1 - f_{i,host}))$$

que provou ser menos sensível a códons raros no genoma. A função *max* foi introduzida devido ao interesse deste trabalho em encontrar valores positivos para TRIT.

4. Resultados

Os resultados apresentados nesta seção estão divididos em dois capítulos:

- Artigo I, intitulado: "*Betacoronavírus genome analysis reveals evolution toward specific codons usage: Implications for SARS-CoV-2 mitigation strategies*", foi publicado no periódico Journal of Medical Virology, em 2021. O trabalho apresenta uma análise do uso de códons de todos os Betacoronavírus, com foco em SARS-CoV-2, através da metodologia de *codon usage*. O objetivo deste trabalho foi entender o padrão evolutivo ao longo das espécies virais até alcançar o nível de aprimoramento evolutivo apresentado pelo genoma de SARS-CoV-2. Todos os dados gerados durante este trabalho, incluindo os que não foram usados no artigo publicado, são apresentados no capítulo 4.1.

- Artigo II intitulado: "*SARS-CoV-2 epidemic in Brazil: how variants displacement have driven distinct epidemic waves*", foi publicado na revista Virus Research, em 2022. O presente trabalho descreve o impacto do surgimento das variantes, em especial as VOCs, no Brasil, ao longo da pandemia de Covid-19, e sua relação com o número de casos infectados e número de mortes. Todos os dados gerados durante este trabalho, serão apresentados no capítulo 4.2.

4.1. Análise da composição genômica dos Coronavírus

Neste capítulo foi avaliada a composição genômica dos Coronavírus. Para este objetivo, foi utilizado inicialmente os HCOV, coronavírus com registro de infecção em humanos, e posteriormente todos os Betacoronavirus, família do vírus SARS-CoV-2.

A composição genômica dos HCOV foi inicialmente realizada por pares, dinucleotídeos, e posteriormente por presença dos nucleotídeos nos códons (Tabela 3 e Figura 4), o objetivo desta análise foi buscar por similaridades genéticas que sugerissem a capacidade de infecção à um novo hospedeiro.

Os resultados da análise de dinucleotídeos mostrou TG como o mais frequente, sendo o dinucleotídeo mais usado por 6 dos 7 vírus analisados, somente a variante HCOV-OC43 apresentou GC como o dinucleotídeo mais frequente, porém com TG como segundo mais usado. SARS-CoV-2 apresentou uma grande similaridade com HCOV-SARS, como esperado, tendo como resultado TG como o mais usado e CG como o menos utilizado. Além disso, houve também uma grande similaridade da frequência entre SARS-CoV-2, HCOV-SARS, HCOV-MERS e HCOV-229e, quando comparados às outras variantes, com média de 39% de similaridade, um possível indicativo de relação devido a capacidade de infecção ao hospedeiro humano.

TG	1.38044694	TG	1.41034426	TG	1.321386123	TG	1.339074806	GC	1.316430721	TG	1.254135477	TG	1.425386538
CA	1.274290716	CA	1.307564291	CA	1.201357189	AC	1.306099597	TG	1.293378258	GC	1.168119146	CA	1.341091581
AC	1.236465055	CT	1.199275129	CT	1.161966768	CA	1.282654279	CA	1.247134351	CC	1.165444731	AC	1.271034558
CT	1.18099503	AC	1.173283757	GC	1.16057512	AA	1.168483338	CC	1.149561555	CT	1.13961666	GC	1.185848098
GC	1.084750593	GC	1.163047416	AC	1.107620721	GT	1.122425058	CT	1.062122445	AC	1.105925277	AA	1.154238575
AA	1.069538666	AA	1.037968514	AA	1.055179717	GC	1.110354415	AA	1.04899481	CA	1.097439917	GT	1.09846302
GT	1.055032748	TT	1.013661153	GT	1.02403689	CT	1.10115822	AC	1.042451995	AA	1.088074095	CT	1.091506857
TT	1.043992133	AG	0.9916574114	AG	1.001078798	CC	1.021964097	GT	1.017451842	GT	1.066725714	TT	1.033582097
AG	0.9928093761	GT	0.9805711615	TT	0.9725445342	TT	1.009125067	TT	1.001264975	TT	0.9760298056	CC	0.916461369
GG	0.9509508427	GA	0.9491245142	CC	0.9406948339	GG	0.9199276459	AG	0.989290888	AG	0.9598581075	AG	0.8811822461
GA	0.9202355913	GG	0.9369514115	GG	0.9370496943	AG	0.8775759251	AT	0.9508115605	TA	0.9571696153	GA	0.8715553461
CC	0.8795316344	AT	0.8602117859	TA	0.8986799268	TA	0.8768244568	TA	0.9265870621	AT	0.9243419498	GG	0.8610839224
TA	0.8276219608	TC	0.8470090947	GA	0.8933560331	AT	0.836952472	GG	0.8950007422	GG	0.8987315235	AT	0.8218448559
AT	0.8049958157	CC	0.8215692503	AT	0.8875310398	GA	0.8163005596	GA	0.8854764121	GA	0.8938598531	TA	0.7953466827
TC	0.7977047932	TA	0.7956466907	TC	0.8474918769	TC	0.7312685679	TC	0.7113522147	TC	0.7932882028	TC	0.7124206432
CG	0.3975525735	CG	0.4499441108	CG	0.5511734761	CG	0.4133342447	CG	0.4786100929	CG	0.4490646718	CG	0.4900112407

Tabela 3 - Frequência de dinucleotídeos dos HCOV por sequência. Eles estão organizados da esquerda para direita em: SARS-CoV-2 (vermelho), HCOV-SARS (verde claro), HCOV-MERS (verde escuro), HCOV-NL63 (roxo), HCOV-OC43 (azul claro), HCOV-HKU1 (amarelo) e HCOV-229e (azul escuro). A tabela está organizada de cima para baixo, por ordem de dinucleotídeo mais usado por vírus.

Após o cálculo dos dinucleotídeos, foi feita a análise comparativa especificamente do conteúdo GC dos códons, com o objetivo de inferir sobre as pressões evolutivas refletidas nos genomas virais (FIGURA 4). O resultado da análise do conteúdo GC geral no genoma, mostrou que, SARS-CoV-2 apresentou a maior porcentagem de G ou C na primeira posição geral dos seus códons 0.43, similarmente HCOV-SARS teve a maior porcentagem 0.42, para GC na segunda posição dos códons. Enquanto que, HCOV-MERS obteve a maior porcentagem para a presença de conteúdo GC geral e na última posição dos códons, 0.41 e 0.39 respectivamente.

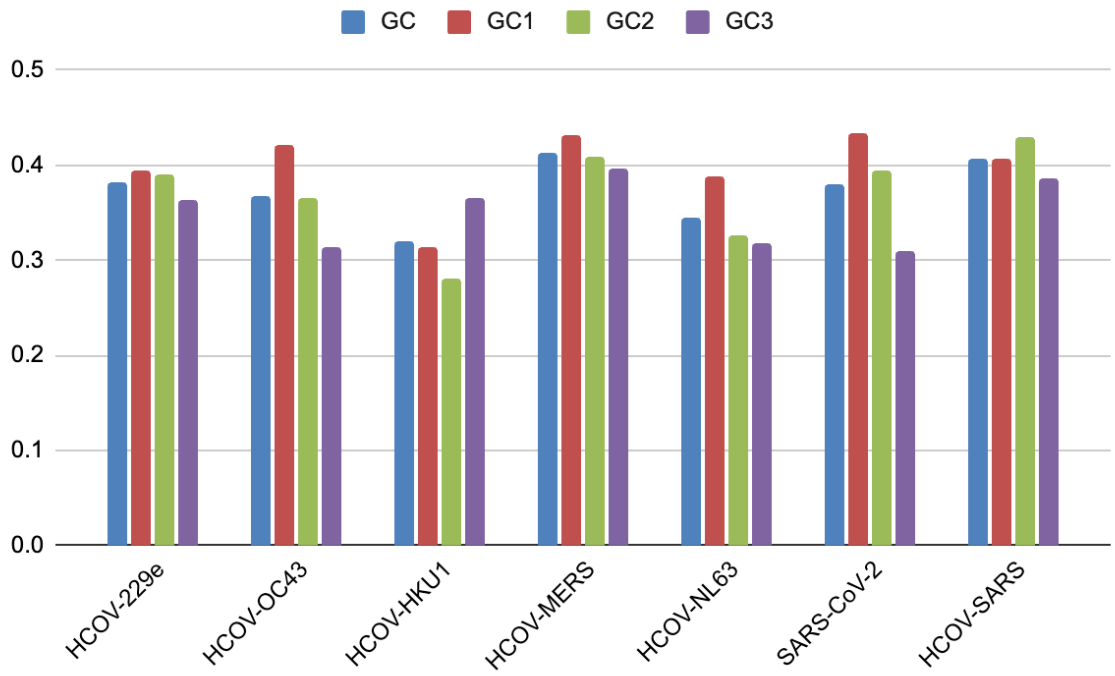

Figura 4 - Conteúdo GC por códons calculados para vírus baseado em sua sequência de referência. GC (Azul) representa a quantidade geral dos nucleotídeos independente da posição no códon. GC1 (vermelho) indica um dos nucleotídeos na posição 1 dos códons. GC2 (verde) um dos nucleotídeos na posição 2 dos códons. GC3 (roxo) um dos nucleotídeos na posição 3 dos códons.

4.1.2. *Relative Synonymous Codon Usage*

A análise do RSCU foi realizada para todos os Betacoronavírus, e seguiu a seguinte ordem, primeiramente foi feita a análise para todos os vírus da família, baseado no resultado foi elaborado uma breve descrição dos códons mais usados por vírus juntamente com um *heatmap* ilustrativo para cada. Os cálculos apresentados foram realizados para as sequências completas e suas respectivas ORF's, a fim de se estabelecer as regiões mais susceptíveis. Em seguida, foi calculado um modelo comparativo do vírus SARS-CoV-2 e hospedeiro humano.

O cálculo do RSCU por hospedeiro, foi realizado através da contagem de Códons obtida no banco de dados, a fórmula do RSCU foi aplicada através de um *script* em R. Os valores disponíveis foram computados e calculados, devido a ausência de outras informações de região, o resultado final apresentado abaixo representa a média geral de uso por organismo.

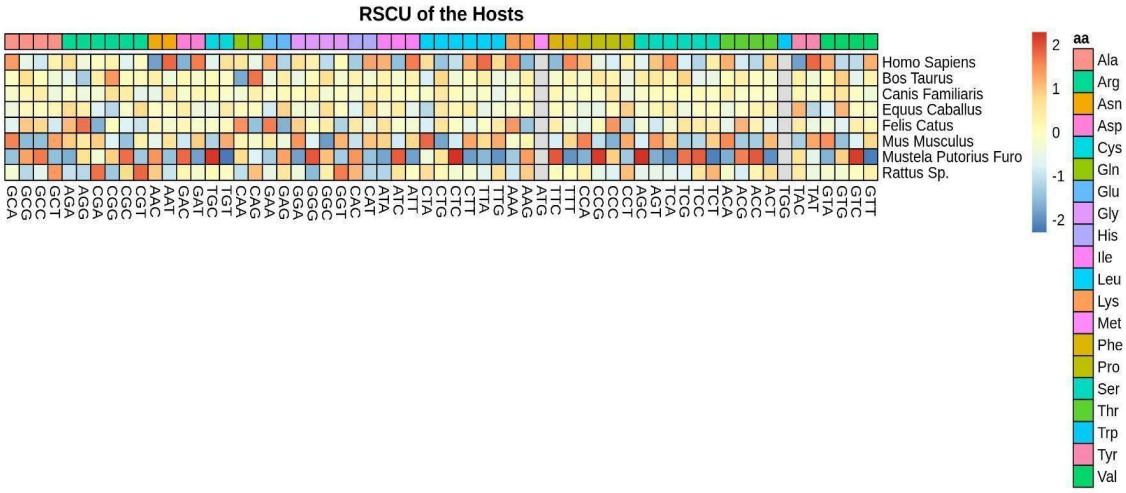

Figura 5 - *Heatmap* dos hospedeiros, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. Os códons com maior valor de RSCU estão com o tom mais forte de vermelho, representando maior seleção para uso por aquele determinado organismo mostrado na coluna lateral. Os aminoácidos e seus códons sinônimos mantêm a mesma cor e são representados na parte superior da figura.

O resultado encontrado, figura 5, mostrou uma relação de similaridade entre os hospedeiros roedores, *Rattus sp*, *Mus Musculus* e *Mustela Putorius Furo*, com a presença de um modelo de seleção bem versátil. *Felis Catus*, apresentou vários pontos azuis, indicando uma

tendência de seleção entre os códons, o que possivelmente indicaria uma maior vulnerabilidade. O *Homo Sapiens* considerado o ser que mais evoluído filogeneticamente do grupo, foi o que apresentou maior tendência de seleção de códons, com a presença fortes cores na seleção de códons específicos para alguns aminoácidos, como AAC(Asn) e TAC(Tyr).

Os valores encontrados para cada codon do hospedeiro foram usados na próxima seção, como modelo comparativo para cada Betacoronavírus analisado. O intuito desta análise foi avaliar a taxa de adaptabilidade viral para cada hospedeiro mamífero estudo, considerando o mimetismo biológico em relação aos códons.

Na figura 6, é mostrado o resultado do RSCU calculado para o vírus *Bat Hp-beta coronavirus*. Os códons que apresentaram uma tendência durante a análise estavam em sua maioria presentes nas ORFs: E, M e ORF7. Foram eles: AAA, GAT, CAA, CCT, CAC, TGC, AGA, GAT, GGA, CAT, CCT, TAT, TTT e GAA. Todos estes códons apresentaram valor equivalente a 1, indicando uma grande disparidade de seleção no processo de transcrição.

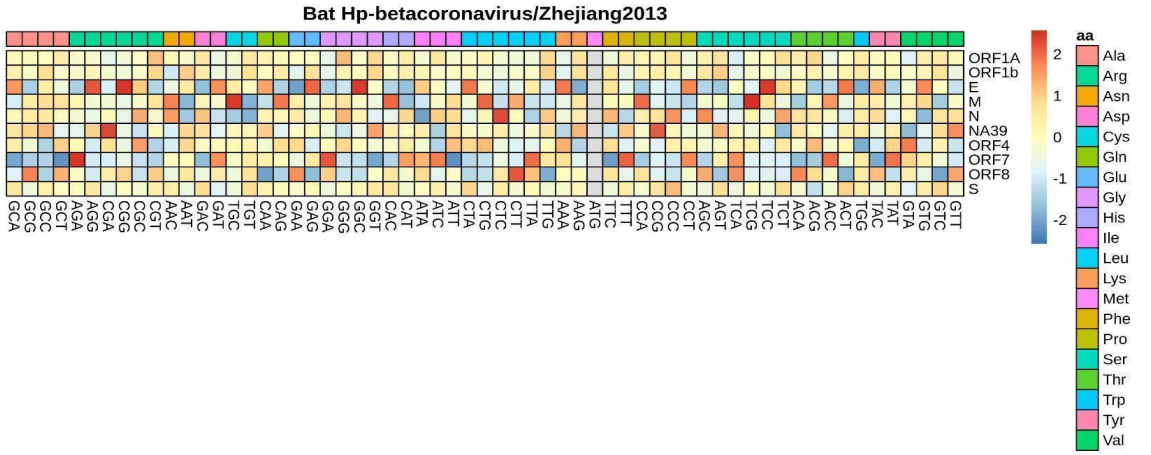

Figura 6 - *Heatmap* dos códons do vírus *Bat Hp-betacoronavirus*, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinônimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

Bovine Coronavirus, figura 7, apresentou uma grande tendência de uso dos códons nas

ORF's: GP05, GP06 e GP07. Eles estão presentes na sua maioria na GP05: AAG, AAT, AGA, ACT, CAT e TCC; GP06: AAC, GAA, GTT, TGT e TTT; GP07: AAA, CAT, CCT e TAT; E: AAA, AAT, GAG e TTT. Sendo em sua maioria códons compostos por A e T. Dois aminoácidos também obtiveram destaque por serem os mais frequentes nos códons de alto valor de RSCU: Asparagina e Glicina.

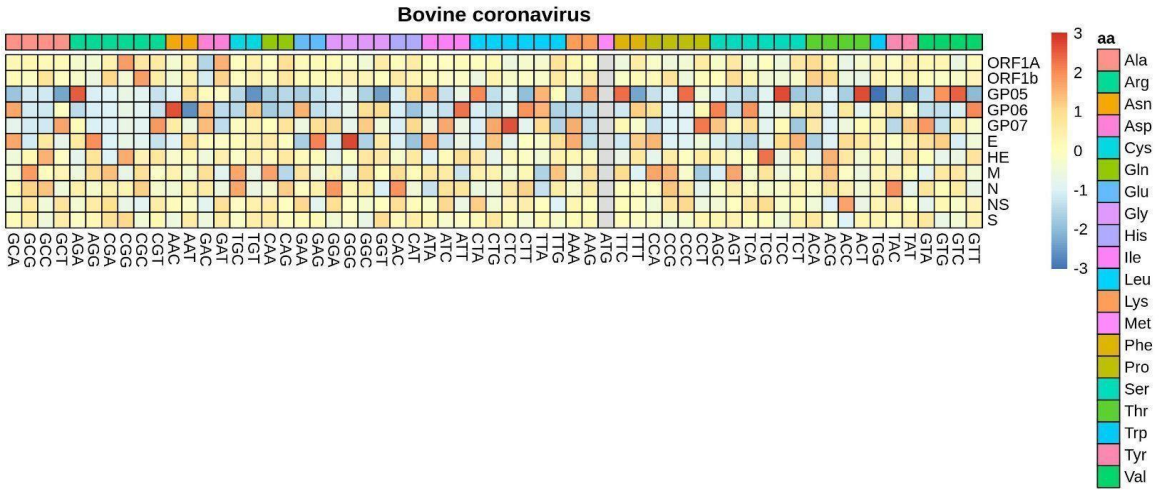

Figura 7 - *Heatmap* dos códons do vírus *Bovine Coronavirus*, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinónimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

O Beta Coronavirus *Erinaceus*, assim como o *Bovine* e *Bat HP*, apresentou um grande número de códons com alto valor de RSCU na ORF E. Os códons com alto valor no geral, por ORF, foram : **ORF E**: GAA, GAT, CAT, CGA e TAT; **ORF M** : GAA e CAT; **ORF N**: TGC; **ORF3A:** AAA, CAT, GAG, GAT, GGA e TGT; **ORF 3B:** GAG, GAT, GGA e TGT; **ORF4A:** ATC, GAG, GAT, GGA, TGT e TTT; **ORF4B:** TTT; **ORF8B:** TAC e TGC. Todos estes códons listados obtiveram valor de RSCU igual a 1, o que indica uma grande tendência de uso em relação a seus codons sinônimos.

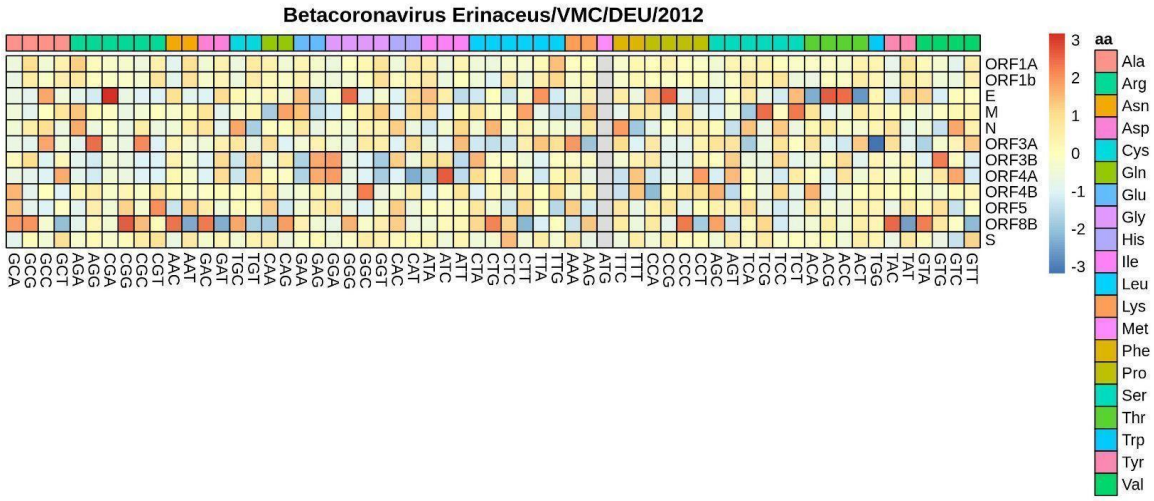

Figura 8 - Heatmap dos códons do Beta Coronavirus *Erinaceus*, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinônimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

O resultado para o HCOV HKU1 mostrou uma grande presença de viés de seleção para genes nas seguintes ORF's, **E:** AAT, AGA, GAA, GAT, GTT e TGT; **M:** TAT e TGT; **N:** TGT; **N2:** TAC; **ORF4:** ATT, CAA, CCT, TGT e TTT (Figura 9).

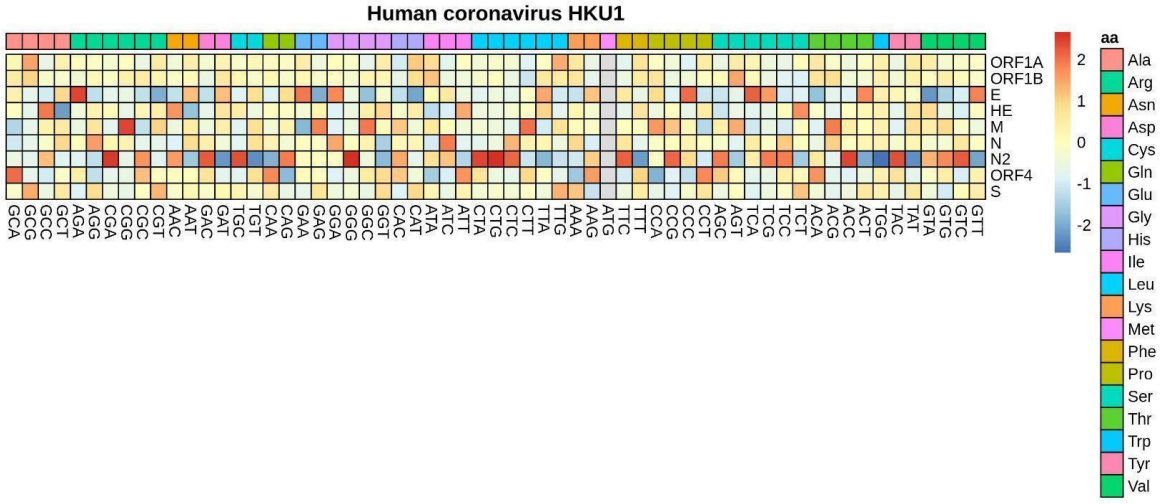

Figura 9 - *Heatmap* dos códons do HCOV HKU1, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinônimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

Rabbit coronavirus HKU14 obteve os seguintes resultados para o cálculo de RSCU, ORF E: AAA, AAT, GAG, GGT e TTT; ORF M: CAA e CAT; ORF NS2A1: AAA, AAT, GAG, GAT, GCT, CAA, CAT, CCT, TAT, TGT, TCA e TTT; NS2A2: AGT e ACT. Diferentemente dos anteriores, HKU14 apresentou a ORF NS2A1 como a mais enriquecida com codons de alto valor de RSCU (Figura 10).

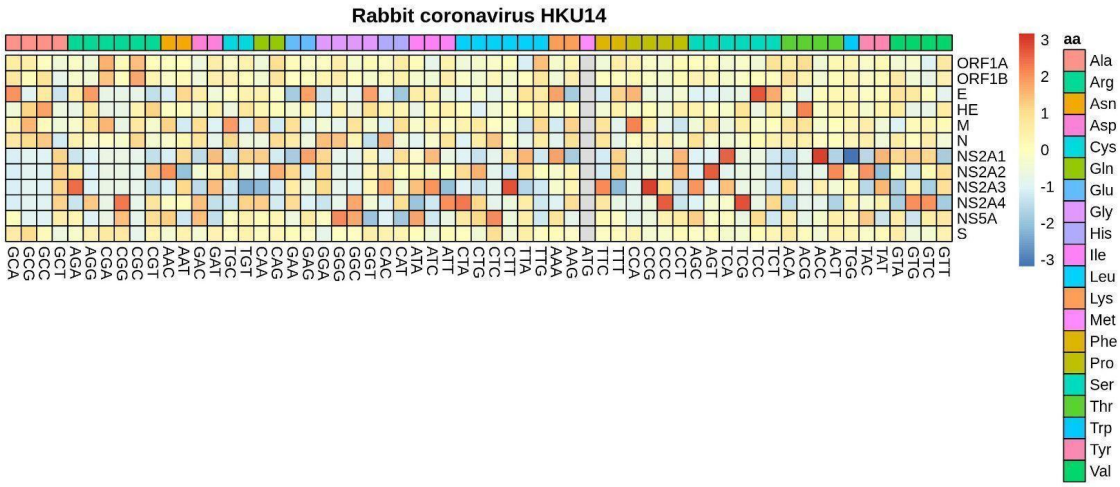

Figura 10 - Heatmap dos códons do *Rabbit Coronavirus HKU14*, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinónimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

O Beta Coronavirus HKU24 foi um dos virus que apresentou a menor quantidade de códons com valores iguais a 1, o que representa uma maior distribuição de uso entre os códons sinônimos. Os códons com os maiores valores de RSCU foram, **ORF E:** AAT, AGA, GAA e TTT; **ORF N2:** CAT; **ORF NS4:** TAT, TGT e TTT; **ORF NS5:** GAG, CAA e CAT (Figura 11).

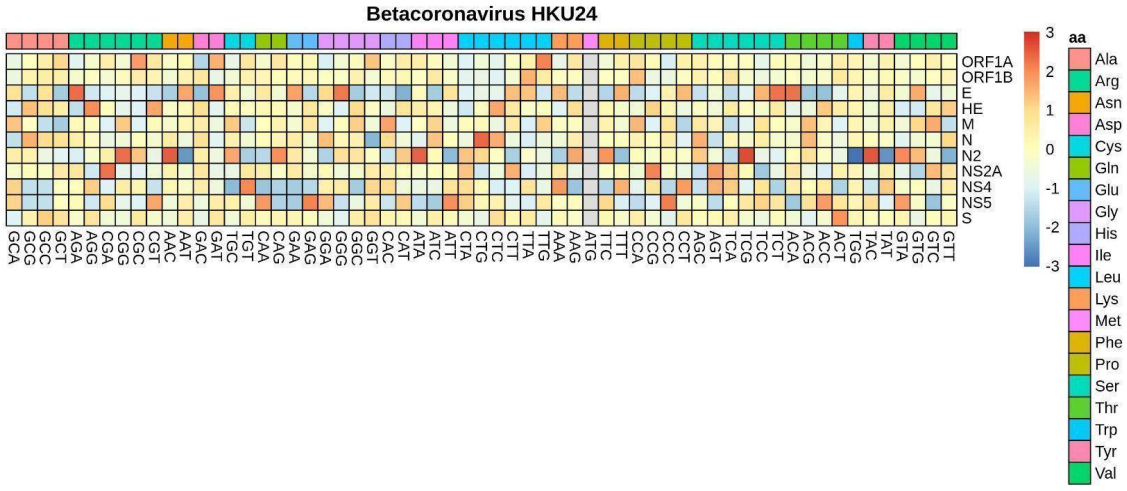

Figura 11- Heatmap dos códons do Betacoronavirus HKU24, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinónimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

O resultado para *Tylonycteris Bat Coronavirus HKU4* foi em sua maioria presente em duas ORF's E e NS3A. Foram eles, **ORF E:** AAG, GAC, CAA, CAT e TCG; **ORF M:** TGT; **ORF N:** TGC; **ORF NS3A:** AAA, AGA, GAT, GCT, CAT, TGT e TTT; **ORF NS3B:** TGT (Figura 12).

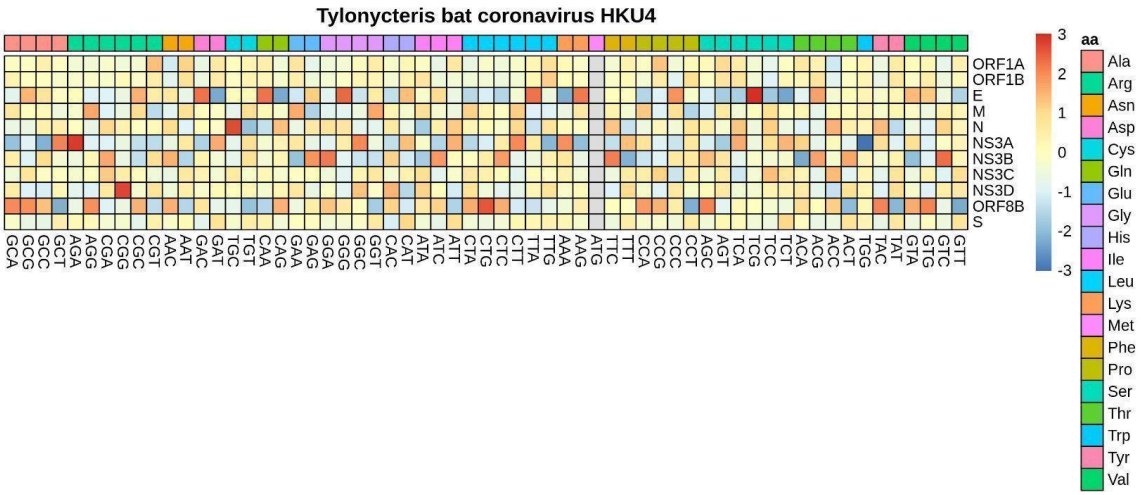

Figura 12 - Heatmap dos códons do *Tylonycteris bat coronavirus HKU4*, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinónimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

Rousettus Bat Coronavirus HKU9 apresentou os seguintes códons com alto valor de RSCU, **ORF E:** AAA, GAA, CAA, CCT e TAT; **ORF M:** CAC e TGT; **ORF NS7B:** AAG (Figura 13).

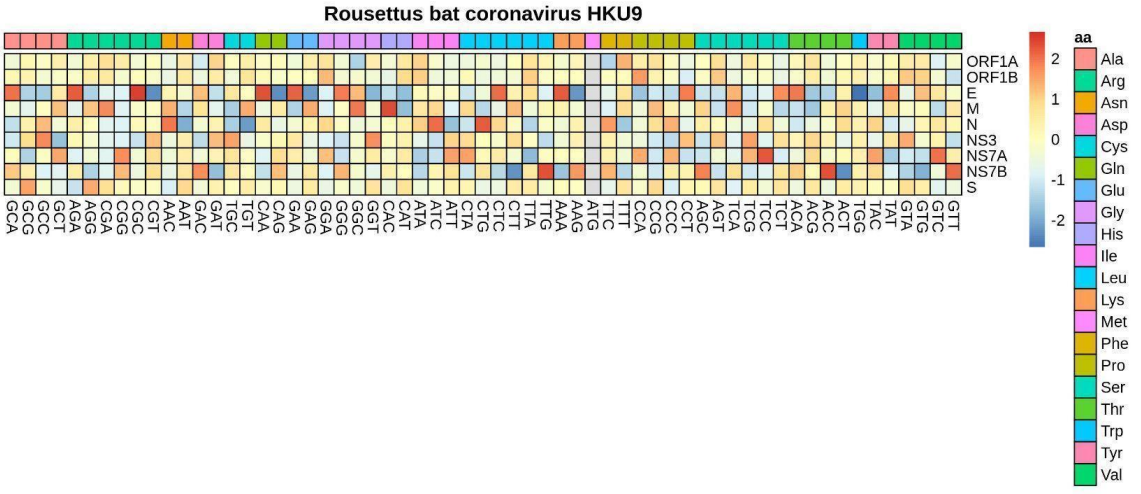

Figura 13 - *Heatmap* dos códons do *Rousettus bat Coronavirus HKU9*, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinónimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

O resultado do RSCU para o *HCOV MERS* foi: **ORF E:** AAA e TGT; **ORF 3:** AAA, AAT, ATT, GGT, CAT e TTT; **ORF4A:** TGT; **ORF5:** GAG; **ORF8B:** GAC (Figura 14).

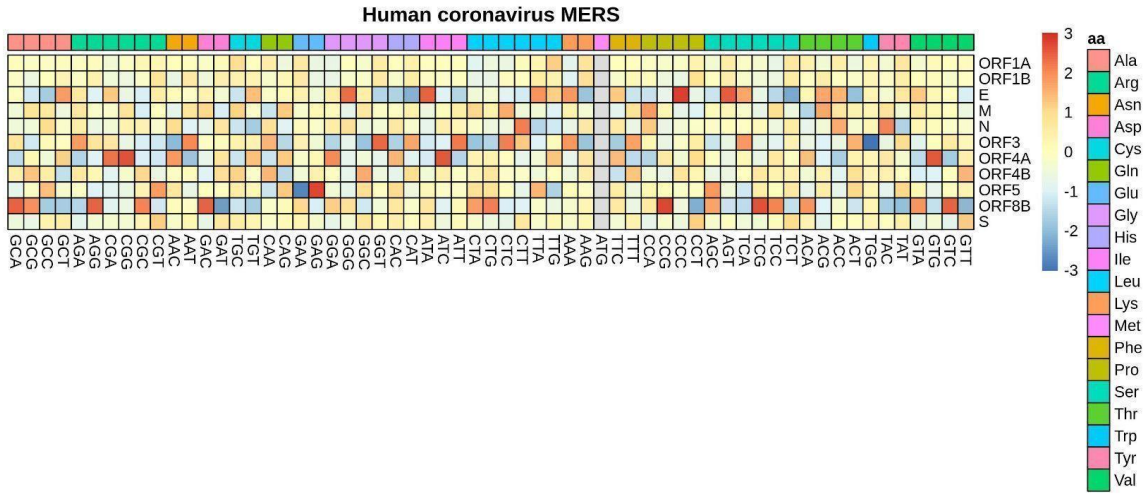

Figura 14 - *Heatmap* dos códons do HCOV MERS, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinónimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

Os codons selecionados como tendenciosos para o vírus *HCOV OC43* foram: **ORF E:** AAA, AAT, GAG e TTT; **ORF HE:** TTT; **ORF M:** CAA; **ORF NS12:** CAT, CCT e TAT (Figura 15).

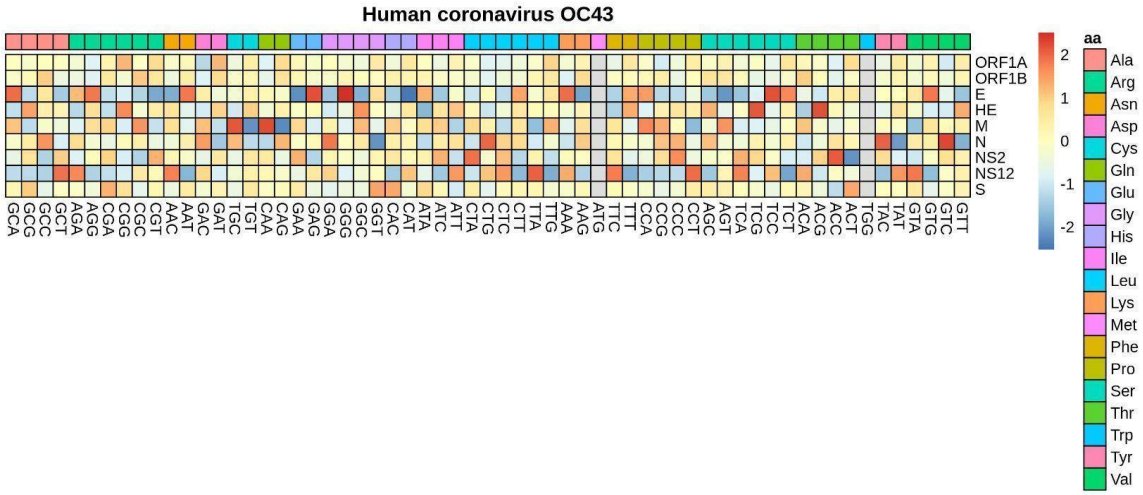

Figura 15 - *Heatmap* dos códons do *HCOV OC43*, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinónimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

HCOV SARS teve um resultado que sugere uma grande variação de uso de códons, figura 16, com presença em várias de suas orfs, indicando uma alta adaptabilidade. Seus códons com alto valor de RSCU foram, **ORF E:** AAA, GAA, GAT e TAC; ORF M: CAC e TGT; **ORF3B:** AAC, GAG, GGC, TAC e TTT; **ORF6:** AAG, AAT, CAT e TAT; ORF7A: AAT e GAC, **ORF7B:** AAA, AAT, GCC, CAG, CCT, TAT e TCA; **ORF8A:** AAT, ACT, GAA, GAT, GCA, CGC e CCT; **ORF8B:** AGC, GAA, CAA e TTT; **ORF9B:** CAT, TAC e TTC.

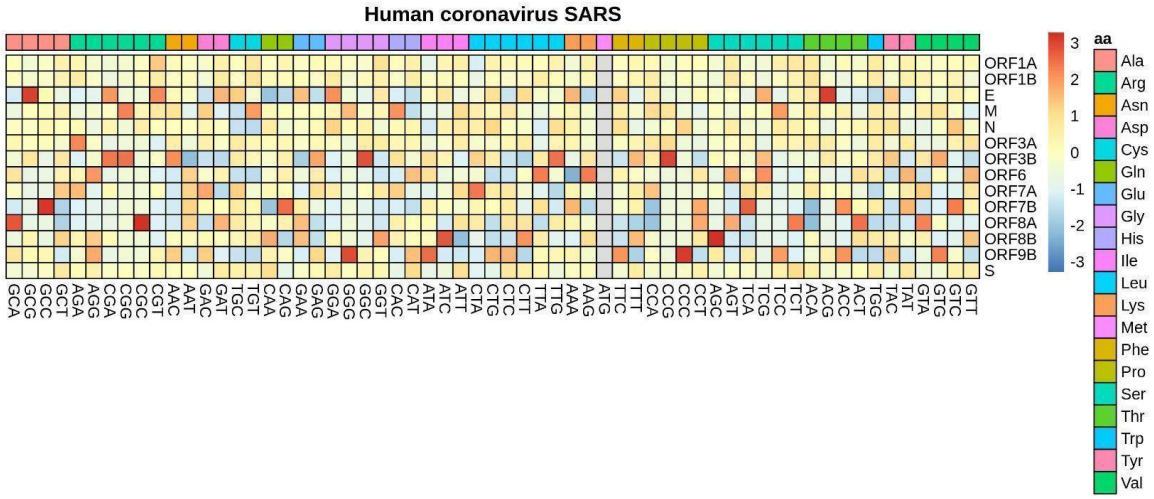

Figura 16 - *Heatmap* dos códons do HCOV SARS, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinónimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

O vírus SARS-CoV-2 foi o vírus que entre os Beta Coronavirus analisados, teve a maior quantidade de códons com valores igual a 1, ou seja com maior tendência de uso (Figura 17). Foram eles, **ORF E:** AAA, GAT, GGT, CCT e TAC; **ORF M:** TGT; **ORF6:** AGG, ACT, GCA, GTT, CAT, CCA e TTT; **ORF7B:** AAT, ACT, GAA, GCC, GTT, CAA, TAT e TCA; **ORF8:** AAA e AAT; **ORF10:** ATA, GAT, GGC, CAA, CCG e TGC. As ORFS com a maior quantidade de códons com valores iguais a 1 foram E, ORF6 e 7B.

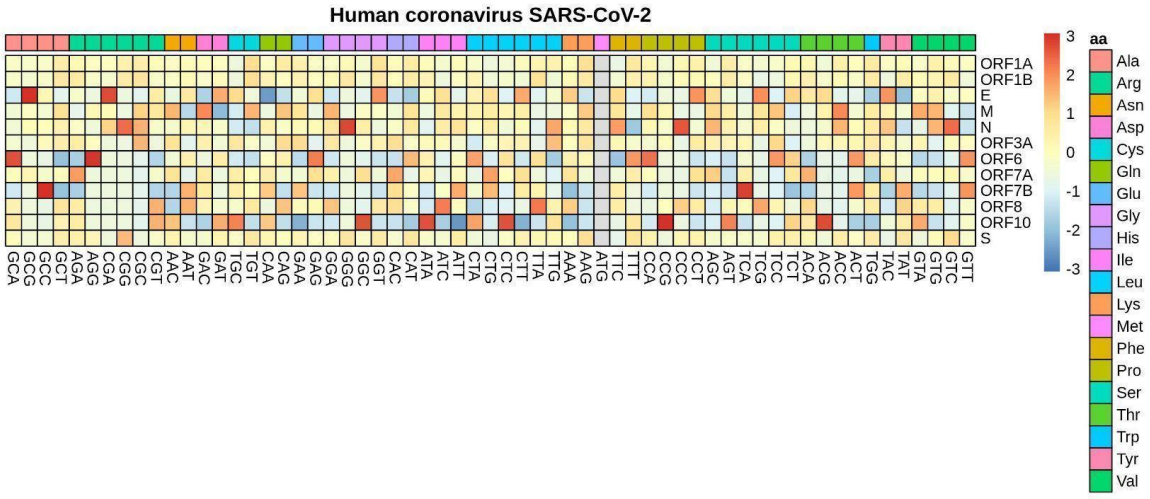

Figura 17 - *Heatmap* dos códons do vírus SARS-CoV-2, a escala apresentada é relativa a variância da média do valor de RSCU encontrado para cada hospedeiro. O códon de maior valor de RSCU entre os sinónimos, está com o tom mais forte de vermelho, assim como de forma oposta ao tom azul, representando a seleção para uso na determinada região do organismo mostrado na coluna lateral. Os aminoácidos e seus códons são representados na parte superior da figura.

4.1.3. Identificação de códons alvos: SARS-CoV-2

Nesta seção foi abordada uma análise focada no vírus SARS-CoV-2 em relação ao hospedeiro humano, o objetivo da presente análise foi hipotetizar os modelos de competição usados pelo vírus para transcrever seus códons.

A análise composicional dos códons com os maiores valores de RSCU revelaram que 5 dos códons mais usados (códons ótimos) pelo SARS-CoV-2 são: TGT (Cys), GAA (Glu), TTT (Phe), CAA (Gln) e AAT(Asn), figura 19, todos ricos de AT (no mínimo 1 nucleotídeo é A ou T dos códons). No genoma humano, os cinco códons mais usados são: CAG (Gln), CAC (His), GAG (Glu), AAG (Lys) e TAC(Tyr), estes sendo, em sua maioria, ricos em GC.

Os valores calculados de RSCU foram plotados como um heatmap, agrupando organismos de acordo com a distância euclidiana entre os valores de observados de RSCU (linha) e colorindo de acordo com os valores normalizados encontrados em cada RSCU, por códon (coluna) (Figura 18).

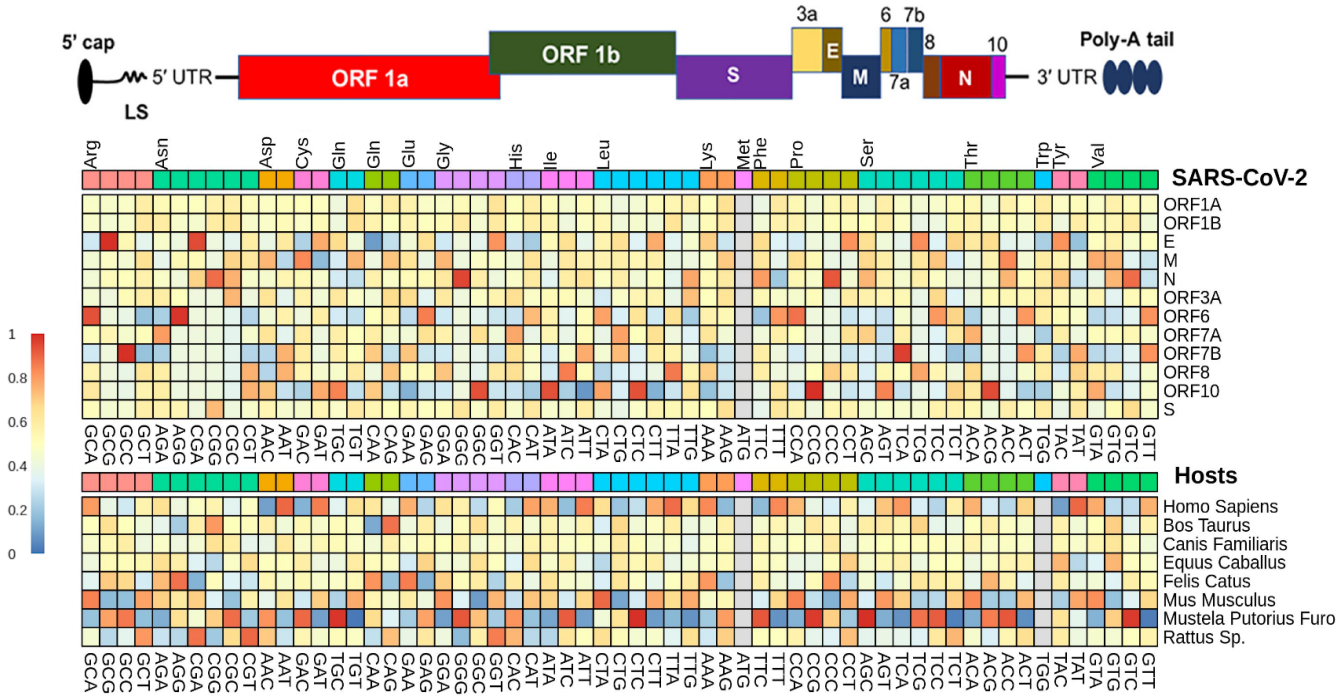

Figura 18 - *Heatmap* dos valores observados de RSCU para SARS-CoV-2 e Hospedeiros, representando os mais usados como vermelho e os menos como azul. Os ramos foram calculados a partir da matriz de distância euclidiana. O tamanho dos ramos é relativo à distância entre o hospedeiro humano e os sete modelos virais de estudo, este cálculo foi realizado sem viés filogenético.

Após comparar os valores RSCU do vírus SARS-CoV-2 com os hospedeiros, foi feita a busca por códons alvos, que aparentam ser mais usados pelo vírus e menos pelo *Homo Sapiens*, de forma a elucidar o mecanismo adaptativo. Dentre estes, 7 foram encontrados como de maior tendência de uso viral, localizados em três diferentes ORFs: CCG, ACG, CTC localizados na ORF10, GGC e TAC na ORF E e GAC na ORF M.

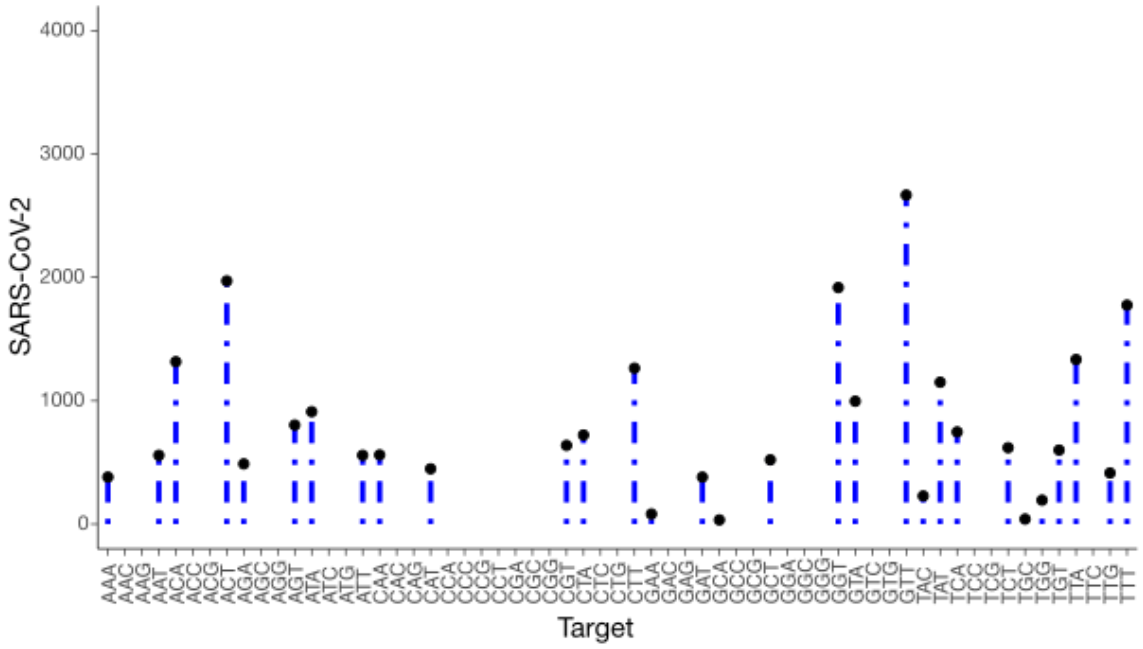

Figura 19 - *Dotplot* dos Códons Ótimos do vírus SARS-CoV-2.

4.1.4. Codon usage em relação a disponibilidade de tRNA do hospedeiro

A metodologia de RSCU leva em consideração a co-evolução do *codon usage* e tRNA *pool* disponível, assumindo que, códons sinônimos são guiados pela disponibilidade de tRNA, enquanto que, o número de cópias de tRNA no genoma é dirigido pela demanda biológica, isso define a frequência relativa de códons sinônimos. Esta teoria foi utilizada nesta seção como base para a relação entre RSCU e tRNA.

O RSCU foi desenvolvido sem levar em consideração o tamanho e a composição do tRNA *pool* disponível para a tradução na célula. Dentro do contexto traducional tem 5 tipos de recursos biológicos que um hospedeiro viral deve competir: ribossomo, fatores de iniciação, fatores de tradução, fatores de liberação e tRNAs maduros.

Considerando a estratégia viral para acessar os recursos de tradução na célula hospedeira, existem duas possíveis opções de disputa:

(1) Competir pelos os recursos mais abundantes, mas que têm grande demanda;
(2) Disputar recursos menos abundantes, porém com menor demanda.

Em ambos cenários há um tipo de competição. Para hipotetizar como o vírus seleciona

entre as duas alternativas acima, é necessário mapear a disponibilidade de tRNA para cada códon, em cada tipo de hospedeiro. No entanto, esta contagem ainda não está disponível publicamente. Por esse motivo, foi utilizado a estratégia de contagem do número de genes que transcrevem cada tipo de tRNA e levando em consideração o mecanismo de compartilhamento de tRNAs entre os códons sinônimos que terminam em pirimidinas. Na figura 20, é mostrado o número de tRNA que traduzem cada um dos 61 códons (desconsiderando códon de parada) no genoma humano.

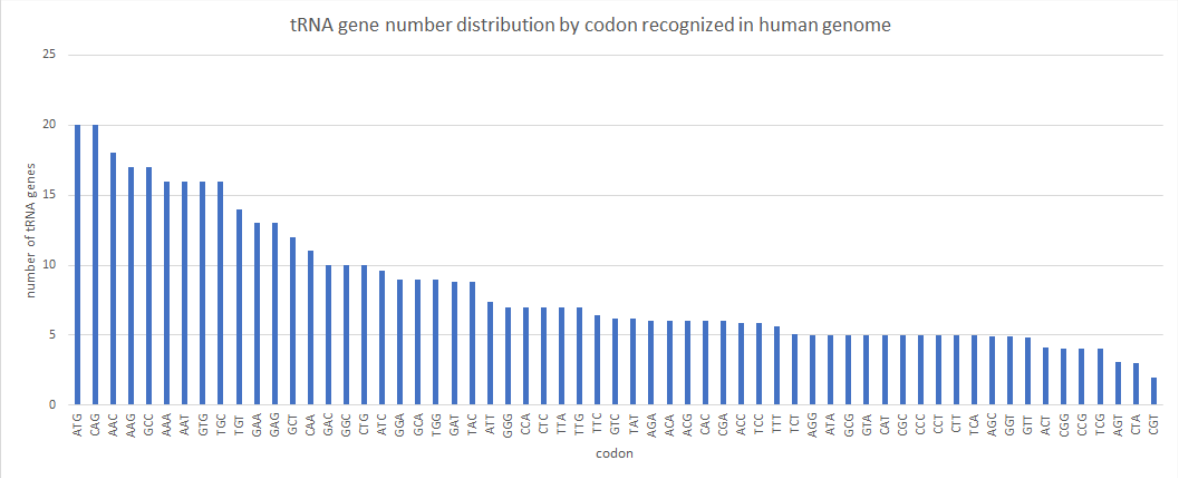

Figura 20 - Contagem de genes codificantes para tRNAs associados aos respectivos códons no genoma humano. Os códons foram ordenados no mais abundante ao menos.

Para visualizar qual tRNA apresenta maior disponibilidade, foi calculado a frequência relativa de genes por tRNA's, figura 21.

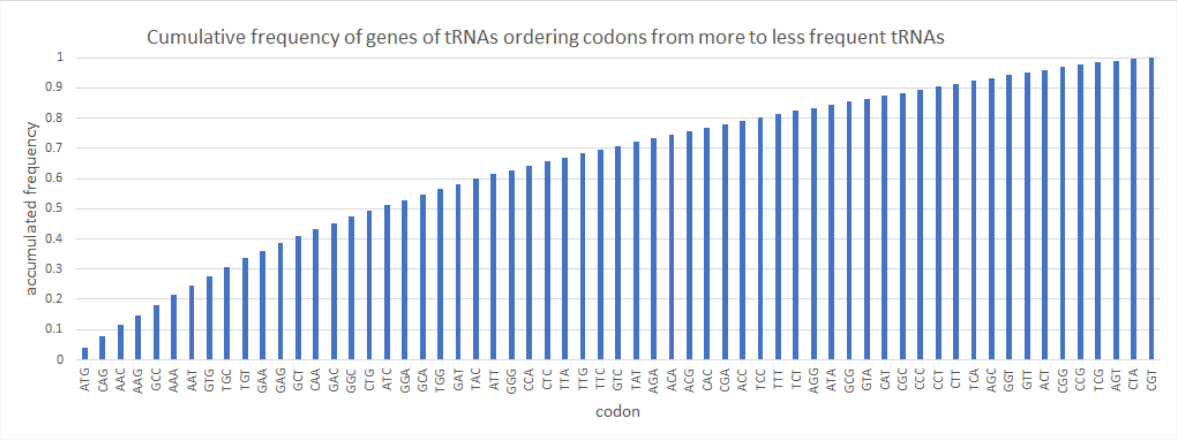

Figura 21 - Frequência acumulada de códons por tRNA**.** Os codons foram ordenados do mais frequente para o menos, da direita para a esquerda.

Foram encontrados 26% dos códons (16/61) com de grande disponibilidade de recurso para o hospedeiro humano, classificados como códons mais ou menos abundantes, isso indicando uma taxa relativa maior que 1. Esse valor representa a disponibilidade de recursos de tRNA para a tradução de acordo com os dados calculados por hospedeiro(Tabela 4).

Códon	**Frequência do tRNA**
TGC	6,87%
TGT	6,37%
AAG	4,66%
AAA	3,19%
TTC	3,15%
GAG	2,70%
GCA	2,70%
TTT	2,49%
GTG	2,45%
CAG	2,45%
CTG	2,45%

ATG	2,21%
GCG	2,21%
GAC	2,17%
GGC	2,07%
GAA	1,96%

Tabela 4 - SARS-CoV-2 códons mais abundantes**.** Os códons por tRNA foram ordenados de cima para baixo, do mais frequente ao menos.

Baseado nos resultados encontrados, assume-se que o uso dos códons pelo vírus em estudo foi ajustado à abundância de tRNA no hospedeiro humano, devido a classe de tRNA abundantes terem alcançado cerca de 50% do total analisado. No entanto, para a encontrar o valor final é preciso análise comparativa entre vírus e hospedeiro. Com esse objetivo, foi desenvolvido o cálculo de TFI apresentado, das sequências de todos os HCOV e o genoma humano (Tabela 5).

Métrica	Espécies Virais							
	Human	SARS-CoV-2	229E	SARS	OC43	MERS	NL63	HKU1
F_{rich} (%)	41.54%	33.5%	32.9%	33.2%	31.7%	31.2%	30.1%	29.1%
D_{ideal} (%)	8.46%	16.5%	17%	16.8%	18.2%	18.7%	19.8%	20.8%
D_{host} (%)	0.0%	19.35%	20.7%	20.7%	23.4%	24.8%	27.4%	29.8%
TFI (%)	100.0%	80.65%	79.2%	70.9%	76.5%	75.1%	72.6%	70.1%

Tabela 5 - Resultado de TFI calculados. Os dados são referentes ao modelo traducional dos sete Coronavírus estudados e o hospedeiro humano. Alguns nomes virais foram reduzidos para melhor enquadramento na tabela, foram eles: 229e (HCOV 229e), SARS (HCOV SARS), OC43 (HCOV OC43), MERS (HCOV MERS), NL63 (HCOV NL63) e HKU1 (HCOV HKU1).

Os resultados relativos ao modelo tradicional dos sete Coronavírus estudados e o hospedeiro humano indicaram o vírus SARS-CoV-2 como o vírus com maior índice de adaptação ao hospedeiro humano, baseado na seleção de recursos traducionais.

4.1.5. Índice fitness traducional (TFI)

Foi feita uma busca pelos tRNA's dos hospedeiros em análise, correspondentes aos códons dos Beta Coronavirus com maiores valores de RSCU. Através desta análise de correlação foi possível encontrar 4 códons que são tRNA abundante e 3 que tem menor disponibilidade de tRNA. Para sumarizar estes resultados foi calculado o índice TFI, como um indicativo para relação de adaptação SARS-CoV-2 e hospedeiros, o resultado é apresentado na tabela 6.

Hospedeiros	**SARS-COV-2 TFI**
Equus Caballus	80.24%
Bos taurus	77.00%
Canis Familiaris	76.83%
Felis Catus	74.71%
Mus musculus	74.71%
Homo Sapiens	74.40%
Mustela putorius furo	73.87%
Rattus sp.	73.00%

Tabela 6 - Índice de fitness traducional. Os valores foram calculados baseados nas informações de todos os hospedeiros disponíveis.

Nossos resultados apontaram um alto TFI de SARS-CoV-2 para todos hospedeiros testados, com média de 70%, sugerindo que outros mamíferos poderiam ser ideais ambientes de desenvolvimento de uma infecção viral.

4.1.6. Coronavírus: terapia de inibição por tRNA (TRIT)

O índice TRIT foi calculado para o vírus SARS-CoV-2. Os melhores três alvos terapêuticos em relação aos valores tRNAs disponíveis, em ordem decrescente são: GTT (Val), ACT (Gly) and GGT (Trp) (Figura 22). Estes três TRITs têm menor abundância de tRNAs cognatos no hospedeiro, este achado representa os tRNAs que foram encontrados poucas cópias de genes no genoma do hospedeiro, para códons sinônimos, sendo assim, alvos candidatos com maior eficácia contra a infecção por SARS-CoV-2. Dois destes três códons, codificam os aminoácidos Valina e Glicina, já identificados como abundantes em infecções por SARS-CoV-2 (WAN et al., 2020). O terceiro códon codifica para Treonina um aminoácido essencial como Valina.

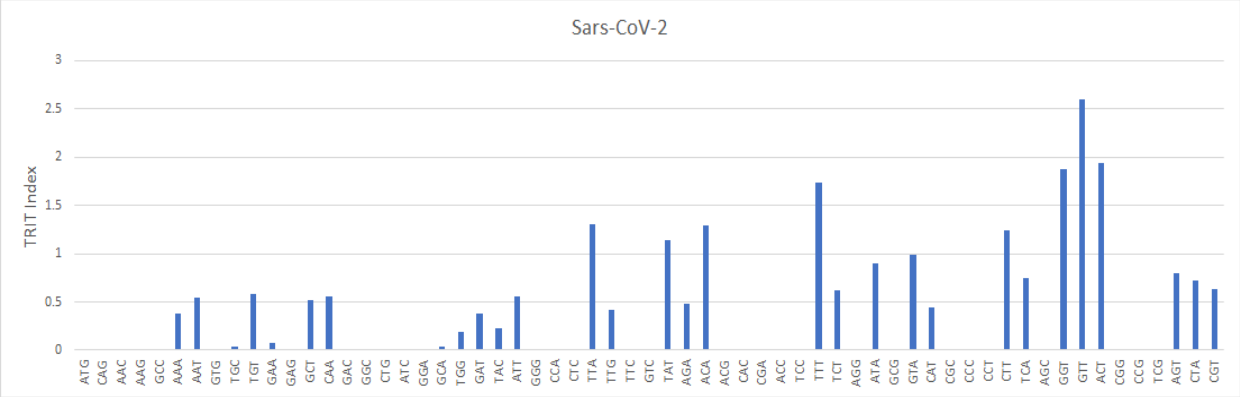

Figura 22. Códons (30) com valor positivo de TRIT para o vírus SARS-CoV-2.

4.2. As variantes de SARS-CoV-2 em circulação no Brasil

Este capítulo traz o impacto das variantes em circulação no Brasil, dentre as várias detectadas até a data do presente trabalho, foram selecionadas apenas as variantes que tiveram o maior impacto na saúde pública, gerando uma nova ou contribuindo de forma consistente, para as 3 ondas de infecções estabelecidas no país. Como consequência dos esforços para gerar dados genômicos que contribuíram para o controle da pandemia, no fim de Fevereiro de 2022, mais de

240 tipos de linhagens de SARS-CoV-2 tinham sido identificadas no Brasil. 11 destas foram classificadas como: VOC's (*Alpha*, *Gamma*, *Beta*, *Delta* e *Omicron*); VOI's (*Lambda* e *Mu*); VUM's (B.1.1.318) e FMV's (*Zeta*, *Eta* e B.1.1.519). As relações ancestrais entre estas linhagens e as principais variantes identificadas no Brasil estão representadas no diagrama abaixo(Figura 23).

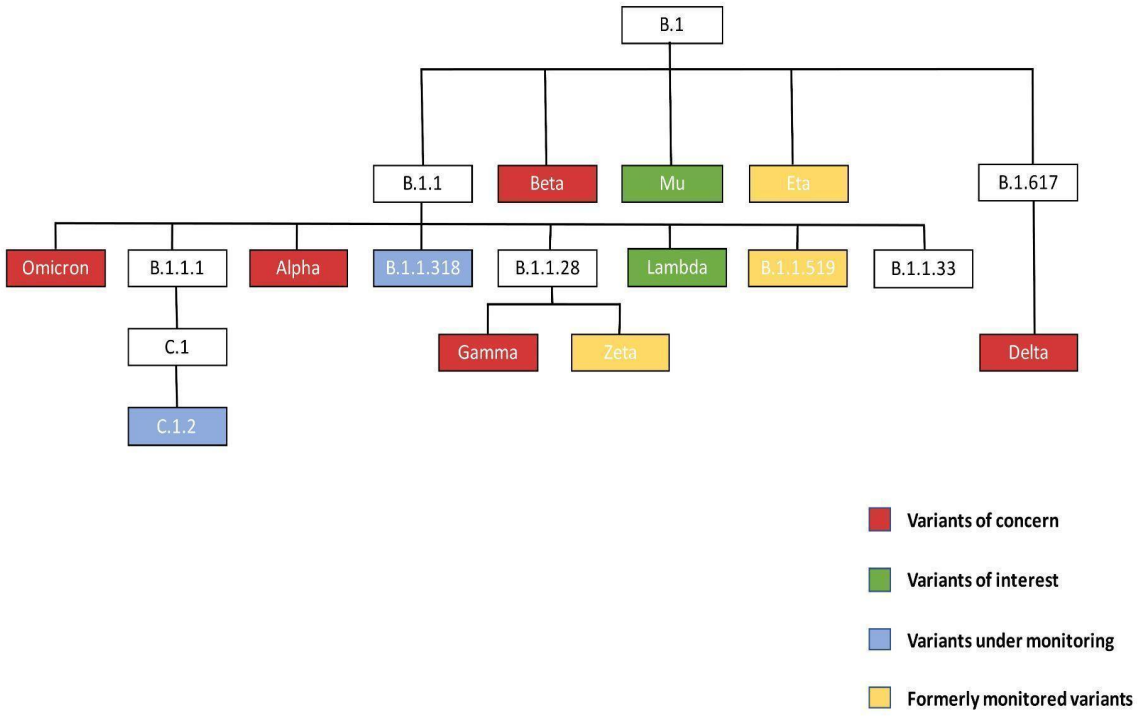

Figura 23 - Diagrama das variantes detectadas no Brasil. O diagrama representa a relação evolucionária entre doze variantes VOCs, VOIs, VUMs e FMVs, detectadas no Brasil desde o início da pandemia até 19 de Fevereiro de 2022. As cores indicam a classificação da variante de acordo com a Organização Mundial da Saúde (OMS, 2022). A B.1 a principal linhagem ancestral. B.1.1.28 teve um crucial impacto, gerando as variantes que apresentaram em seguida alta frequência entre os genomas virais detectados no Brasil (*Gamma* e *Zeta*).

4.2.2. O cenário pandêmico brasileiro após Introdução das VOC's

COVID-19 no Brasil foi caracterizada por 3 diferentes ondas epidêmicas, que causaram mais de 27 milhões de casos com 670 mil mortes até 19 de Fevereiro de 2022 (Figura 24). A primeira onda foi de Fevereiro de 2020 até Novembro 2020, e foi caracterizada pela co-circulação de diferentes linhagens causadas por múltiplos eventos de introdução que

ocorreram através do tempo. A segunda onda foi entre Dezembro de 2020 até Dezembro de 2021, foi dirigida pela circulação de várias VUM's, como a P.2 (*Zeta*), e VOC's como Gama (P.1), que começou a ser detectada em Janeiro de 2021. Ainda, foram detectados esporadicamente várias VOCs, VOIs, VUMs e FMVs adicionais durante esta onda, incluindo *Alpha*, *MU* e *Eta,* C.1.2, B.1.1.318, *Lambda* e B.1.1.519, elas foram registradas com baixa frequência e pouco foi atribuído ao ressurgimento de casos deste período.

Em Abril de 2021, a variante *Delta* começou a ser detectada no país, esta VOC foi a sucessora da *Gamma*, se tornando assim a variante dominante em circulação no cenário nacional, em Outubro de 2021 (igura 25A).

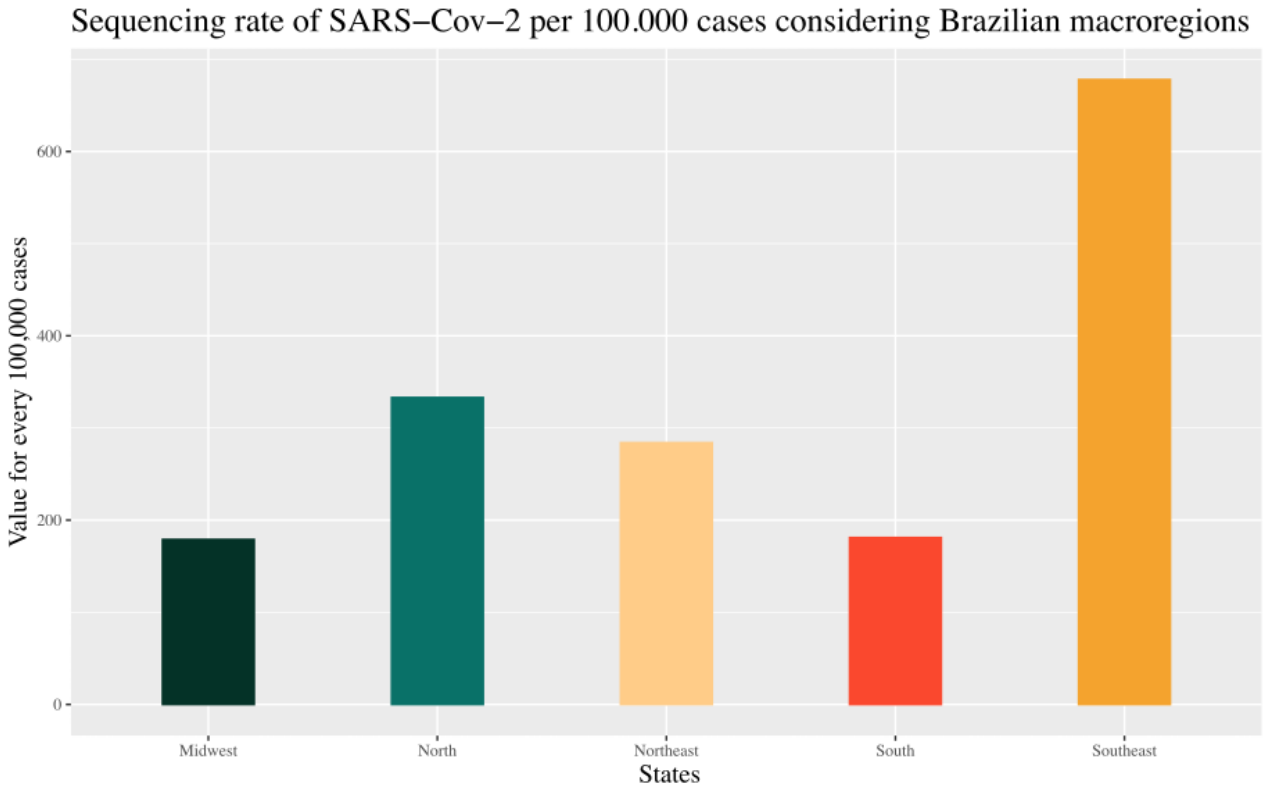

Figura 24 - Número de sequências de SARS-CoV-2 por cada macrorregião brasileira. Os dados foram calculados na escala de sequências por 100 mil casos. Cada cor indicando uma das 5 macrorregiões do Brasil.

Os meses que sucederam a detecção desta VOC emergente no Brasil, entre Setembro e Dezembro de 2021, foram marcados por baixo nível de incidência de casos de infecções e mortes reportadas.

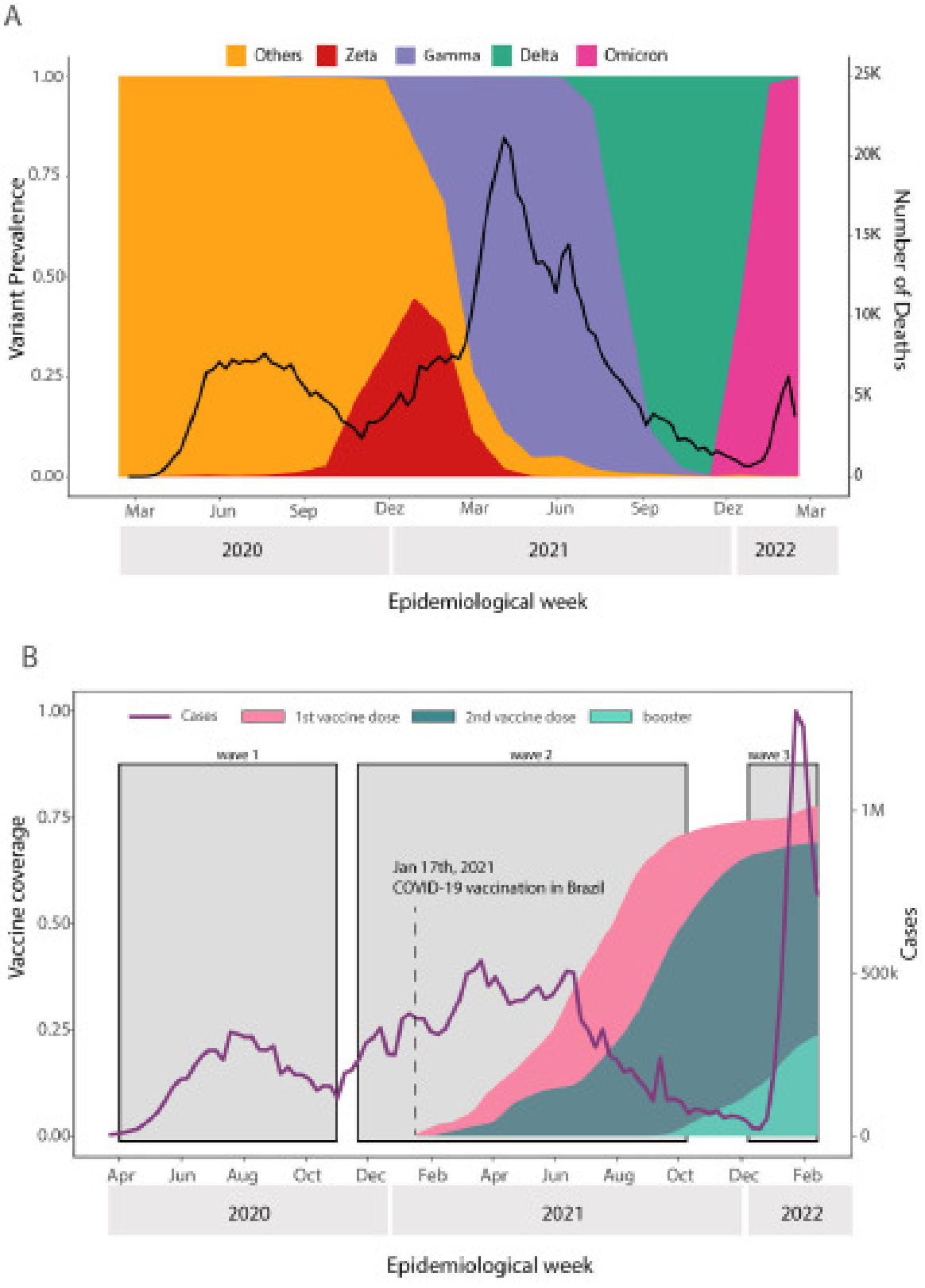
A
Others
Zeta
Gamma
Delta
Omicron
Variant Prevalence
Number of Deaths
1.00
0.75
0.50
0.25
0.00
25K
20K
15K
10K
5K
0
Mar
Jun
Sep
Dez
Mar
Jun
Sep
Dez
Mar
2020
2021
2022
Epidemiological week
B
Cases
1st vaccine dose
2nd vaccine dose
booster
wave 1
wave 2
wave 3
Jan 17th, 2021
COVID-19 vaccination in Brazil
Vaccine coverage
Cases
1M
500k
0
Apr
Jun
Aug
Oct
Dec
Feb
Apr
Jun
Aug
Oct
Dec
Feb
2020
2021
2022
Epidemiological week

Figura 25. Dinâmica da epidemia de SARS-CoV-2 no Brasil. **A**) Número diário de casos de COVID-19 e taxa de vacinação ao longo do tempo monitorado; **B**) Progressão da proporção de variantes no Brasil ao longo da primeira, segunda e terceira onda de infecções, mostrando a rápida sucessão de VOCs através do tempo, adicionalmente mostrando o número de mortes diárias causadas.

Após a descoberta da variante *Omicron* em Botswana, na África do Sul, no fim de Novembro de 2021, o primeiro caso de Omicron foi também detectado no Brasil. Desde esse ocorrido, o número de casos teve um grande crescimento e prevalência entre a taxa de casos detectados no país, até o início de Janeiro de 2022. Foi observado que, houve uma clara tendência de substituição entre a *Delta* pela *Omicron,* em nível nacional e internacional de infecções causadas por SARS-CoV-2.

Um mês após a detecção da *Omicron* em solo nacional, esta variante emergente já foi associada com uma sólida transmissão entre todas as regiões brasileiras e também associada com o aumento inicial de infectados e mortos COVID-19 (Figura 25B). Apesar disso, foi observado que a onda de infecção causada pela *Omicron* teve um rápido pico seguido por declínio, provavelmente refletindo o sucesso no avanço no programa de vacinação nacional.

5. Discussão

A COVID-19 é uma doença emergente que despertou a atenção mundial pelo impacto catastrófico causado pelo SARS-CoV-2. De 2019 até abril de 2022, a COVID-19 alcançou o número de mais de seis milhões de mortes no mundo, com o Brasil ocupando a posição de terceiro país com maior número de casos (OMS, 2022; MAGALHAES et al., 2020; OUDE MUNNINK et al., 2021).

Estudos prévios buscaram identificar hospedeiros para SARS-CoV-2. Em destaque, um deles detectou antecipadamente o furão (Mink) como um hospedeiro secundário para SARS-CoV-2 que, posteriormente, causou uma epidemia em humanos (Oreshkova 2020). Esse achado levantou várias questões sobre a importância da vigilância genômica em relação aos hospedeiros e como a maioria dos hospedeiros secundários ainda resta desconhecida. É fato que, mesmo reduzindo o universo de possíveis hospedeiros para somente mamíferos, ainda assim seria economicamente inviável a vigilância em potenciais hospedeiros (OUDE MUNNINK et al., 2021). Como um possível meio de redução deste universo amostral, propôs-se a análise de *codon usage* somada à disponibilidade de tRNA como um indicativo de possíveis novos hospedeiros.

Neste trabalho, propôs-se a utilização da bioinformática associada à vigilância genômica, e que usando dados públicos, seja um meio de avaliar o processo infeccioso causado pelo SARS-CoV-2, além de identificar possíveis novos hospedeiros para esse vírus. Para isso, foram analisadas todas as sequências referências de Beta Coronavírus e correlacionadas com os códons dos hospedeiros, e posteriormente a disponibilidade de tRNAs. O intuito dessa análise foi comparar as sequências e a adaptabilidade do vírus a diferentes hospedeiros, e isso de forma a elucidar sua aquisição adaptativa ao hospedeiro humano e a regiões do genoma sob maior pressão tendenciosa, por meio de análise de cada códon, aminoácido, ORF e sequência.

Os valores de RSCU calculados, combinados com a disponibilidade de tRNA, apontaram para seis códons e três ORFs como preferidas na seleção do vírus SARS-CoV-2 em relação ao hospedeiro humano. ORFs: CCG, ACG, CTC localizado na ORF10; GGC, TAC localizado na ORF E; e GAC na ORF M. Com estes resultados foi possível notar que, apesar da maioria dos trabalhos, apenas, apresentar foco na região *Spike*, outras ORFs têm possivelmente contribuído tanto quanto para a adaptação do SARS-CoV-2. Quando foram analisados todos os Beta Coronavírus, foi notado um alto nível de tendência de uso de códons da ORF E, apresentando na

maioria dos Beta Coronavirus a maior quantidade de códons com os valores de RSCU iguais a 1. Este resultado sugere que as regiões implicadas vinham passando por pressão seletiva ao longo dos anos, desse modo, evidenciando um processo evolutivo que começara muito antes, e que resultou no TFI atingido por SARS-CoV-2.

O resultado da análise de TFI para oito hospedeiros diferentes revelou um nível de adaptação muito alto, acima de 70% do vírus SARS-CoV-2 (tabela 6). Isso evidenciou um grande grupo de hospedeiros secundários possíveis, o que, conforme discutido previamente neste trabalho, é um reservatório necessário para mutações que levam à adaptação em uma nova espécie.

De 2019 até 2022 foram identificadas mais de 240 linhagens do vírus SARS-CoV-2 no Brasil (WHO, 2022). As variantes que apresentaram maior risco devido às suas características, VOCs, ficaram sob monitoramento internacional e alerta das autoridades de saúde. O Brasil teve um total de três ondas de transmissão bem estabelecidas e mais de 27 milhões de casos (WHO, 2022). Apesar dos esforços para que o conhecimento da introdução de uma nova variante acontecesse quase em tempo real, o impacto dela ainda foi sofrido nacionalmente, causado pela disseminação em massa e a capacidade cada vez mais desenvolvida de adaptação das variantes.

Ao observar a introdução de uma nova variante em um ambiente, foi possível verificar: uma substituição da variante prévia, evidenciando um mecanismo de seleção natural entre os vírus, e alteração do perfil de suscetibilidade imunológica da população devido à evasão do patógeno à resposta imune do hospedeiro. Outro interessante aspecto desta seleção é a viremia, conforme citado por Dennehy (2017), em que é mais vantajoso ao parasita que o hospedeiro sobreviva para que ele consiga completar seu ciclo, como exemplo a Ômicron, a VOC mais recente conhecida, cuja taxa de mortalidade é inferior à Delta, sua antecessora.

No presente trabalho foram combinados dados epidemiológicos e genéticos com uma abordagem bioinformática aplicada a dados públicos. A análise de códons teve resultados promissores sobre evolução da adaptação viral, porém limitada à espécie, tendo como necessidade a complementaridade de um novo trabalho que possa analisar também por variantes. Os dados de tRNAs para hospedeiros foram também limitados somente aos disponíveis em um banco de dados público (http://gtrnadb.ucsc.edu/).

A vigilância genômica se mostrou uma poderosa ferramenta informativa sobre vírus emergentes, utilizada no mundo todo. Porém os resultados apontaram para uma necessidade de

implementação de um maior número de sequenciamentos por região, de forma que amplie a cobertura pelo país, devido a disparidade de sequências em relação ao tamanho de pais, subnotificando o número de infectados e a circulação de variantes emergentes.

O modelo de estudo desta tese, SARS-CoV-2 durante a pandemia de 2019 a 2022, trouxe luz à possibilidade de aplicação da bioinformática combinada a vigilância genômica. Contribuindo assim, para o risco nacional dos arbovírus e a susceptibilidade geográfica brasileira, além de monitoramento de possíveis novos vírus emergentes e reemergentes em todo o mundo.

6. Considerações Finais

O presente trabalho busca contribuir para que possíveis novos cenários de emergência viral possam ser mais bem controlados. Dessa forma, a primeira etapa do trabalho expõe uma investigação de genomas virais por meio da bioinformática, para tanto, foram usadas, apenas, sequências públicas de referência, tendo como resultado evidências sobre adaptação e comportamento viral.

Na segunda etapa deste trabalho, foi feita a correlação do efeito da circulação de variantes virais e como isso impacta em todo um país, apenas usando dados públicos e um computador de uso pessoal. Ambos evidenciaram que o poder da bioinformática aplicada à vigilância genômica, e como ela tem sido importante para a compreensão de epidemias virais.

Os resultados destas metodologias são promissores. E trazem a possibilidade da aplicação da mesma metodologia para outros vírus de natureza zoonótica. Como perspectivas deste trabalho é pretendido a inclusão de novas técnicas de investigação genômica, como algoritmos de *Machine Learning*, que têm apresentado resultados promissores para diversos problemas biológicos complexos. Em conclusão, as análises de vigilância genômica se mostraram uma poderosa ferramenta na luta contra patógenos virais emergentes e reemergentes.

Referências Bibliográficas

ACKERMANN, M. *et al.* Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19. **New England Journal of Medicine**, v. 383, n. 2, p. 120–128, 9 jul. 2020.

ALANAGREH, L.; ALZOUGHOOL, F.; ATOUM, M. The Human Coronavirus Disease COVID-19: Its Origin, Characteristics, and Insights into Potential Drugs and Its Mechanisms. **Pathogens**, v. 9, n. 5, p. 331, 29 abr. 2020.

BANHO, C. A. *et al.* Impact of SARS-CoV-2 Gamma lineage introduction and COVID-19 vaccination on the epidemiological landscape of a Brazilian city. **Communications Medicine**, v. 2, n. 1, p. 41, dez. 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.

BUTT, A. M. *et al.* Evolution of codon usage in Zika virus genomes is host and vector specific. **Emerging Microbes & Infections**, v. 5, n. 1, p. 1–14, jan. 2016.

CHAFEKAR, A.; FIELDING, B. C. MERS-CoV: Understanding the Latest Human Coronavirus Threat. **Viruses**, v. 10, n. 2, 24 2018.

CHAN, J. F.-W. *et al.* Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. **Emerging Microbes & Infections**, v. 9, n. 1, p. 221–236, 1 jan. 2020a.

CHAN, J. F.-W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. **The Lancet**, v. 395, n. 10223, p. 514–523, 15 fev. 2020b.

CHEN, B. et al. Overview of lethal human coronaviruses. **Signal Transduction and Targeted Therapy**, v. 5, n. 1, p. 89, dez. 2020.

CHEN, L.; ZHONG, L. Genomics functional analysis and drug screening of SARS-CoV-2. **Genes & Diseases**, p. S2352304220300544, abr. 2020.

CONTRERAS-MOREIRA, B.; VINUESA, P. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Applied and* **Environmental Microbiology**, v. 79, n. 24, p. 7696–7701, dez. 2013.

CORMAN, V. M. et al. Hosts and Sources of Endemic Human Coronaviruses. In: **Advances in Virus Research**. [s.l.] Elsevier, 2018. v. 100p. 163–188.

COSTA, Z. G. A. *et al.* Evolução histórica da vigilância epidemiológica e do controle da febre

amarela no Brasil. **Revista Pan-Amazônica de Saúde**, v. 2, n. 1, p. 11–26, mar. 2011.

DE ARAÚJO, J. M. G. *et al.* A retrospective survey of dengue virus infection in fatal cases from an epidemic in Brazil. **Journal of Virological Methods**, v. 155, n. 1, p. 34–38, jan. 2009.

DE GROOT, R. J. *et al.* Middle East Respiratory Syndrome Coronavirus (MERS-CoV): Announcement of the Coronavirus Study Group. **Journal of Virology**, v. 87, n. 14, p. 7790–7792, jul. 2013.

DE LIMA, S. T. S. *et al.* Fatal Outcome of Chikungunya Virus Infection in Brazil. **Clinical Infectious Diseases,** v. 73, n. 7, p. e2436–e2443, 5 out. 2021.

DE OLIVEIRA FIGUEIREDO, P. *et al.* Re-Emergence of Yellow Fever in Brazil during 2016–2019: Challenges, Lessons Learned, and Perspectives. **Viruses**, v. 12, n. 11, p. 1233, 30 out. 2020.

DE SOUZA, A. S. *et al.* Severe Acute Respiratory Syndrome Coronavirus 2 Variants of Concern: A Perspective for Emerging More Transmissible and Vaccine-Resistant Strains. **Viruses**, v. 14, n. 4, p. 827, 16 abr. 2022.

DEJNIRATTISAI, W. *et al.* Antibody evasion by the P.1 strain of SARS-CoV-2. **Cell**, v. 184, n. 11, p. 2939- 2954.e9, maio 2021.

DENNEHY, J. J. Evolutionary ecology of virus emergence: Virus emergence. **Annals of the New York Academy of Sciences**, v. 1389, n. 1, p. 124–146, fev. 2017.

DEVER, T. E.; GREEN, R. The elongation, termination, and recycling phases of translation in eukaryotes. **Cold Spring Harbor Perspectives in Biology**, v. 4, n. 7, p. a013706, 1 jul. 2012.

DIECKMANN, M. A. *et al.* EDGAR3.0: comparative genomics and phylogenomics on a scalable infrastructure. **Nucleic Acids Research**, v. 49, n. W1, p. W185–W192, 2 jul. 2021.

DIEHL, W. E. *et al.* Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013–2016 Epidemic. **Cell**, v. 167, n. 4, p. 1088- 1098.e6, 3 nov. 2016.

DUDAS, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. **Nature**, v. 544, n. 7650, p. 309–315, abr. 2017.

DUMACHE, R. *et al.* SARS-CoV-2: An Overview of the Genetic Profile and Vaccine Effectiveness of the Five Variants of Concern. **Pathogens**, v. 11, n. 5, p. 516, 26 abr. 2022.

FARIA, N. R. *et al. Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus, Brazil*. preprint. [S.l.]: Epidemiology, 3 mar. 2021. Disponível em: <http://medrxiv.org/lookup/doi/10.1101/2021.02.26.21252554>. Acesso em: 13 jun. 2022.

FIGUEIREDO, L. T. M. Human Urban Arboviruses Can Infect Wild Animals and Jump to Sylvatic Maintenance Cycles in South America. **Frontiers in Cellular and Infection Microbiology**, v. 9, p. 259, 17 jul. 2019.

FLORES-VEGA, V. R. *et al.* SARS-CoV-2: Evolution and Emergence of New Viral Variants. **Viruses**, v. 14, n. 4, p. 653, 22 mar. 2022.

FRIAS, D. *et al*. Human Retrovirus Codon Usage from tRNA Point of View: Therapeutic Insights. **Bioinformatics and Biology Insights**, v. 7, p. BBI.S12093, jan. 2013.

FUNG, T. S.; LIU, D. X. Human Coronavirus: Host-Pathogen Interaction. **Annual Review of Microbiology**, v. 73, n. 1, p. 529–557, 8 set. 2019.

GEOGHEGAN, J. L.; HOLMES, E. C. Predicting virus emergence amid evolutionary noise. **Open Biology**, v. 7, n. 10, p. 170189, out. 2017.

GIOVANETTI, M. *et al.* Pan-genomics of virus and its applications. **Pan-genomics: Applications, Challenges, and Future Prospects**. [S.l.]: Elsevier, 2020. p. 237–250. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/B9780128170762000111>. Acesso em: 13 jun. 2022.

HAMRE, D.; PROCKNOW, J. J. A New Virus Isolated from the Human Respiratory Tract. **Proceedings of the Society for Experimental Biology and Medicine**, v. 121, n. 1, p. 190–193, 1 jan. 1966.

HINNEBUSCH, A. G.; LORSCH, J. R. The Mechanism of Eukaryotic Translation Initiation: New Insights and Challenges. **Cold Spring Harbor Perspectives in Biology**, v. 4, n. 10, out. 2012.

HO, J. M. *et al*. Drugging tRNA aminoacylation. **RNA biology**, v. 15, n. 4–5, p. 667–677, 2018.

HUANG, Y.-J. S.; HIGGS, S.; VANLANDINGHAM, D. L. Emergence and re-emergence of mosquito-borne arboviruses. **Current Opinion in Virology**, v. 34, p. 104–109, fev. 2019.

IBRAHIM, B. *et al.* A new era of virus bioinformatics. **Virus Research**, v. 251, p. 86–90, jun. 2018.

JENKINS, G. M.; HOLMES, E. C. The extent of codon usage bias in human RNA viruses and its evolutionary origin. **Virus Research**, v. 92, n. 1, p. 1–7, mar. 2003.

JI, W. et al. Cross‐species transmission of the newly identified coronavirus 2019‐nCoV. **Journal of Medical Virology**, v. 92, n. 4, p. 433–440, abr. 2020.

KARLIN, S.; BURGE, C. Dinucleotide relative abundance extremes: a genomic signature.

Trends in genetics: TIG, v. 11, n. 7, p. 283–290, jul. 1995.

KHAILANY, R. A.; SAFDAR, M.; OZASLAN, M. Genomic characterization of a novel SARS-CoV-2. **Gene Reports**, v. 19, p. 100682, jun. 2020.

LAM, T. T.-Y. *et al.* Dissemination, divergence and establishment of H7N9 influenza viruses in China. **Nature**, v. 522, n. 7554, p. 102–105, jun. 2015.

LI, G. *et al.* Coronavirus infections and immune responses. **Journal of Medical Virology**, v. 92, n. 4, p. 424–432, abr. 2020a.

LI, R. *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). **Science**, v. 368, n. 6490, p. 489–493, 1 maio 2020b.

LI, R. *et al.* SOAP: short oligonucleotide alignment program. **Bioinformatics**, v. 24, n. 5, p. 713–714, 1 mar. 2008.

LOWE, R. *et al.* The Zika Virus Epidemic in Brazil: From Discovery to Future Implications. **International Journal of Environmental Research and Public Health**, v. 15, n. 1, p. 96, 9 jan. 2018.

LU, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. **Lancet (London, England)**, v. 395, n. 10224, p. 565–574, 22 2020.

MAGALHAES, T. *et al.* The Endless Challenges of Arboviral Diseases in Brazil. **Tropical Medicine and Infectious Disease**, v. 5, n. 2, p. 75, 9 maio 2020.

MALIK, Y. A. Properties of Coronavirus and SARS-CoV-2. **The Malaysian Journal of Pathology**, v. 42, n. 1, p. 3–11, abr. 2020.

MANOKARAN, G. *et al.* Attenuation of a dengue virus replicon by codon deoptimization of nonstructural genes. **Vaccine**, v. 37, n. 21, p. 2857–2863, 9 maio 2019.

MCINTOSH, K. *et al.* Recovery in tracheal organ cultures of novel viruses from patients with respiratory disease. **Proceedings of the National Academy of Sciences**, v. 57, n. 4, p. 933–940, 1 abr. 1967.

MENG, B. *et al.* Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. **Cell Reports**, v. 35, n. 13, p. 109292, 29 jun. 2021.

MITTAL, A.; KHATTRI, A.; VERMA, V. Structural and antigenic variations in the spike protein of emerging SARS-CoV-2 variants. **PLOS Pathogens**, v. 18, n. 2, p. e1010260, 17 fev. 2022.

MUSSO, D.; GUBLER, D. J. Zika Virus. *Clinical Microbiology Reviews*, v. 29, n. 3, p. 487–524,

jul. 2016.

MUELLER, S. et al. A codon-pair deoptimized live-attenuated vaccine against respiratory syncytial virus is immunogenic and efficacious in non-human primates. **Vaccine**, v. 38, n. 14, p. 2943–2948, 23 mar. 2020.

NAKAMURA, Y. Codon usage tabulated from international DNA sequence databases: status for the year 2000. **Nucleic Acids Research**, v. 28, n. 1, p. 292–292, 1 jan. 2000.

NOGUEIRA, R. M. *et al.* Isolation of dengue virus type 2 in Rio de Janeiro. **Memorias Do Instituto Oswaldo Cruz**, v. 85, n. 2, p. 253, jun. 1990.

NUNES, M. R. T. *et al.* Emergence and potential for spread of Chikungunya virus in Brazil. **BMC medicine**, v. 13, p. 102, 30 abr. 2015.

NUNES, P. C. G. *et al.* 30 years of fatal dengue cases in Brazil: a review. **BMC Public Health**, v. 19, n. 1, p. 329, dez. 2019.

NÜRENBERG, E.; TAMPÉ, R. Tying up loose ends: ribosome recycling in eukaryotes and archaea. **Trends in Biochemical Sciences**, v. 38, n. 2, p. 64–74, fev. 2013.

ORELLE, C. *et al.* Identifying the targets of aminoacyl-tRNA synthetase inhibitors by primer extension inhibition. **Nucleic Acids Research**, v. 41, n. 14, p. e144–e144, 1 ago. 2013.

Oreshkova N, Molenaar RJ, Vreman S, Harders F, Oude Munnink BB, Hakze-van der Honing RW, Gerhards N, Tolsma P, Bouwstra R, Sikkema RS, Tacken MG, de Rooij MM, Weesendorp E, Engelsma MY, Bruschke CJ, Smit LA, Koopmans M, van der Poel WH, Stegeman A. SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. **Euro Surveill**. 2020 Jun;25(23):2001005. doi: 10.2807/1560-7917.ES.2020.25.23.2001005. Erratum in: Euro Surveill. 2021 Mar;26(12): PMID: 32553059; PMCID: PMC7403642.

OUDE MUNNINK, B. B. *et al.* The next phase of SARS-CoV-2 surveillance: real-time molecular epidemiology. **Nature Medicine,** v. 27, n. 9, p. 1518–1524, set. 2021.

PERLMAN, S.; NETLAND, J. Coronaviruses post-SARS: update on replication and pathogenesis. **Nature Reviews Microbiology**, v. 7, n. 6, p. 439–450, jun. 2009.

ROZEWICKI, J. *et al.* MAFFT-DASH: integrated protein sequence and structural alignment. **Nucleic Acids Research**, v. 47, n. W1, p. W5–W10, 2 jul. 2019.

RUDORF, S. Efficiency of protein synthesis inhibition depends on tRNA and codon compositions. **PLOS Computational Biology**, v. 15, n. 8, p. e1006979, 1 ago. 2019.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing

phylogenetic trees. **Molecular Biology and Evolution**, v. 4, n. 4, p. 406–425, 1 jul. 1987.

SHARP, P. M.; TUOHY, T. M. F.; MOSURSKI, K. R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. **Nucleic Acids Research**, v. 14, n. 13, p. 5125–5143, 1986.

SMITH, J. M. Optimization Theory in Evolution. **Annual Review of Ecology and Systematics**, v. 9, n. 1, p. 31–56, 1 nov. 1978.

TUSCHE, C.; STEINBRÜCK, L.; MCHARDY, A. C. Detecting patches of protein sites of influenza A viruses under positive selection. **Molecular Biology and Evolution**, v. 29, n. 8, p. 2063–2071, ago. 2012.

THOMSON, B. J. Viruses and apoptosis. **International Journal of Experimental Pathology**, v. 82, n. 2, p. 65–76, abr. 2001.

URBANOWICZ, R. A. *et al.* Human Adaptation of Ebola Virus during the West African Outbreak. **Cell**, v. 167, n. 4, p. 1079- 1087.e5, 3 nov. 2016.

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

VAN HEMERT, F. *et al.* **Euclidean Distance Analysis Enables Nucleotide Skew Analysis in Viral Genomes**. Research Article. Disponível em: <https://www.hindawi.com/journals/cmmm/2018/6490647/>. Acesso em: 30 ago. 2020.

VIANA, R. *et al.* Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. **Nature**, v. 603, n. 7902, p. 679–686, mar. 2022.

WAN, Y. et al. The Wuhan Pneumonia Outbreak: High Isoleucine and High Valine Plus Glycine Contents Are Features of the Proteins of COVID-19 Virus. 20 mar. 2020.

WERMELINGER, E. D. *et al.* Métodos e procedimentos usados no controle do Aedes aegypti na bem-sucedida campanha de profilaxia da febre amarela de 1928 e 1929 no Rio de Janeiro. **Epidemiologia e Serviços de Saúde**, v. 25, n. 4, p. 837–844, out. 2016.

WHITLOW, Z. W.; CONNOR, J. H.; LYLES, D. S. Preferential translation of vesicular stomatitis virus mRNAs is conferred by transcription from the viral genome. **Journal of Virology**, v. 80, n. 23, p. 11733–11742, dez. 2006.

WHITLOW, Z. W.; CONNOR, J. H.; LYLES, D. S. New mRNAs Are Preferentially Translated during Vesicular Stomatitis Virus Infection. **Journal of Virology**, v. 82, n. 5, p. 2286–2294, 1 mar. 2008.

WHO - **World Health Organization.** Acessado 01 set 2020.

WÖLFEL, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. **Nature**, v. 581, n. 7809, p. 465–469, maio 2020.

WOO, P. C. Y. *et al.* Characterization and Complete Genome Sequence of a Novel Coronavirus, Coronavirus HKU1, from Patients with Pneumonia. **Journal of Virology**, v. 79, n. 2, p. 884–895, 15 jan. 2005.

WU, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. **Cell Host & Microbe**, v. 27, n. 3, p. 325–328, 11 Mar. 2020.

ZANOTTO, P. M. DE A.; LEITE, L. C. DE C. The Challenges Imposed by Dengue, Zika, and Chikungunya to Brazil. **Frontiers in Immunology**, v. 9, p. 1964, 28 ago. 2018.

ZHANG, Y.-Z.; HOLMES, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. **Cell**, v. 181, n. 2, p. 223–227, 16 abr. 2020.

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

ZHU, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. **New England Journal of Medicine**, v. 382, n. 8, p. 727–733, 20 fev. 2020.

Apêndice A. Trabalhos participados durante o doutorado

1. Hanspers K, Kutmon M, Coort SL, Digles D, Dupuis LJ, Ehrhart F, Hu F, Lopes EN, Martens M, Pham N, Shin W, Slenter DN, Waagmeester A, Willighagen EL, Winckers LA, Evelo CT, Pico AR. **Ten simple rules for creating reusable pathway models for computational analysis and visualization**. PLoS Comput Biol. 2021 Aug 19;17(8):e1009226. doi: 10.1371/journal.pcbi.1009226.

2. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, A Miller R, Digles D, Lopes EN, Ehrhart F, Dupuis LJ, Winckers LA, Coort SL, Willighagen EL, Evelo CT, Pico AR, Kutmon M. **WikiPathways: connecting communities**. Nucleic Acids Res. 2021 Jan 8;49(D1):D613-D621. doi: 10.1093/nar/gkaa1024.

3. Giovanetti Marta, Alcantara Luiz Carlos Junior, Dorea Alfredo Souza, Ferreira Qesya Rodrigues, Marques Willian de Almeida, Junior Franca de Barros Jose, Adelino Talita Emile Ribeiro, Tosta Stephane, Fritsch Hegger, Iani Felipe Campos de Melo, Mares-Guia Maria Angélica, Salgado Alvaro, Fonseca Vagner, Xavier Joilson, Lopes Elisson Nogueira, Soares Gilson Carlos, Castro Amarante Maria Fernanda de, Azevedo Vasco, Kruger Alícia, Correa Matta Gustavo, Paineiras-Domingos Laisa Liane, Colonnello Claudia, Bispo de Filippis Ana Maria, Montesano Carla, Colizzi Vittorio, Barreto Fernanda Khouri, **Promoting Responsible Research and Innovation (RRI) During Brazilian Activities of Genomic and Epidemiological Surveillance of Arboviruses**, Frontiers in Public Health, Volume 9, 2021. DOI=10.3389/fpubh.2021.693743

4. Lopes EN, Fonseca V, Frias D, Tosta S, Salgado Á, Assunção Vialle R, Paulo Eduardo TS, Barreto FK, Ariston de Azevedo V, Guarino M, Angeletti S, Ciccozzi M, Junior Alcantara LC, Giovanetti M. **Betacoronaviruses genome analysis reveals evolution toward specific codons usage: Implications for SARS-CoV-2 mitigation strategies**. J Med Virol. 2021 Sep;93(9):5630-5634. doi: 10.1002/jmv.27056.

5. Luiz Carlos Junior Alcantara, Elisson Nogueira, Gabriel Shuab, Stephane Tosta, Hegger

Fristch, Victor Pimentel, Jayme A. Souza-Neto, Luiz Lehmann Coutinho, Heidge Fukumasu, Sandra Coccuzzo Sampaio, Maria Carolina Elias, Simone Kashima, Svetoslav Nanev Slavov, Massimo Ciccozzi, Eleonora Cella, José Lourenco, Vagner Fonseca, Marta Giovanetti, **SARS-CoV-2 epidemic in Brazil: how variants displacement have driven distinct epidemic waves**, Virus Research, 2022, 198785, ISSN 0168-1702, https://doi.org/10.1016/j.virusres.2022.198785.

6. Ágata Lopes-Ribeiro, Franklin Pereira Araujo, Patrícia de Melo Oliveira, Lorena de Almeida Teixeira, Geovane Marques Ferreira, Alice Aparecida Lourenço, Laura Cardoso Corrêa Dias, Caio Wilker Teixeira, Henrique Morais Retes, Élisson Nogueira Lopes, Alice Freitas Versiani, Edel Figueiredo Barbosa-Stancioli, Flávio Guimarães Da Fonseca, Olindo Assis Martins Filho, Moriya Tsuji, Vanessa Peruhype-Magalhães and Jordana Grazziela Coelho-dos-Reis, **In silico and in vitro arboviral MHC Class I-Restricted-epitope signatures reveal immunodominance and poor overlapping patterns,** Frontiers in Immunology, section Viral Immunology. 2022, https://doi.org/10.3389/fimmu.2022.1035515.