UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE BIOCIÊNCIAS

CAIO ANDREY BEZERRA JANUÁRIO

DIVERSIDADE E ESTRUTURAÇÃO GENÉTICA DO GENE SPIKE DE SARS-COV-2 CIRCULANTE NO ESTADO DE PERNAMBUCO.

Recife
2022

CAIO ANDREY BEZERRA JANUÁRIO

DIVERSIDADE E ESTRUTURAÇÃO GENÉTICA DO GENE SPIKE DE SARS-COV-2 CIRCULANTE NO ESTADO DE PERNAMBUCO.

Monografia apresentada à Coordenação do Curso de Ciências Biológicas Bacharelado da Universidade Federal de Pernambuco, como Requisito Parcial para obtenção do Título de Bacharel em Ciências Biológicas.

Orientador: Dr. Marcos da Silveira Regueira Neto

Coorientador: Dr. Valdir de Queiroz Balbino

Recife

2022

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Bezerra Januário, Caio Andrey.
Diversidade e estruturação genética do gene spike de SARS-CoV-2 circulante no estado de Pernambuco / Caio Andrey Bezerra Januário. - Recife, 2022.
51 p. : il., tab.

Orientador(a): Marcos da Silveira Regueira Neto
Cooorientador(a): Valdir de Queiroz Balbino
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Biociências, Ciências Biológicas - Bacharelado, 2022.
Inclui referências, apêndices.

1. COVID-19. 2. Nordeste. 3. Variantes. 4. Vigilância Epidemiológica. I. Silveira Regueira Neto, Marcos da. (Orientação). II. Queiroz Balbino, Valdir de. (Coorientação). III. Título.

500 CDD (22.ed.)

CAIO ANDREY BEZERRA JANUÁRIO

DIVERSIDADE E ESTRUTURAÇÃO GENÉTICA DO GENE SPIKE DE SARS-COV-2 CIRCULANTE NO ESTADO DE PERNAMBUCO.

Monografia apresentada à Coordenação do Curso de Ciências Biológicas Bacharelado da Universidade Federal de Pernambuco, como Requisito Parcial para obtenção do Título de Bacharel em Ciências Biológicas.

Aprovada em: ___/___/____

COMISSÃO EXAMINADORA

_____

Orientador: Dr. Marcos da Silveira Regueira Neto

UFPE

_____

Examinador Externo: Dr. Wilson José da Silva Júnior

Hospital Israelita Albert Einstein

_____

Examinador Interno: Dr. Rodrigo César Gonçalves de Oliveira

UFPE

AGRADECIMENTOS

Não existe espaço para colocar os nomes de todos aqueles que ajudaram a me tornar Eu.

Tudo que consegui conquistar é dedicado a minha família, meus pais, dona Gorete e seu Pedro, que sempre acreditaram em mim e não importa a ideia maluca que eu tenha, sei que eles sempre vão estar ao meu lado, a meus irmãos mais velhos, Nityeska e Wlademir que mesmo sem nunca entender direito o que eu faço da minha vida sempre foram abertos a ouvir e me apoiar em todas as pequenas conquistas que eu consigo, não importa quanto anos eu tenha sempre vou precisar de vocês, a meu irmão mais novo Henrique e meu sobrinho Gabriel que são meus orgulhos, sem vocês eu não estaria aqui hoje.

Quero agradecer a minha companheira Ana Carla, que nessa reta final da graduação foi fundamental em me ajudar a não surtar e me aguentou nos piores momentos e sempre consegue me ajudar e me colocar no eixo, sem você eu também não estaria aqui.

Sou grato a todas as pessoas que conheci ao longo da minha trajetória na graduação, em especial a meus amigos que me fizeram amar o dia a dia da universidade e me ensinaram tanto, deixo um agradecimento especial a Arthur, Edgar, Amanda, Beca, Lucas, Ellyson, Pollyana, Eduarda e Hugo.

Sou muito feliz por todas as pessoas que conheci no LABBE e que se tornaram tão especial para mim, Leandro, Dayane, Jussara, Wilson, Bandeira, Keyla, Karol e todos os novatos, que me ensinam e ensinaram tanta coisa, hoje muito do que aprendi foi através de vocês.

Um agradecimento mais que especial ao meu orientador Marcos Regueira, que me ensinou muito, com conselhos sobre a academia e sobre a vida, peço desculpas por ser um orientando difícil às vezes de lidar mas sou eternamente grato por toda paciência e lições que você me proporcionou, assim como meu coorientador Valdir Balbino que foi fundamental na minha formação.

Por fim queria agradecer a mim, por todos os momentos que não vi solução ou caminho para seguir, quando achei que não era capaz, mas não desisti, então obrigado Caio.

“Antes, a questão era descobrir se a vida precisava ter algum significado para ser vivida. Agora, ao contrário, está evidente que ela será melhor vivida se não tiver significado” (ALBERT CAMUS).

RESUMO

No ano de 2019, todo o mundo foi surpreendido pela COVID-19, causada por um vírus zoonótico, e que foi declarada como pandemia pela OMS em março de 2020, causando cerca de 100.000 mortes dentro de um mês após o anúncio da instituição. Em Pernambuco, o primeiro caso da doença se deu em 12 de março de 2020, e a epidemia foi caracterizada com a maior concentração de casos na capital pernambucana, seguindo posteriormente com uma rápida e intensa disseminação pelo interior do estado. Entretanto, a falta de amostragem de dados genômicos completos e de alta cobertura no Estado dificultava a investigação da diversidade genética e evolutiva do vírus. Dessa forma, o objetivo deste trabalho foi acompanhar o comportamento desse vírus e do gene Spike presente em sua superfície, assim como a ancestralidade do mesmo, além de acompanhar como a diversidade desse gene está se apresentando nas amostras sequenciadas em Pernambuco. Além disso, entender como esses clados e linhagens estão estruturadas geneticamente com o objetivo de compreender como os clados são compostos a nível genético e se há a sua sobreposição, e se os eventos mutacionais ocorridos no gene de interesse podem ou não ter conferido maior adaptabilidade do vírus e uma maior capacidade de infecção. Para isso, foram obtidos os genomas do vírus referente ao Estado na plataforma GISAID, e a partir da montagem de um dataset com o conjunto de dados filtrados, foram realizados a classificação desses clados e linhagens, seguido pela diversidade genética do gene nesses clados e a sua estruturação populacional. Em seguida, foi realizada uma análise filogenética desses genomas. Essas análises indicaram que os clados e linhagens tiveram predominância em épocas diferentes no estado de Pernambuco ao longo da pandemia, assim como as mutações sofridas e como o vírus tem se estruturado filogeneticamente durante esse período. No ano de 2020, a predominância foi do clado 20B (B.1.1 e B.1.1.28); e no período de 2021 até início de 2022, do clado 21J (AY.99.2) e 20J (P.1), que apresentou um maior impacto em Pernambuco. Além disso, análises corroborativas, como a distância de Nei's, entre outras, estão dando indícios que as linhagens tidas como única, podem estar se tornando distintas dentro da própria linhagem.

Palavras-chave: COVID-19; Nordeste; Variantes; Vigilância Epidemiológica.

ABSTRACT

In the year 2019, the whole world 2019, the whole world was affected by COVID-19, by a zoonotic virus, and which was declared as a pandemic by the WHO in 2020, institution of 10000 deaths within the month after the announcement of the date of a virus zoonotic. In Pernambuco, the first case of the disease occurred on intense 12, 2020, and the epidemic was the state with the highest concentration of people from Pernambuco, followed later by a rapid spread throughout the interior of Pernambuco. However, the lack of diversity of complete genomic data and high difficulty in investigating the diversity and evolution of the virus. Thus, the objective of this work was to follow the behavior of this virus and the Spike gene present on its surface, as well as its ancestry, in addition to following how the diversity of this gene is presenting itself in the samples sequenced in Pernambuco. In addition, to understand how these clades and lineages are genetically structured in order to understand how clades are composed at the genetic level and if there is an overlap, and whether mutational events that occurred in the gene of interest may or may not have conferred greater adaptability to the gene. virus and a greater capacity for infection. For this, the genomes of the virus referring to the State were obtained in the GISAID platform, and from the assembly of a dataset with the filtered data set, the classification of these clades and lineages were performed, followed by the genetic diversity of the gene in these clades and the its population structure. Then, a phylogenetic analysis of these genomes was performed. These analyzes indicated that the clades and lineages had predominance at different times in the state of Pernambuco throughout the pandemic, as well as the mutations suffered and how the virus has been phylogenetically structured during this period. In the year 2020, the predominance was of clade 20B (B.1.1 and B.1.1.28); and in the period from 2021 to early 2022, from clade 21J (AY.99.2) and 20J (P.1), which had a greater impact in Pernambuco. In addition, corroborative analysis, such as Nei's distance, among others, are indicating that lineages considered unique may be becoming distinct within the lineage itself.

Keywords: COVID-19; Epidemiological surveillance; North East; Variants.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Nas últimas décadas, o número de zoonoses foi constantemente um desafio para a saúde pública mundial. Essas zoonoses são caracterizadas, muitas vezes, pela aproximação de espécies selvagens ou domesticadas devido a perda de habitat ou a restrição do espaço físico (UNEP et al.,2016). Esta situação possibilita que alguns vírus, por sofrerem rápidas mutações e se tornarem capazes de infectar outras espécies, ultrapassem essa barreira e se tornem um patógeno de outros organismos. No ano de 2019, uma nova espécie de coronavírus se tornou o patógeno humano causador de uma das maiores crises sanitárias já enfrentadas, sendo atualmente o vírus zoonótico de importância significativa dentro da epidemiologia (MISHRA et al., 2021).

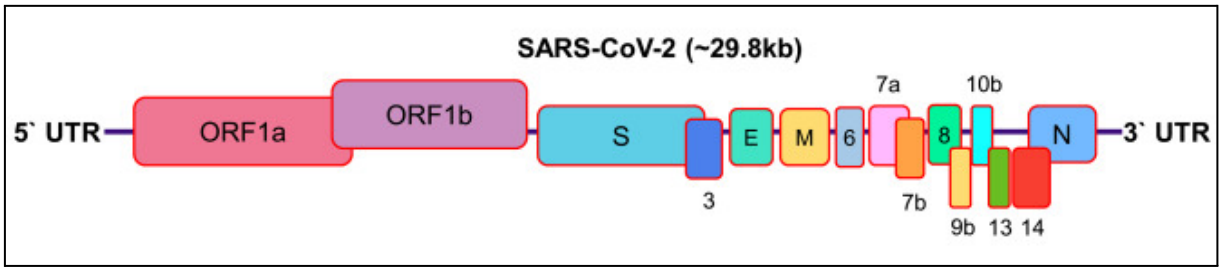
Os coronavírus são vírus envelopados, aproximadamente esféricos, de RNA de sentido positivo (+ssRNA), e variam entre 80 nm a 220 nm de diâmetro com projeções semelhantes a picos em sua superfície com tamanho genômico de 29,9 kb (LU et al., 2020). Esses picos característicos deram origem ao seu nome, uma vez que ao microscópio eletrônico, sua aparência se assemelha com uma coroa. São pertencentes à família *Coronaviridae*, da ordem *Nidovirales*, e possuem quatro gêneros: *Alphacoronavirus* (alfa-CoV), *Betacoronavirus* (beta-CoV), *Gammacoronavirus* (gamma-CoV) e *Deltacoronavirus* (delta-CoV). Os *Alfacoronavírus* e *Betacoronavírus* podem ser encontrados principalmente em mamíferos, enquanto *Gammacoronavírus* e *Deltacoronavírus* são encontrados com maior facilidade em aves, embora há casos de infecção por *Gammacoronavírus* também em alguns cetáceos. Em humanos, *Alfacoronavírus* e *Betacoronavírus* geralmente causam doenças respiratórias, enquanto em outros animais é mais comum causarem doenças gastrointestinais (DECARO et al., 2008).

Uma das características importantes sobre os coronavírus é a presença de um genoma com altas taxas de mutação, uma vez que possuem mecanismos complexos de revisão e reparo de erros de replicação (DENISON et al., 2011), o que facilita sua sobrevivência e a seleção de vírus recombinantes em diferentes populações (CORMAN et al., 2018). Além disso, apresentam em sua estrutura quatro genes estruturais: gene de membrana (M), de envelope (E), do nucleocapsídeo (N) e o gene spike (S) (Figura 1). O gene S codifica a glicoproteína Spike (S), que é a mais importante para o início da infecção, uma vez que é responsável por conectar-se aos receptores do hospedeiro através do domínio RDB (*receptor-binding domain*), além de neutralizar os anticorpos e mediar a fusão à membrana e a entrada na célula. Didaticamente ela é dividida em duas regiões: S1, que intervêm na ligação na ligação do vírus ao receptor da célula hospedeira, e a S2, que é responsável pela fusão do vírus à membrana da célula (RAHMAN et al., 2021).

No código genético viral, além dos genes estruturais são encontrados os *Open Reading Frames* (ORFs), que produzem as proteínas não estruturais envolvidas na construção do vírus durante sua replicação na célula hospedeira, que são a ORF1a e ORF1b (LI et al., 2020), entretanto existem outras ORFs que

intercalam os genes estruturais e não tem função na replicação viral, que são: ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8a, ORF8b e ORF9b (YOUNT et al., 2005), essas proteinas estruturais ou acessorias estão ligadas na regulação durante a replicação viral (TAN et al., 2021).

Figura 1 - Estrutura genômica do SARS-CoV-2.



Fonte: Adaptado de Kirtipal (2020).

Em humanos, a primeira vez que o coronavírus foi isolado ocorreu no final da década de 1960. Entretanto, foi somente após o surgimento da espécie SARS-CoV, responsável pela Síndrome Respiratória Aguda Grave em 2002, que os coronavírus foram colocados como potenciais ameaças à saúde humana mundial (CHAU et al., 2021). Em dezembro de 2019, na cidade de Wuhan, localizada na província de Hubei na China, iniciou-se outro surto de insuficiência respiratória, semelhante ao que aconteceu com o SARS-CoV, porém que surpreendeu o mundo ao se espalhar de forma tão rápida entre os continentes. A nova doença, causada pelo vírus SARS-CoV-2, ficou conhecida como COVID-19 (do inglês – *COronaVIrus Disease* 2019 – doença por coronavírus do ano 2019), e declarada como pandemia em 11 de março de 2020 pela Organização Mundial da Saúde (OMS), causando cerca de 100.000 infectados dentro de um mês após o anúncio da instituição (CALLAWAY et al., 2020).

Desde o início da pandemia da COVID-19 foram observadas diversas mutações em seu genoma. Essas mutações são eventos naturais da replicação viral principalmente em vírus de RNA, e em sua maioria se apresentam de forma neutra, não mostrando vantagem ou desvantagem para o vírus (LOKMAN et al., 2020). Entretanto, algumas mutações podem atribuir novas propriedades químicas às proteínas virais, que alteram a forma como o vírus costuma se comportar nas infecções. Dentre essas mutações, as que ocorreram na proteína Spike (S) são as mais relevantes no sentido clínico e epidemiológico, uma vez que essa proteína apresenta um importante papel no processo infeccioso (LOKMAN et al., 2020). Uma dessas mutações na proteína S resultou numa maior afinidade da ligação com o receptor ACE2 humano. O receptor ACE2 (Enzima Conversora da Angiotensina 2) é amplamente expresso em células epiteliais dos alvéolos pulmonares, em enterócitos do intestino delgado, em células endoteliais venosas e arteriais (HAMMING et al., 2004) e em células da mucosa oral (XU et al., 2020). Outra mutação importante que propiciou uma maior disseminação desse vírus é aquisição da capacidade de utilizar a furina, uma protease transmembrana que é expressa em todos os tipos celulares,

para "pré ativar" as glicoproteínas da superfície viral, o que por sua vez facilita a propagação do vírus célula-célula (SHANG et al., 2020), e a escapar da resposta de anticorpos do hospedeiro.

Quando é necessário lidar com uma grande quantidade de informações é inevitável a formação dos bancos de dados, que são de forma geral um local onde agrupamos informações de um determinado tipo, no caso de SARS-CoV-2 é necessário a criação de bancos de dados, desde de informações de amostras do vírus, o foco desse trabalho, até exomas dos pacientes e outras informações retiradas de da biologia do patógeno e hospedeiros, estruturados em milhões de gigabytes, na pandemia recente o banco de dados mais conhecido e importante para essa doença é o *Global Initiative on Sharing All Influenza Data* (GISAID) (ELBE et al., 2017) que agrupa informações primordiais para os estudos envolvendo esse vírus assim como o da influenza, no presente trabalho em diversos momentos foi criado bancos de dados específicos, assim como bancos auxiliares em algumas etapas (MOREIRA et al., 2002).

Os dados coletados e sequenciados precisam passar por análises posteriores para só então os cientistas poderem criar interpretações e ações em relação a pandemia, tanto compreendendo como o patógeno está se comportando dentro das populações assim como o possível impacto dessa variação, com isso foi e continua sendo crucial a união com a bioinformática para a compressão em relação ao dados (VERLI et al., 2014), ferramentas como DnaSP (Rozas et al., 2017) para análises de polimorfismo, FastQC (Brown et al., 2017) usado para verificação de qualidade da sequências brutas saídas do sequenciador, Cutadapt (Martin et al., 2011) utilizado para limpeza da sequências, esses são um pequeno exemplo de como a bioinformática é necessária não só para SARS-CoV-2 mas para em relação a todos os dados biológicos construídos nessa geração tecnológica que são gerados cada vez mais dados em um espaço de tempo muito curto (RAY et al., 2021).

Com o rastreamento da disseminação dessas mutações identificadas no SARS-CoV-2 nas diferentes partes do mundo, através do acesso aos bancos de dados e plataformas genômicas, como o *GISAID*, por exemplo, foi possível identificar os diferentes grupos genéticos virais, também denominado linhagens, e o surgimento de mutações adicionais, que promove diferença dentro de cada grupo genético, as chamadas variantes. Dessa forma, foi possível observar a crescente modificação do SARS-CoV-2 ao longo do tempo, assim como a possibilidade de cocirculação de diferentes clados com diferentes linhagens e variantes em diversos países e a migração entre eles (O'TOOLE et al., 2021).

A OMS, que acompanha a evolução do vírus da COVID-19, inclusive avaliando fatores como a transmissibilidade, a virulência, entre outros, classificou as variantes circulantes globalmente como variantes de preocupação (VOC, do inglês "*variant of concern*") e variantes de interesse em saúde pública (VOI, do inglês "*variant of interest*"). Entre as VOC, tem-se as variantes Alfa (B.1.1.7), identificada inicialmente no Reino Unido, Beta (B.1.351), descoberta na África do Sul, Gama (B.1.1.28.1), originária do Brasil (Manaus) e Delta (B.1.617.2), identificada na Índia.

Já no grupo das VOI, tem-se as variantes Eta (B.1.525), detectada em diversos países, Epsilon (B.1.427/B.1.429), identificada nos Estados Unidos da América (Califórnia), Zeta (B.1.1.28.2), originária do Brasil (Rio de Janeiro), Teta (B.1.1.28.3), detectada nas Filipinas e no Japão, Iota (B.1.526), descoberta nos Estados Unidos da América, Kapa (B.1.617.1) detectada na Índia, e Lambda (C.37), originária do Peru (Tabela 1) (RAMBAUT et al., 2020).

Essas linhagens estão agrupadas em diferentes clados, que são descritos como ramos dentro de uma filogenia que possuem um ancestral comum e todos os seus descendentes. Os principais clados definidos foram apresentados na tabela 1.

Tabela 1 - Sistema de nomenclatura dos genomas de SARS-CoV-2 e como eles se relacionam, clados do Nextstrain, as linhagens do Pangolin e as variantes nomeadas pela OMS.

Clado do Nextstrain	Linhagens no Pangolin	OMS
20I (Alfa, V1)	B.1.1.7	α Alfa
20H (Beta, V2)	B.1.351	β Beta
20J (Gama, V3)	P.1	γ Gama
21A (Delta)	B.1.617.2	δ Delta
21I (Delta)	AY.47	δ Delta
21J (Delta)	AY.99.2	δ Delta
21B (Kappa)	B.1.617.1	κ Kappa
21C (Épsilon)	B.1.427,B.1.429	ε Épsilon
21D (Eta)	B.1.525	η Eta
21F (Iota)	B.1.526	ι Iota
21G (Lambda)	C.37	λ Lambda
21H (Mu)	B.1.621	μ Mu
21K (Ômícron)	BA.1	o Omicron
21L (Ômícron)	BA.2	o Omicron
22A (Ômícron)	BA.4	o Omicron
22B (Ômícron)	BA.5	o Omicron
22C (Ômícron)	BA.2.12.1	o Omicron
22D (Ômícron)	BA.2.75	o Omicron
20E (EU1)	B.1.177	
20B / **S :** 732A	B.1.1.519	
20A / **S :** 126A	B.1.620	
20A .EU2	B.1.160	
20A / **S :** 439K	B.1.258	
20A / **S :** 98F	B.1.221	
20C / **S :** 80Y	B.1.367	

20B / **S :** 626S	B.1.1.277
20B / **S :** 1122L	B.1.1.302

Fonte: covariants.org (2022).

No Brasil, a primeira notificação de caso confirmado de COVID-19 foi no dia 26 de fevereiro de 2020, no Estado de São Paulo. A epidemia, que teve início com várias introduções independentes e posteriormente com a transmissão comunitária, foi primeiramente impulsionada principalmente pelas linhagens B.1.1.28 e B.1.1.33, que devido ao nível de importância o Ministério da Saúde Brasileiro determinou no primeiro trimestre de 2020 estado de Emergência em Saúde Pública de Importância Nacional (BRASIL et al., 2020). Essas linhagens permaneceram prevalentes até outubro de 2020, quando começou-se a circular duas variantes de origem nacional: P.1 e P.2, originadas da linhagem B.1.1.28. Após quatro meses do surgimento das duas variantes brasileiras, elas corresponderam juntas a 75% dos sequenciamentos no território nacional. Em relação a notificação de VOC e VOI no Brasil, das sete variantes pela OMS, foram notificadas as quatro variantes classificadas como VOC e duas classificadas como VOI: Zeta e Lambda (COVID et al., 2021).

Em Pernambuco, um Estado do nordeste brasileiro, o primeiro caso de COVID-19 se deu em 12 de março de 2020, e a epidemia foi caracterizada com a maior concentração de casos na capital pernambucana, seguindo posteriormente com uma rápida e intensa disseminação pelo interior do estado. Apesar do alto número de casos durante a pandemia, a amostragem de dados genômicos no estado ainda é precária, o que impede o investigação da diversidade genética e evolutiva do SARS-CoV-2, e por sua vez dificulta o acompanhamento e a prevenção de futuros surtos de COVID-19 (SILVEIRA et al., 2022). Dessa forma, acompanhar as mudanças do vírus no estado de Pernambuco é um fator fundamental para as decisões sanitárias, assim como para reduzir o impacto causado pela pandemia de COVID-19.

2. OBJETIVOS

2.1 OBJETIVO GERAL

Analisar a diversidade e a estruturação genética do gene Spike de cepas de SARS-CoV-2 circulantes no estado de Pernambuco do início da pandemia até os primeiros meses de 2022.

2.2 OBJETIVOS ESPECÍFICOS

- Realizar a designação das linhagens e clados das sequências incluídas no estudo;
- Isolar o gene spike no conjunto de dados;
- Avaliar a diversidade genética do gene spike;

- Compreender a estruturação populacional das linhagens dentro dos clados levando-se em conta o gene Spike;
- Realizar uma análise filogenética do conjunto de dados do gene spike de SARS-CoV-2 circulante no estado de Pernambuco.

3. METODOLOGIA

3.1 Montagem do Banco de Dados

As sequências genômicas do vírus SARS-CoV-2 foram obtidas do banco de dados internacional GISAID, *Global Initiative on Sharing Avian Influenza Data*, que passou a armazenar os dados também de coronavírus em dezembro de 2019, no início da epidemia. O conjunto de dados referente a todos os genomas publicados no estado de Pernambuco foram resgatados no dia 18 de julho de 2022, do período da primeira submissão em 28 de julho de 2020 ao da última submissão em 6 de junho de 2022, sendo assim recuperados 4.109 genomas. Entretanto, esse conjunto de dados foi submetido a um processo de filtragem, permanecendo assim os que tivessem o genoma completo, são os que têm >29.000nt e com <1% de bases indefinidas (Ns) e com alta cobertura, que é definido como tendo <1% Ns e <0,05% de mutações, restando 1.863 genomas de Pernambuco para compor o dataset.

Essas sequências de nucleotídeos foram selecionadas e alinhadas com o genoma de referência Wuhan-Hu1 (número de acesso GenBank NC_045512), através do software para alinhamento múltiplo de sequências MAFFT (KATOH et al., 2019). Neste software é utilizado método progressivo de alinhamento.

3.2 Classificação dos Clados e Linhagens

Para identificar as linhagens e clados, os genomas presentes no dataset foram submetidos à análise de classificação das linhagens por meio do método de classificação proposto por Rambaut et al. (2020), através do software PANGOLIN v.1.1.14 (*Phylogenetic Assignment of Named Global Outbreak Lineages*) e para os respectivos clados foi utilizado o software Nextstrain v.1.16.5 (HADFIELD et al., 2018) A partir dessas plataformas, obteve-se os clados e as linhagens dos dados de SARS-CoV-2. A codificação dos clados, de maneira geral, é mais abrangente e aborda a diversidade geral das linhagens que as compõem, enquanto que as linhagens correspondem melhor às variantes que geram os surtos (ALM et al., 2020). Dessa forma, os clados foram utilizados como as populações nas análises posteriores por abranger uma maior diversidade genética.

Nessa etapa foram submetidos os genomas completos, pois para a classificação das linhagens e clados é levado em consideração diversas partes do genoma, logo utilizar poucas regiões ou genomas não completos pode gerar classificações equivocadas.

Posterior a esta etapa, o dataset foi trimado manualmente através do Aliview (LARSSON et al., 2014) com o objetivo de isolar somente a região de interesse: o

gene Spike, localizado na posição dos pares de base do 21563 ao 25384, dentro do genoma completo de referência (NC_045512.2) composto por 29.903 pares de bases.

3.3 Diversidade Genética

Para compreender a diversidade genética foram levados em consideração os clados como populações. Os parâmetros utilizados, de forma geral, foram: número de haplótipos (h), a diversidade de haplótipos (Hd), a diversidade nucleotídica (Pi), o número de sítios polimórficos, o número de mutações e quantas são compartilhadas entre as populações, o número de sítios singleton (S) e parcimônio-informativos (Ps), o número médio de substituições nucleotídicas por sítio entre populações (Dxy),o número de sítios segregantes (S), o número total de substituições por sítio entre populações (Da) e o número médio de diferenças de nucleotídeos (k). Essas análises foram feitas com o dataset completo, e também para cada clado separadamente. Todas essas medidas foram realizadas com o software DnaSP v. 5.0 (LIBRADO et al., 2009), utilizando os pares de base referente ao gene Spike.

Além disso, foram submetidas aos testes de neutralidade D de Tajima e Fs de Fu através do pacote de software Arlequin v.3.5 (EXCOFFIER et al., 2010) para 10.000 amostras simuladas, a fim de se observar se os clados se encontram em expansão ou regressão. Também na plataforma Arlequin v. 3.5, a fim de compreender a estrutura genética das populações, foi medido a diferenciação genética entre as populações por meio do Índice de Fixação (Fst) (Weir and Cockerham, 1984; Michalakis and Excoffier, 1996), com a execução de 100.000 gerações de cadeia de Markov.

Para compreender melhor a diferenciação populacional em relação aos clados foi gerado o diagrama baseado na distância de Nei's também no Arlequin v. 3.5, em que foram utilizadas 100.000 rodadas de cadeia de Markov.

3.4 Estruturação Populacional

Foram criados diferentes datasets com as respectivas sequências de cada clado, a fim de se observar como esses clados estão estruturados. Para os clados que se encontravam com mais de 300 sequências, foi criado dataset reduzido com duas amostras por linhagem presentes neste clado. As escolhas das amostras foram feitas de forma aleatória.

Foram selecionados nesses datasets os dados de SNPs (*Single nucleotide polymorphism*) para a análise de estruturação utilizando o FastStructure (Raj et al., 2014), que implementa a análise pelo método Bayesiano. Em cada um dos datasets criados foram testadas 30 populações possíveis. Esse software foi escolhido por proporcionar maior agilidade em relação ao tempo de análise com grande volume de dados em comparação com outros programas como o Structure.

A partir dos resultados, foi produzido um script no programa R (TEAM et al., 2020) em conjunto com o pacote Pophelper v.2.3.1 (FRANCIS et al., 2017) para a

visualização dos dados de forma mais detalhada. O pacote Pophelper é uma ferramenta para auxiliar nas montagens dos resultados das análises da estrutura populacional. Pode ser usado para a análise de resultados gerados a partir de outros programas de atribuição de população, como Admixture, Structure e Tess.

3.5 Análise Filogenética

Através das ferramentas disponibilizadas pelo Nextstrain (HADFIELD et al., 2018), que possibilita a produção das árvores filogenéticas e detecção das mutações anotadas de SARS-CoV-2 e outros vírus de interesse, foi realizada a análise filogenética das sequências por meio da ferramenta Augur (HUDDLESTON et al., 2021), que também constrói árvores filogenéticas através do IQ-TREE usando método de Máxima Verossimilhança (ML - *Maximum Likelihood*) e faz rastreamento da sua evolução ao longo do tempo utilizando o TreeTime. O melhor modelo para construção foi o GTR. O resultado foi observado em outra ferramenta do Nextstrain, o Auspice (HADFIELD et al., 2018) que foi criado para se observar os resultados filogenéticos.

4. RESULTADOS

3.1 Classificação das linhagens e clados

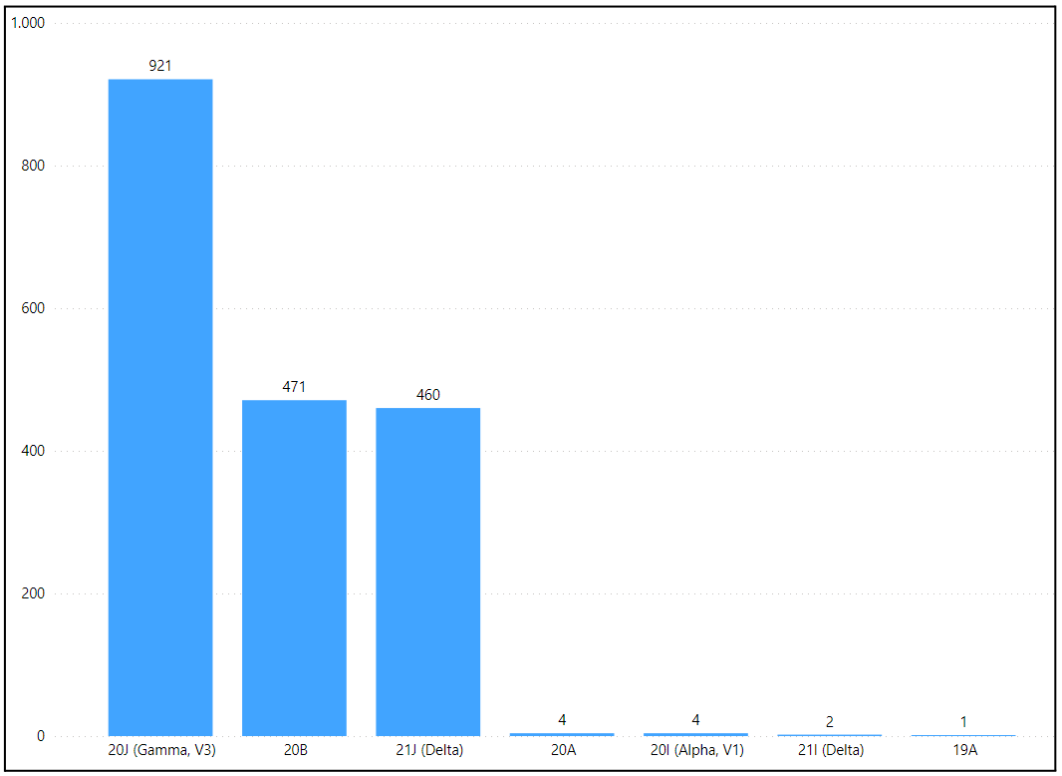
A partir dos 1.863 genomas completos utilizados para a classificação, identificou-se que ao longo da pandemia no estado de Pernambuco estavam presentes sete clados, que compreendem mais de trinta linhagens, no período do final de 2020 até o fim do primeiro semestre de 2022. Entre os principais clados estão 21J e 20J, que são conhecidos nos meios de comunicação como as variantes Delta e Gamma respectivamente (Tabela 2). Como esperado, os clados que apresentaram maior quantidade de sequências também tiveram, por consequência, a maior quantidade de linhagens identificadas. Os três clados mais diversos foram: 21J portando 14 linhagens, sendo o mais diverso, seguido pelo clado 20B com 10 linhagens e por último o 20J com 6 linhagens presentes, os outros clados apresentaram baixa quantidade de linhagens. Em relação a classificação das linhagens, foi observado que as duas linhagens mais presentes foram a P.1, que obteve mais de 40% de representatividade no conjunto de dados, seguida pela AY.99.2 com quase 20%, as outras linhagens ficaram abaixo de 10% (Figura 3), quando observamos os clados, notamos que tres clados 20J, 20B e 21J foram os mais expressivos dentro do nossos dados (FIgura 2). Por se tratar de três anos de pandemia compilados em um dataset, era de se esperar uma quantidade elevada de linhagens e clados. Ao longo dos meses ocorrem flutuações no seu domínio na população naquela faixa de tempo, e com base nos dados foi possível observar a distribuição das linhagens e clados ao longo dos meses nesses três anos de pandemia e demonstrar parcialmente a situação de Pernambuco (Figura 4).

Tabela 2 - Relação das quantidade de genomas utilizados, clados e linhagens que foram classificados dentro do nosso conjunto de dados através do software Pangolin e Nextstrain.

N.º de genomas	Clado do Nextstrain	Linhagens do Pangolin
1	19A	B
4	20A	B.1, B.1.212
470	20B	P.2, B.1.1, B.1.1.117, B.1.1.192, B.1.1.28, B.1.1.33, B.1.1.348, B.1.1.371, B.1.1.398, N.9
4	20I (Alpha,V1)	B.1.1.7
922	20J (Gamma,V3)	P.1, P.1.14, P.1.2, P.1.4, P.1.7, P.1.8
2	21I (Delta)	AY.47
460	21J (Delta)	AY.47, AY.100, AY.101, AY.113, AY.122, AY.124.1, AY.34.1.1, AY.36, AY.42, AY.43, AY.46.3, AY.6, AY.99.1, AY.99.2, B.1.617.2

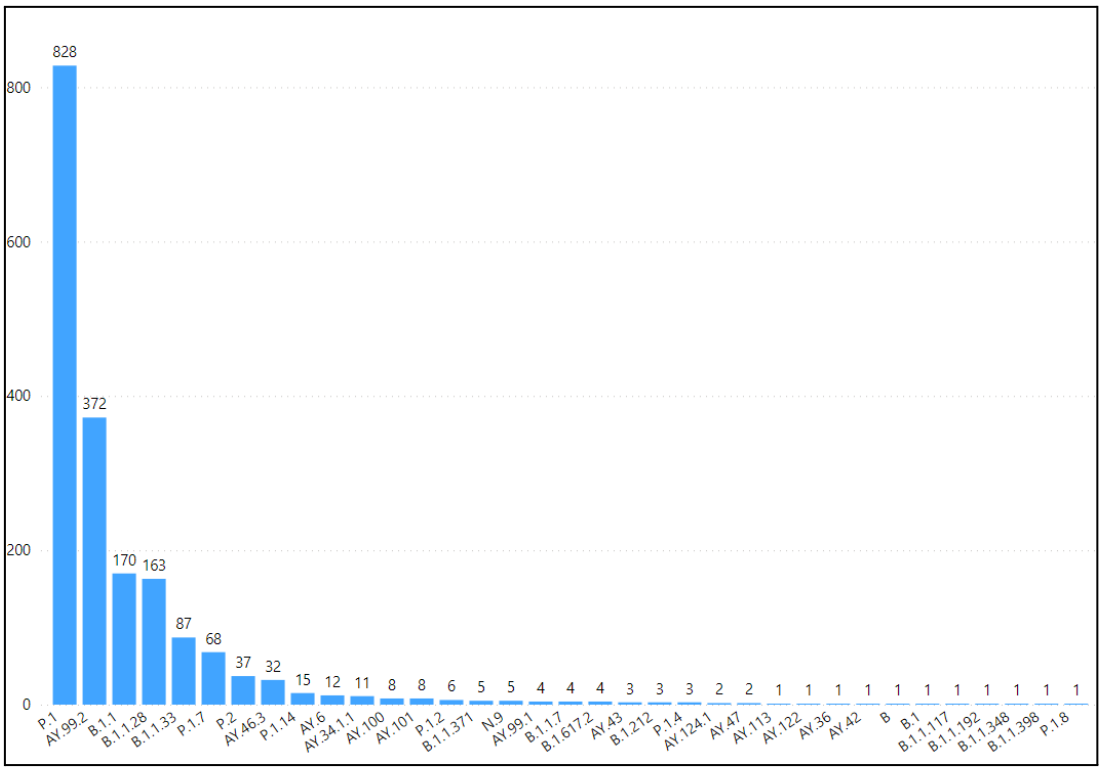
Fonte: O autor (2022).

Figura 2 - Distribuição dos clados encontrados no conjunto de dados, acima das barras de linhagens está o total de genomas.



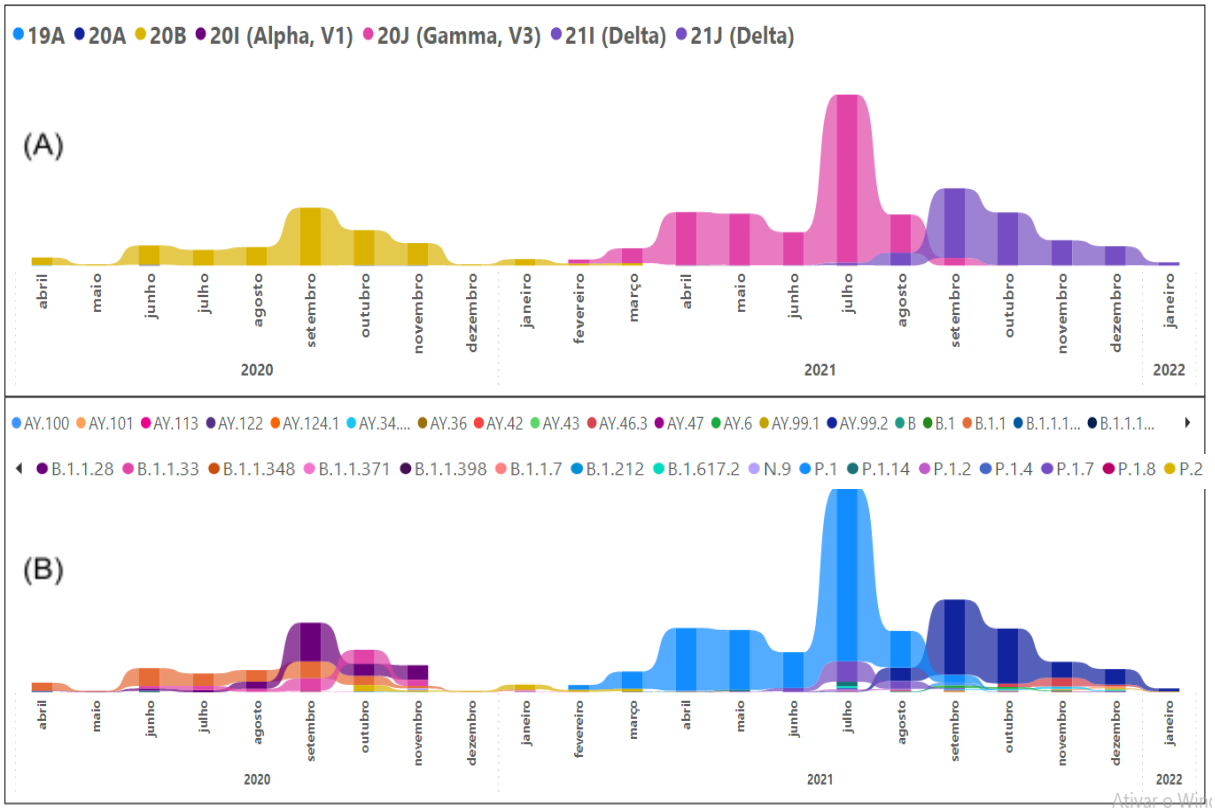
Fonte: O autor (2022).

Figura 3 - Distribuição das linhagens encontradas no conjunto de dados, acima das barras de linhagens está o total de genomas.



Fonte: O autor (2022)

Figura 4 - (A) Presença dos clados encontrados no nosso conjunto de dados ao longo dos três anos de pandemia no estado de Pernambuco. (B) Presença das linhagens encontradas no nosso conjunto de dados ao longo dos três anos de pandemia no estado, sendo em B.1.1 (laranja), B.1.1.28 (roxo), B.1.1.33 (rosa), P.2 (amarelo), P.1 (azul claro) e AY.99.2 (azul escuro).



Fonte: O autor (2022).

3.2 Diversidade Genética

Ao se fazer as análises para compreender a diversidade genética, para o sítios polimórficos foram observados os seguintes resultados para o nosso conjunto de dados da Spike (3.821bp). Foram encontrados 230 mutações, 217 sítios variáveis polimórficos, sendo 97 do tipo variável singleton, com 95,9% desses sítios portando duas bases nucleotídicas possíveis e com restante portando três ou as quatros nos sítios variáveis. Também foi encontrado 120 sítios parcimoniosamente informativos, 93,3% dos sítios sendo composto por duas bases e o restante por três bases (Tabela 3).

Tabela 3 - Relação da quantidade de mutações, sítios polimórficos do tipo singleton e parcimoniosamente informativos, compostos por dois a quatro nucleotídeos possíveis nos sítios variáveis, para nossos dados da Spike.

	Total de mutações	Sítios Polimórficos	Singleton	Parcimoniosamente Informativo
Total	230	217	97	120
Duas bases			93	112
Três bases			3	8
Quatro bases			1	0

Fonte: O autor (2022).

Quando se analisou os clados separadamente, no clado 21J foram observadas 79 mutações e variações polimórficas, sendo 37 sítios variáveis do tipo singleton e 42 sítios parcimoniosamente informativos, ambos presentes somente com dois nucleotídeos possíveis. Já no clado 20J, obteve-se 190 mutações com 181 sítios variáveis polimórficos, sendo 96,8% do tipo singleton composto por dois tipos de nucleotídeos nos sítios presentes e 94,2% do tipo parcimoniosamente informativo. No clado 20B, se observou 110 mutações com 107 sítios polimórficos, com 55 sítios singleton, desses 98,1% dos sítios sendo composto só por duas bases possíveis e 52 sítios parcimoniosamente informativos, sendo 96,1% composto por duas possibilidade de bases. Os outros clados, apesar de apresentarem valores baixos, também foram compilados (Tabela 4).

Tabela 4 - Dados sobre as mutações, sítios polimórficos, singleton e parcimoniosamente informativos separados por clados. Clados que obtiveram 0 em qualquer coluna ou linha foram ocultados para ficar mais fácil a visualização dos resultados. O clado 19A não obteve nenhum resultado por apresentar só um genoma no conjunto de dados.

	Total de mutações	Sítios Polimórficos	Singleton	Parcimoniosamente Informativo
Total	(20A - 2) (20B - 110) (20I - 1) (20J - 190) (21I - 1) (21J - 79)	(20A - 2) (20B - 107) (20I - 1) (20J - 181) (21I - 1) (21J - 79)	(20A - 2) (20B - 55) (20I - 1) (20J - 94) (21I - 1) (21J - 37)	(20B -52) (20J - 87) (21J - 42)
Duas bases			(20A - 2) (20B - 54) (20I - 1) (20J - 91) (21I - 1) (21J - 37)	(20B - 50) (20J - 82) (21J - 42)
Três bases			(20B - 1) (20J - 2)	(20B - 2) (20J - 5)
Quatro bases			(20J - 1)	

Fonte: O autor (2022).

Em complemento, para o dataset com todos os clados, foi encontrado 275 haplótipos (h), com diversidade haplotípica (Hd) de 0,8363, a diversidade nucleotídica (Pi) em 0,00208 e com a média de diferença nucleotídica (k) em 4,127. Quando comparando os clados entre si, os resultados mais expressivos foram a comparação do clado 21J e 20J, ambos os clados apresentam diversidade nucleotídica (Pi) muito similar, com menos de 1% de diferença, apresentam um valor de Dxy médio e total muito baixo, equivalente a 0,00371 e Da 0,00332, mas os dois clados compartilham 23 mutações e apresentam a diferença média de nucleotídeos de 8,811. Na comparação de 21J contra 20B, detectou que eles compartilham 15 mutações, a Pi é bastante similar, apresentam o Dxy de 0,00139 e com diferença média entre nucleotídeos de 3,670. O clado 20B apresentou maior Pi, sendo quase duas vezes maior que a Pi de 20J. Esses dois clados partilham a maior quantidade de mutações dentre os clados, sendo 27 partilhadas (Tabela 5).

Tabela 5 - Valores encontrados na comparação entre clados, para diversidade nucleotídica (Pi), número médio de substituições nucleotídicas por sítio entre populações (Dxy), número total de substituições por sítio entre populações (Da) e o número médio de diferenças de nucleotídeos (K) e as quantas mutações os clados compartilham.

	Pi	Dxy	Da	K	Mutações Compartilhadas
21J/20J	0,00046/0,00038 (0,00186)	0,00371	0,00332	8,811	23
21J/20I	0,00046/0,00015 (0,00049)	0,00229	0,00199	7,540	1
21J/20A	0,00046/0,00031 (0,00047)	0,00101	0,00062	3,285	1
21J/20B	0,00052/0,00059 (0,00097)	0,00139	0,00084	3,670	15
21I/20I	0,00027/0,00013 (0,00197)	0,00355	0,00355	13,250	0
21I/20A	0,00027/0,00027 (0,00128)	0,00216	0,00189	8,000	0
21I/20B	0,00036/0,00058 (0,00059)	0,00196	0,00149	5,500	0
20I/20A	0,00013/0,00027 (0,00119)	0,00193	0,00173	7,250	0
20I/20B	0,00017/0,00057 (0,00060)	0,00200	0,00163	5,752	0
20A/20B	0,00035/0,00058 (0,00058)	0,00053	0,00007	1,515	1
20J/20I	0,00038/0,00020 (0,00042)	0,00439	0,00410	11,190	0
20J/20A	0,00038/0,00020 (0,00040)	0,00306	0,0277	7,676	0
20J/20B	0,00037/0,00065 (0,00156)	0,00296	0,00245	6,376	27

Fonte: O autor (2022).

Na análise intra-populacional para os clados, se observou para frequência de haplótipos que os clados 20J, 21J e 20B apresentaram as maiores frequências, sendo 337, 241 e 162 respectivamente, os outros clados obtiveram frequência

menores que 5 haplótipos cada um. Ao realizar o teste de neutralidade de Tajima o clado 21J foi apresentado 86 sítios com substituições (S) e com valor de Tajima's D de -2,43189, e no teste de neutralidade de Fs de Fu apresentou um Fs de -26,50976. Para o clado 20J foi observado 260 sítios com substituições (S), com valores de Tajima' s D e FS de Fu de -2,75488 e -26,27752 respectivamente, o 20B obteve valores similares aos dois outros clados apresentados anteriormente, com Tajima's D de -2,63846 e Fs de -26,57499. Por fim os clados 20A e 20I apresentaram valores próximos também, valores de Tajima's D bem próximo dos -0,75 mas com Fs um pouco diferente, com 20A apresentando Fs de -3,13549 e 20I com -1,95684, lembrando que ambos os teste foram feitos com 10.000 simulações (Tabela 6).

Tabela 6 - Valores dos testes de neutralidades e os valores de p (significância) respectivos.

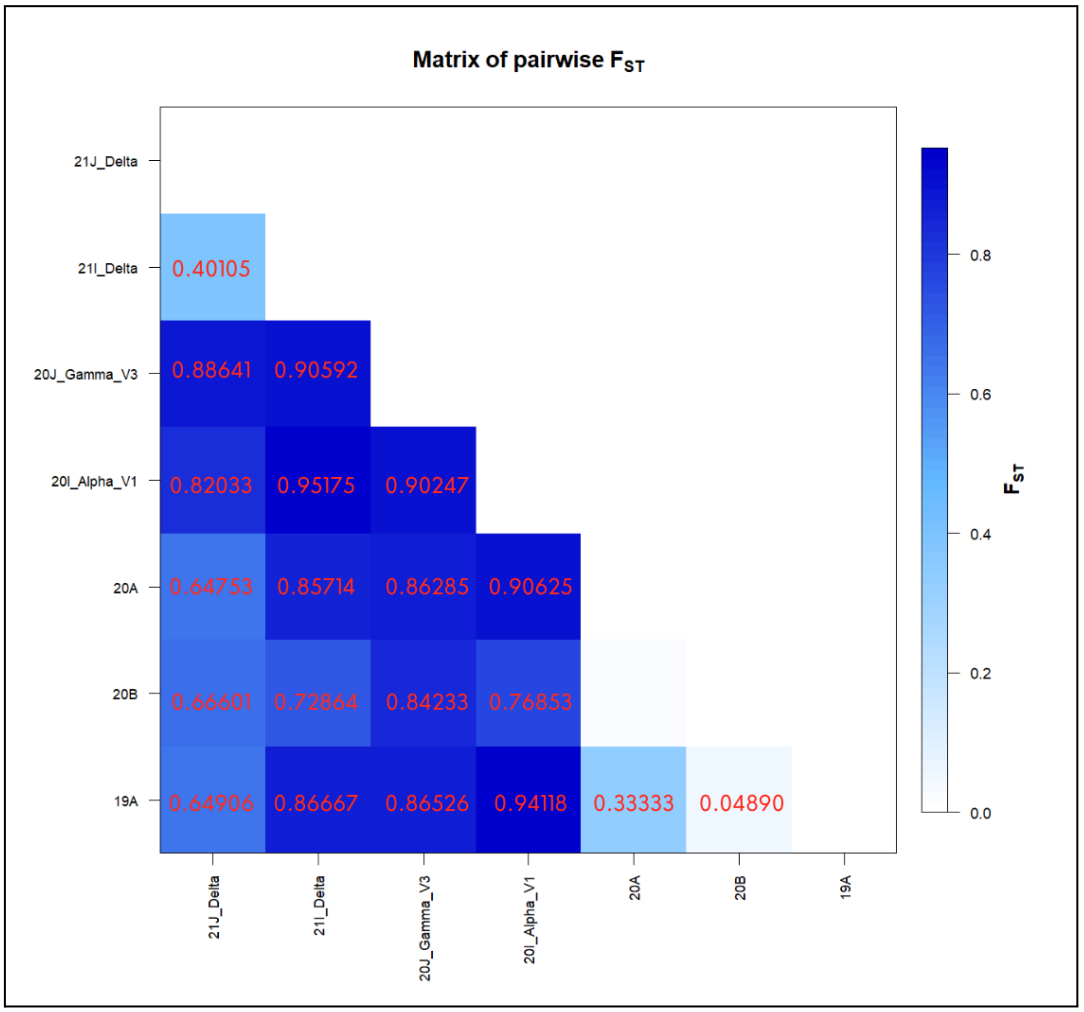
Testes de Neutralidade						
Estatísticas	21J	21I	20J	20I	20B	20B
Tajima's D	-2.43189	0.00	-2.75488	-0.61237	-0.70990	-2.63846
Valor de p de Tajima	0.00	1.00	0.00	0.37790	0.27930	0.00
FS	-26.50976	0.00	-26.27752	-1.95684	-3.13549	-26.57499
Valor de p de FS	0.00	0.25070	0.00010	0.00370	0.00	0.00

Fonte: O autor (2022)

O Fst produzido para observar a distância genética entre os clados apontou que os clados mais distantes geneticamente foram o 21I, 20J e 20I que apresentam grau de quase 0.8 quando comparado aos outros clados. Já 20B e 21J são os que apresentaram maior diferença genética quando comparados aos outros clados (Figura 5). Quando foi analisado o número médio de diferenças de pares de Nei's entre os clados e dentro deles, observou-se que 20J foi considerado o mais distante entre os clados, com valores superiores a 15, a segunda com maior diferença foi a 20I que apresentou diferença significativa em relação a 21J, 20J e 21I. O resultado ficou abaixo de 10 quando comparadas entre si, com os valores da média corrigida ou *Nei's distance* apresentando resultados semelhantes ao entre os clados. Dentro da mesma análise, mas ao observar a diferença dentro dos clados, foi observado valores quase próximos de 2,0 para os clados 21J, 20J e 20B, seguidos por 21I e 20A com valores abaixo de 1.5 mas com valor significativo de diferença dentro dos

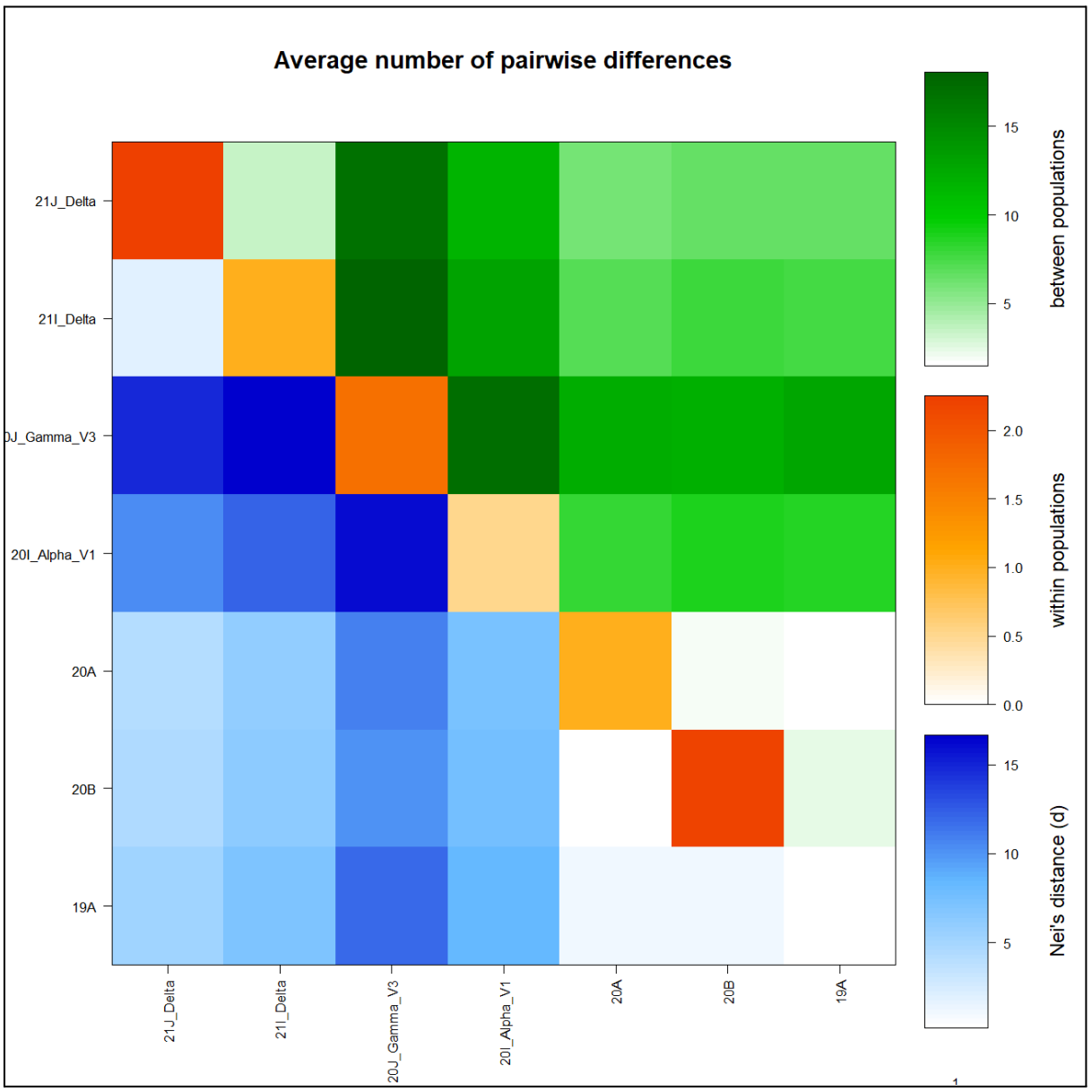
clados, por fim a 20I e 19A obtiveram valores próximos de 0,0 demonstrando pouca diferença dentro do clado (Figura 6).

Figura 5 - Título em português “Matriz de Fst emparelhado”, demonstra o grau de diferença genética entre os clados, valores de 0.0 quando existe pouca diferença genética e valores próximos de 1.0 quando existe um alto grau de diferença genética.



Fonte: O autor (2022).

Figura 6 - Título em português: "Número médio de diferenças emparelhadas". A matriz foi gerada pelo número médio de diferenças de Nei's. A parte superior em verde aponta a média de diferenças encontradas comparando as populações; A parte em azul representa os valores da média corrigida; Por fim, a linha laranja na diagonal aponta o grau de diferença genética encontrada dentro das populações.



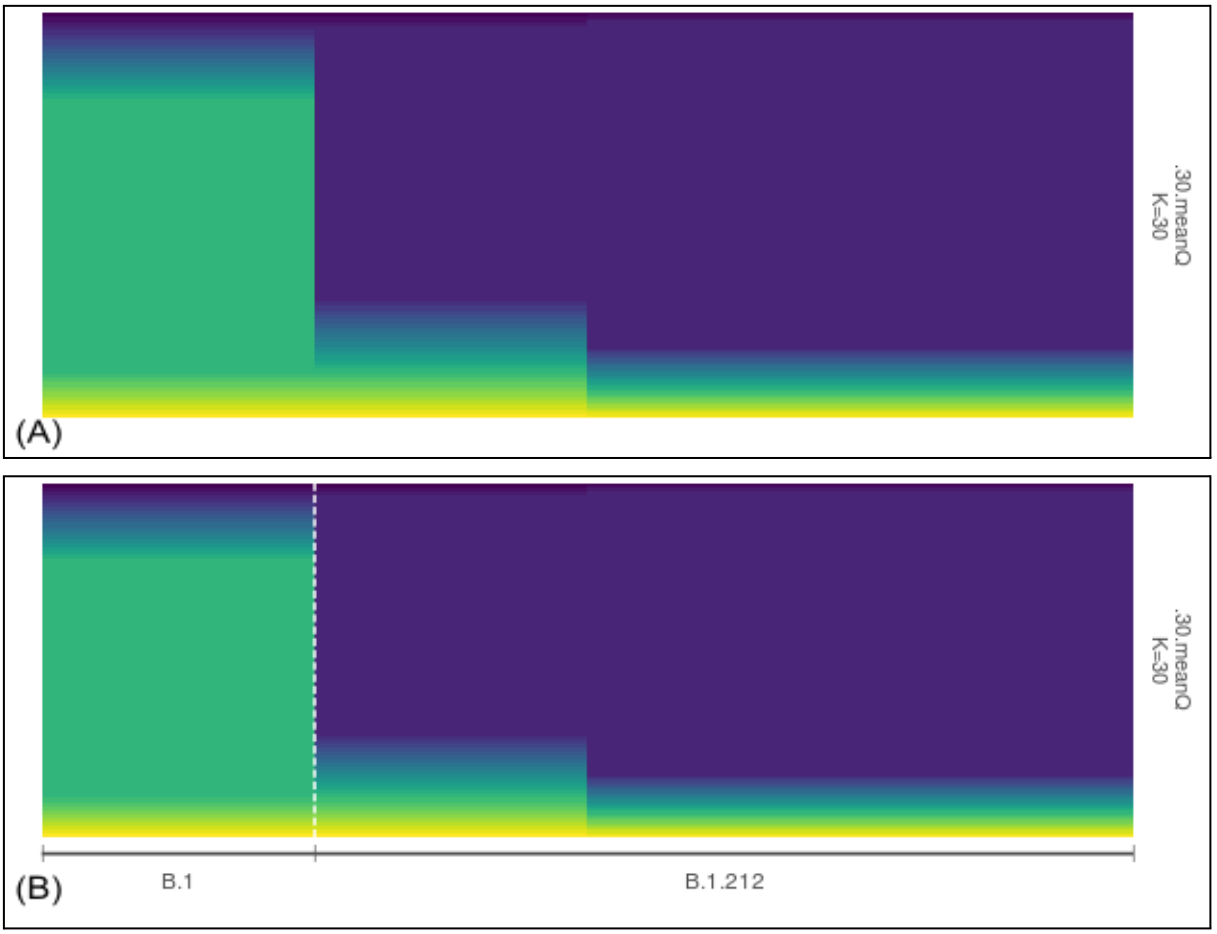
Fonte: O autor (2022).

3.3 Estruturação Populacional

Para se compreender a estruturação dos clados, as sequências referentes ao gene Spike foram separados por clados, e compondo esse conjunto de dados estão presentes somente as frequências dos SNPs. O clado 19A foi retirado da análise pois não continha amostras o suficiente para se observar resultados. Ao se observar o resultado de 20A, foi apresentado como estruturado em dois grupos genéticos

para k igual a 30. O valor de k foi definido pelo próprio programa. Cada linhagem (B.1 e B.1.212) foi estruturada como um grupo genético distinto. (Figura 7).

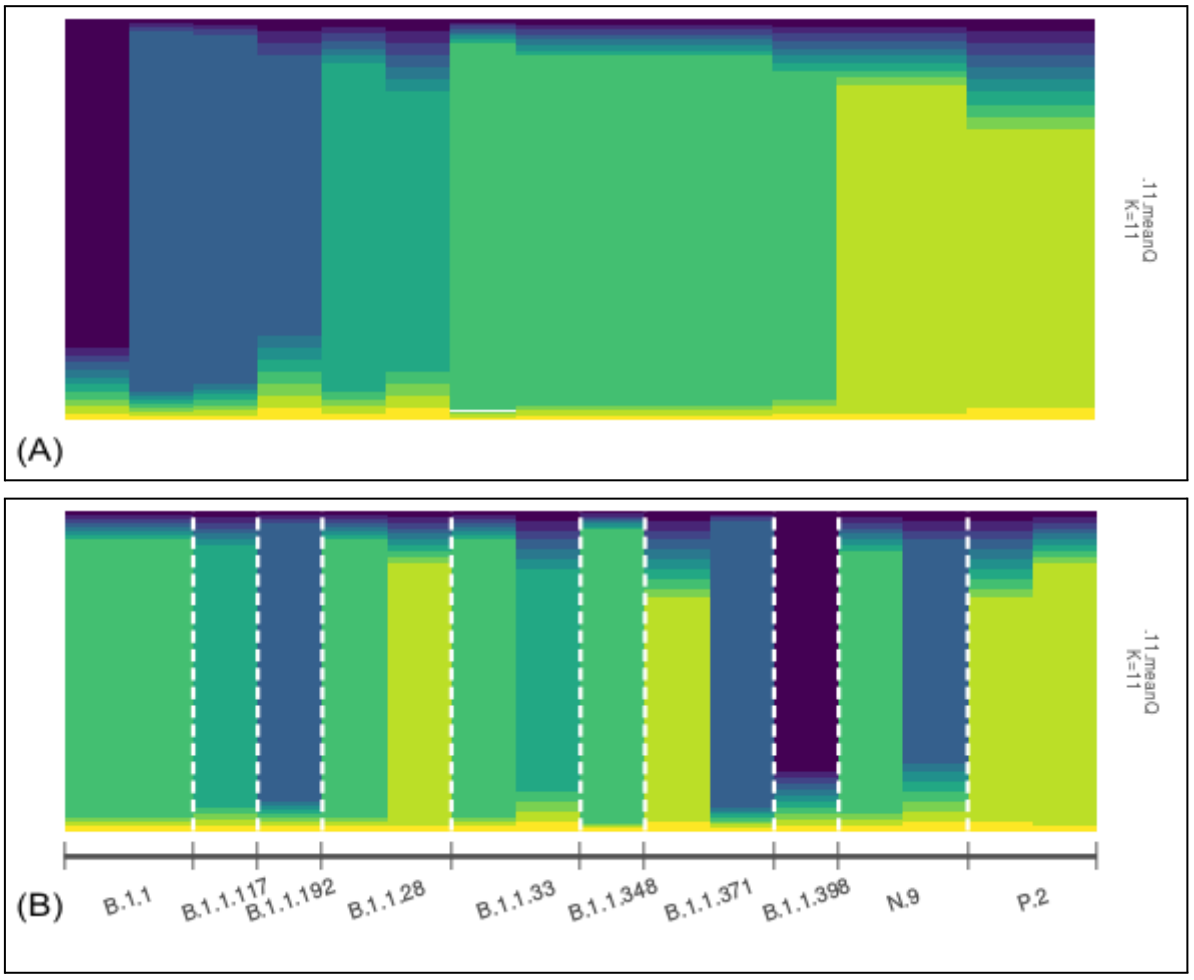
Figura 7 - Estruturação genética encontrada para o clado 20A, na imagem (A) se observa dois grupos genéticos, na imagem (B) a separação com as linhagens presentes na legenda.



Fonte: O autor (2022).

Já para 20B, o software definiu o melhor k sendo 11, onde foi encontrado a maior quantidade de grupos genéticos, equivalente a cinco grupos, onde as linhagens P.2 e B.1.1 são as únicas com a presença de um grupo genético bem estruturado, o restante das linhagens como B.1.1.28, N.9, B.1.1.371 e B.1.1.33 apresentaram dois grupos genéticos dentro da mesma linhagem. As outras linhagens não apresentaram uma quantidade significativa de amostras para se observar suas diferenças (Figura 8).

Figura 8 - Estruturação genética encontrada para o clado 20B, na imagem (A) se encontra cinco grupos genéticos, na imagem (B) a separação com as linhagens presentes na legenda.



Fonte: O autor (2022).

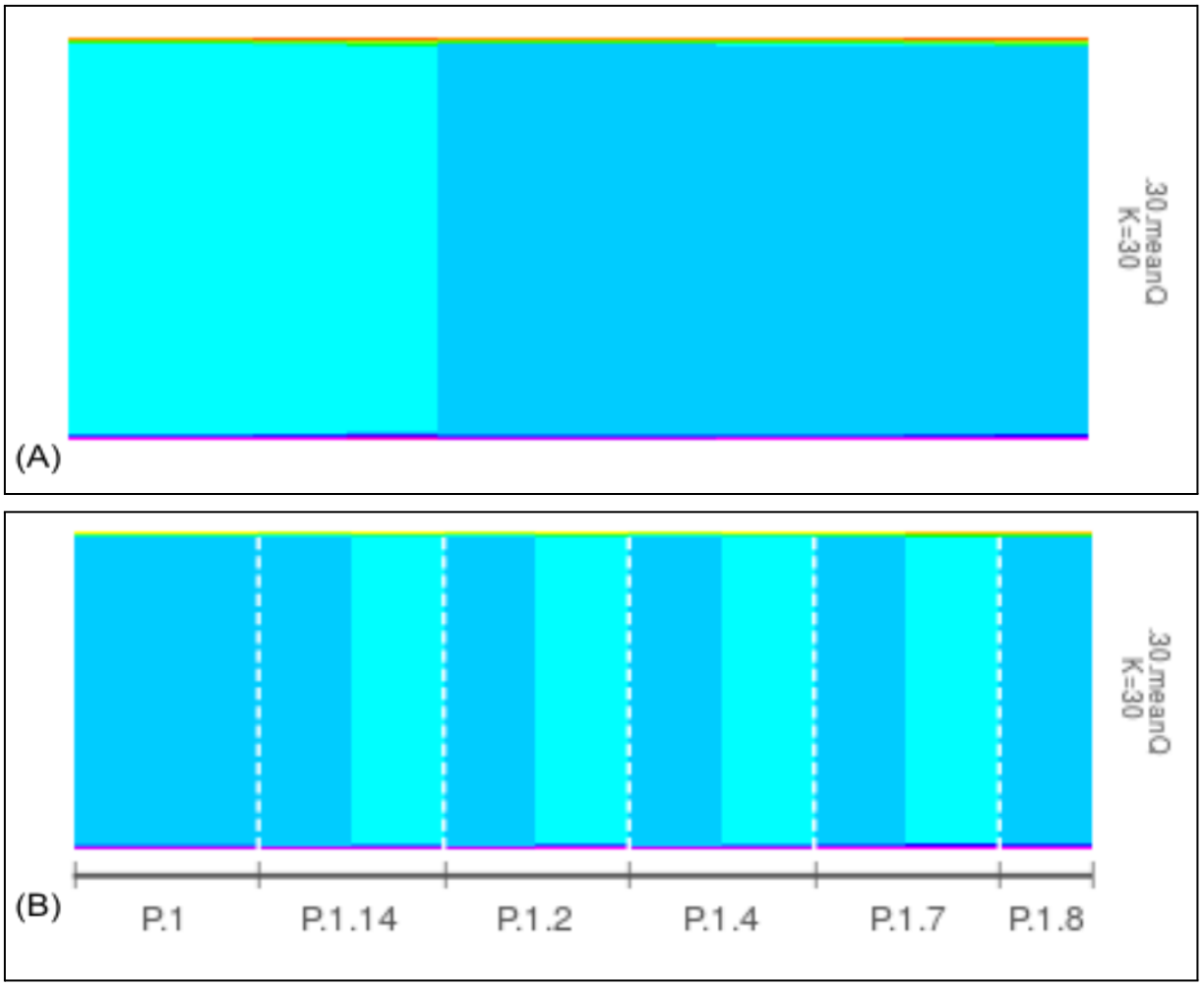
A 20I apresentou um único grupo genético dentro da única linhagem que teve amostra representada a B.1.1.7 (Figura 9).

Figure 9 - Estruturação genética encontrada para o clado 20I, único grupo genético para a única linhagem classificada.

Fonte: O autor (2022).

Foi encontrado em 20J somente dois grupos genéticos dentro das seis linhagens representadas, sendo a linhagem P.1 contendo a presença de um único grupo genético bem estruturado e esse grupo presente em todas as outras linhagens que apresentaram o segundo grupo genético que não aparece em P.1 (Figura 10). O software escolheu o melhor K sendo 30.

Figura 10 - Estrutura gerada para o clado 20J, (A) foi encontrado dois grupos genéticos distintos, (B) como esses grupos estão presentes nas linhagens classificadas.



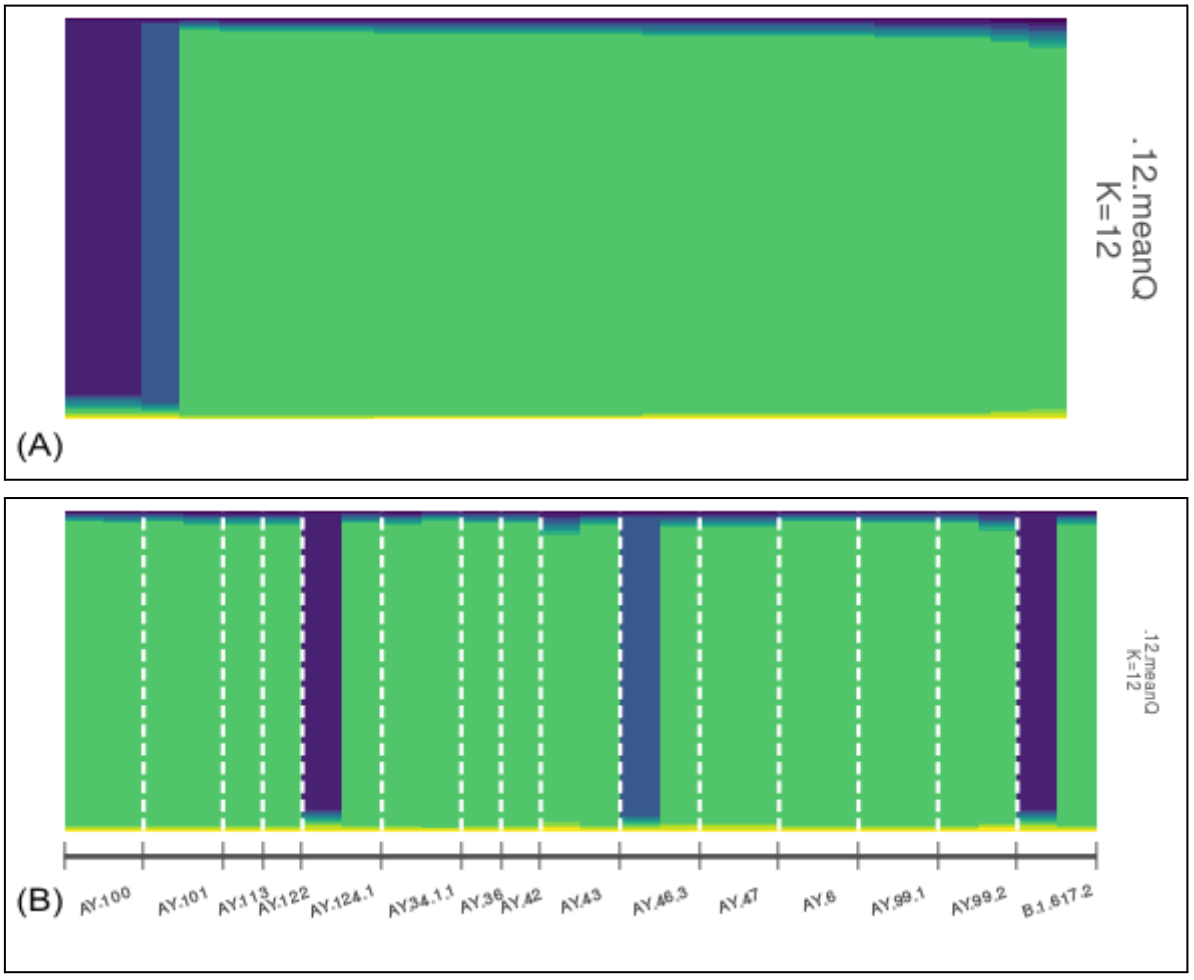
Fonte: O autor (2022).

Na 21I foi observado a formação de um único grupo genético estruturando a única linhagem presente, a AY.47 (Figura 11). Por fim no clado 21J, com (k=12) foi encontrado três grandes grupos genéticos bem definidos pelos SNPs dentro das 15 linhagens representadas, com a maioria das linhagens sendo estruturadas majoritariamente por um grupo genético e três linhagens, AY.123.1, AY.46.3 e B.1.617.2, tendo presença de dois grupos genéticos diferentes (Figura 12). Todas as rodadas de análises para os clados, foram testados k para 30 possíveis grupos genéticos.

Figura 11 - Estrutura genética gerada para o clado 21I, foi encontrado só um grupo genético para a única linhagem classificada.

Fonte: O autor (2022).

Figura 12 - Estrutura genética para o clado 21J, (A) foi encontrado três grupos genéticos, com o verde mais claro sendo predominante em todo o clado, (B) como os grupos estão dispostos em relação às linhagens classificadas.

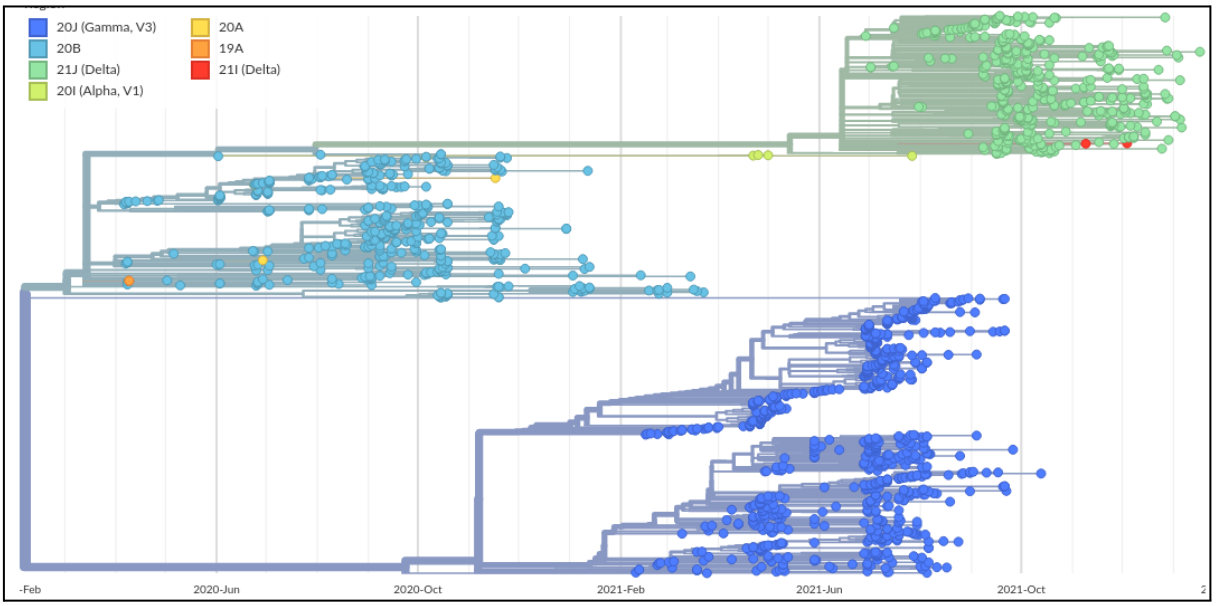


Fonte: O autor (2022).

3.4 Análise Filogenética

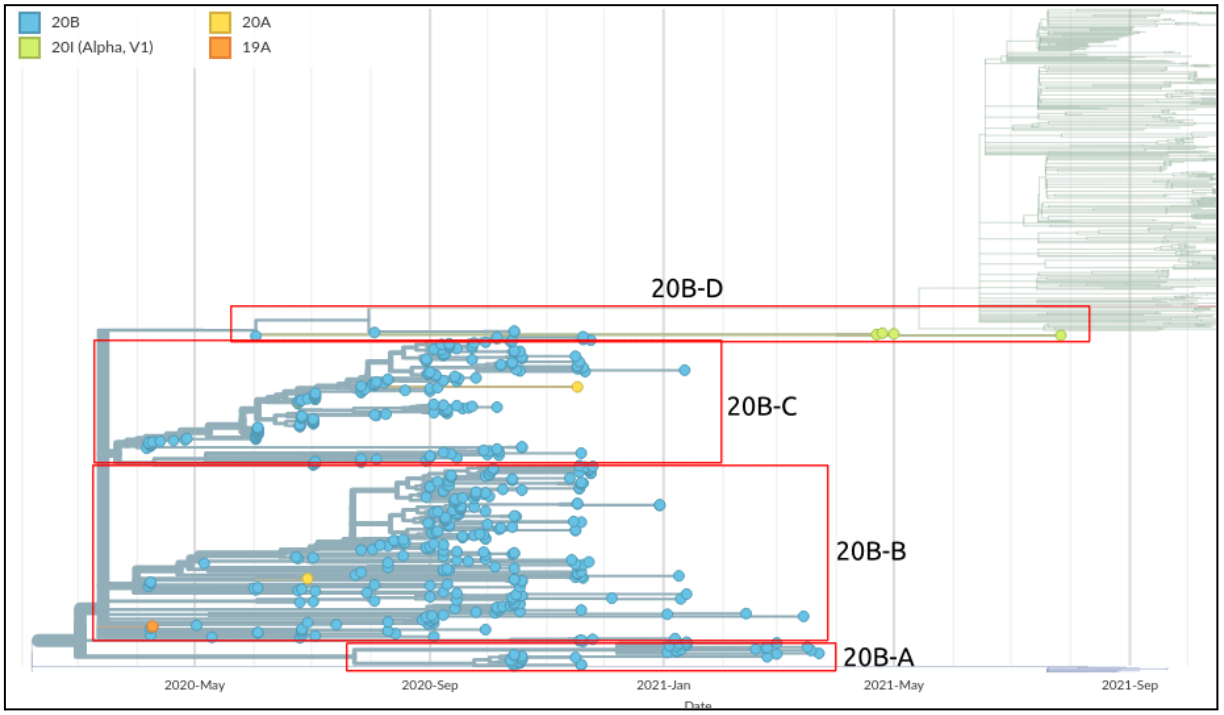
A árvore filogenética, na visão dos clados, tem três grandes ramos, compostos por 20B, 21J e 20J (Figura 13). 20B foi o primeiro clado a ter amostras encontradas dentro da população de Pernambuco, no final de abril de 2020, dentro do seu ramo se encontra os clados 20A, 19A e 20I. A fim de se visualizar melhor os resultados, 20B foi subdividido em quatro sub-ramos, que nós nomeamos como 20B-A, que é composto pela linhagem P.2, 20B-B, que tem majoritariamente a presença da linhagem B.1.1.28, 20B-C que apresenta em sua maioria a linhagem B.1.1, e 20B-D que são um pequeno sub-ramo de quatro amostras de linhagem B.1.1.7. O clado 19A e uma amostra de 20A está presente em 20B-B, o restante das amostras de 20A está presente em 20B-C e o clado 20I está exclusivamente no sub-ramo 20B-D (Figura 14).

Figura 13 - Todos os clados amostrados no conjunto de dados, separados em três grandes ramos, 20J sendo azul escuro, 20B azul mais claro e 21J em verde, os outros clados estão dentro dos ramos desses três grandes conjuntos.



Fonte: O autor (2022).

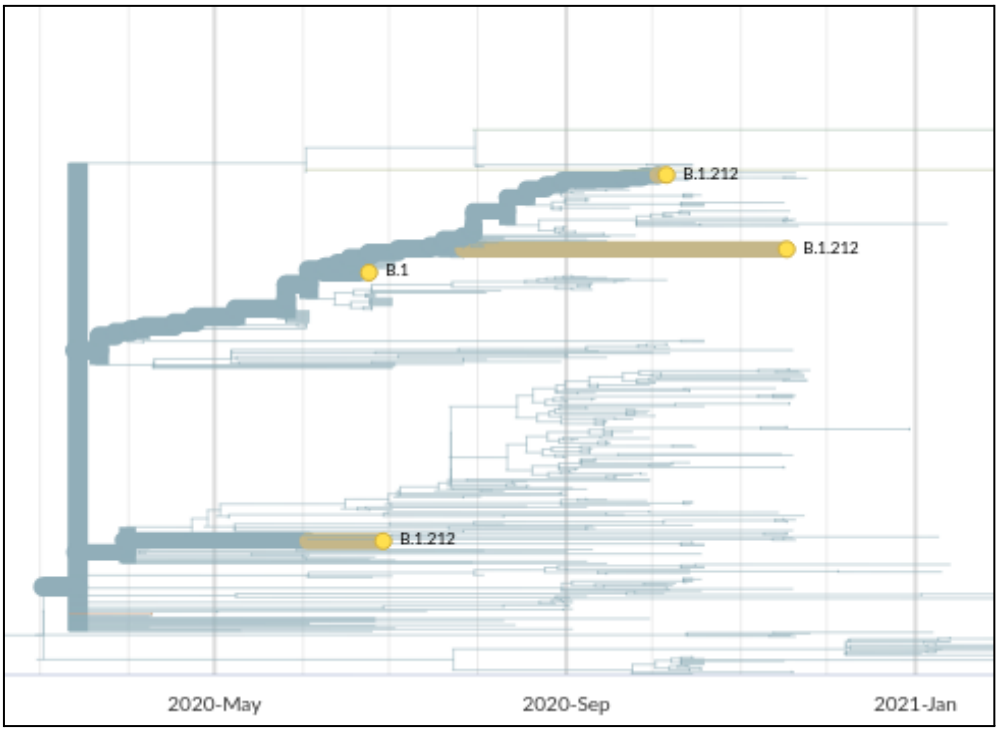
Figura 14 - Primeiro grande ramo a ter amostras depositadas, com a presença de quatro clados filogeneticamente ligados, 20B em azul claro, 20I em verde, 20A em amarelo e 19A em laranja. O ramos foi subdividido em quatro para se compreender melhor.



Fonte: O autor (2022).

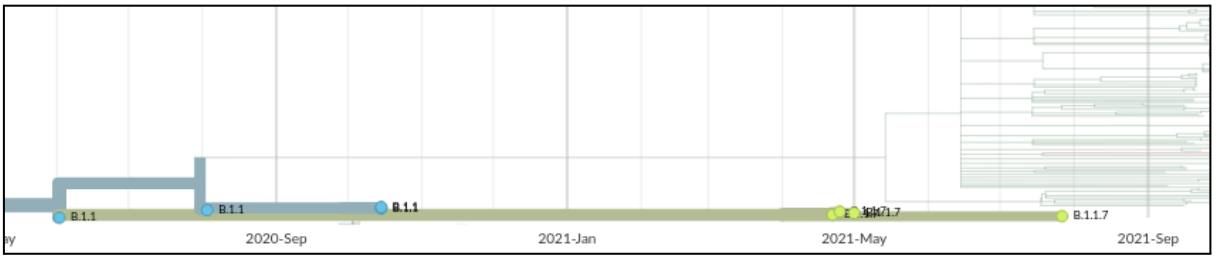
As amostras dos sub-ramos 20B-B e 20B-C foram as pioneiras no estado e a 20B-B é intimamente relacionada a amostra de 19A que pertence a linhagem B. Uma amostra de 20A está mais próxima geneticamente do sub-ramo 20B-B enquanto as outras amostra de 20A estão mais próximas filogeneticamente do sub-ramo 20B-C. A amostra encontrada em 20B-B pertence a linhagem B.1.212 enquanto que a mais relacionada a 20B-C são das linhagens B.1.212 e B.1 (Figura 15). Por fim o sub-ramo 20B-D, presente em junho de 2020, partilha o ancestral comum com o clado 21J e 20I, esse último clado sendo todas suas amostras compostas pela linhagem B.1.1.7 e sendo um grupo monofilético bem resolvido (Figura 16).

Figura 15 - A localização das amostras do clado 20A na filogenia, três amostras filogeneticamente próximas de 20B-C e uma de 20B-B, com presença de duas linhagens B.1.212 e B.1.



Fonte: O autor (2022).

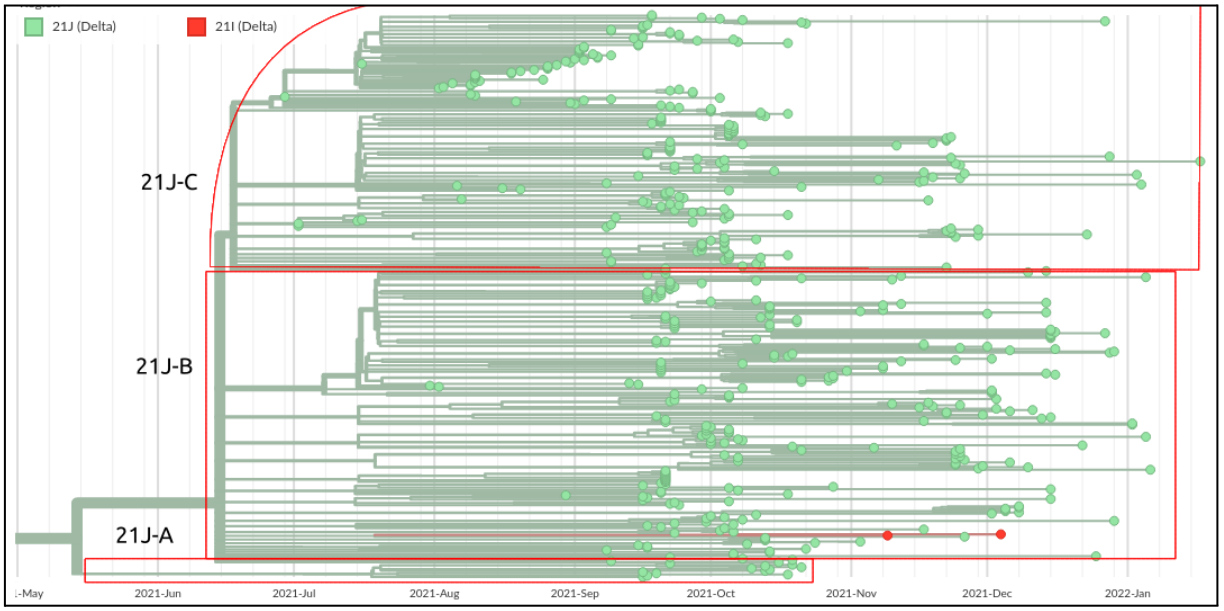
Figura 16 - Sub-ramo 20B-D que se mostrou ancestral do clado 20I, que tem presença só da linhagem B.1.1.7



Fonte: O autor (2022).

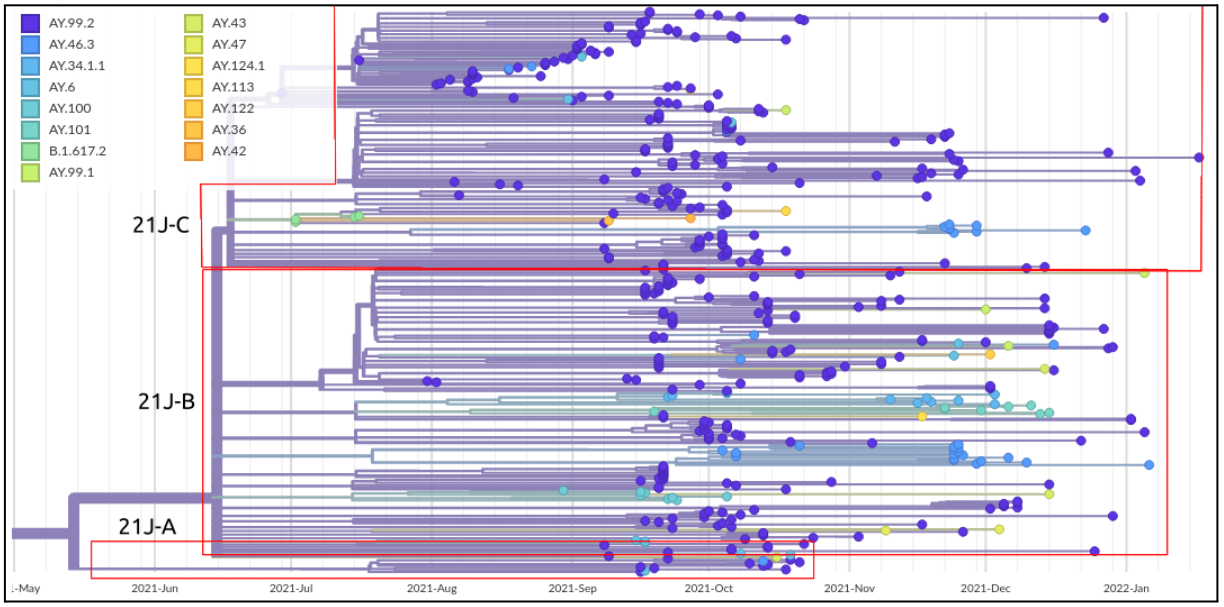
O ramo do clado 21J se apresenta como o mais derivado dos ramos com presença de ancestralidade no estado entre maio e junho de 2021 e com presença até início de 2022, fizemos sua divisão em três sub-ramos, o mais basal 21J-A, composto principalmente pela linhagem AY.99.2, o sub-ramo 21J-B que na nossa filogenia ficou mais próximos geneticamente das amostras do clado 21I, esse último nos nossos dados é composto pela linhagem AY.47 (Figura 17), o sub-ramo 21J-C é o último sub-ramo a divergir do resto do clado, os três sub-ramos de 21J é composto sua maior parte pela linhagem AY.99.2. No 21J-C não se relaciona filogeneticamente com nenhum outro clado e apresenta assim como no restante do clado 21J a linhagem AY.99.2 como a mais presente (Figura 18).

Figura 17 - O segundo maior ramo da filogenia, com presença de dois clados, maior parte de 21J e duas amostras de 21I, o ramo foi separado em três sub-ramos.



Fonte: O autor (2022).

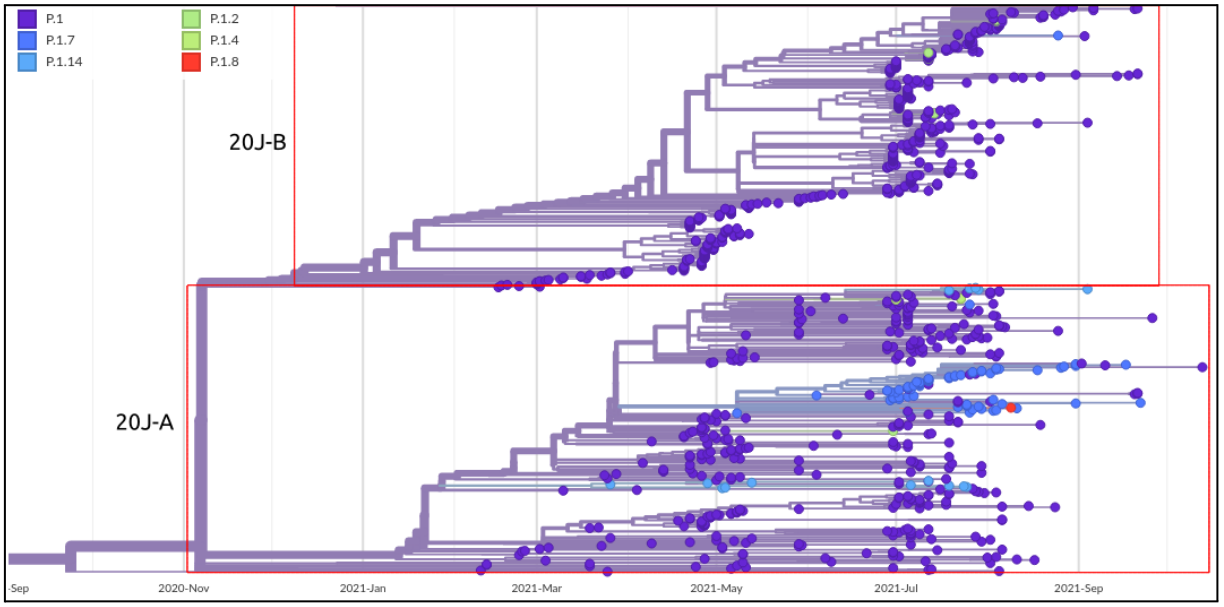
Figura 18 - Os sub-ramos do clado 21J, deixando em evidência a distribuição das linhagens nos três grupos, sendo AY.99.2 em azul escuro a predominante.



Fonte: O autor (2022).

O ramo mais basal encontrado no estado é formado somente pelo clado 20J, que tem sua ancestralidade no estado desde do primeiro ano de pandemia, fizemos sua divisão em dois sub-ramos o 20J-A e 20J-B, esse ramo está filogeneticamente mais distante dos outros clados, a 20J-A é composto principalmente por todas as linhagens presentes no clado 20J, mas sendo majoritariamente de P.1, o 20J-B apresenta três linhagens, em sua maioria sendo P.1 e com amostras de P.1.7 e P.1.2, ambos os sub-ramos compartilham as linhagens P.1, P.1.7 e P.1.2 (Figura 19).

Figura 19 - O ramo mais basal da filogenia, composto exclusivamente pelo clado 20J, com os sub-ramos 20J-A e 20J-B. A linhagem P.1 (roxa), P.1.7 (azul escuro), P.1.14 (azul claro), P.1.2 (verde escuro), P.1.4 (verde claro) e P.1.8 (vermelho).



Fonte: O autor (2022).

5. DISCUSSÃO

Dos clados encontrados no estudo, os três principais foram 20J, 21J e 20B, e dentre as linhagens, P.1, AY.99.2, B.1.1 e B.1.1.28. Ambos tiveram predominância em épocas diferentes no estado de Pernambuco ao longo desses anos de pandemia. Como confirmado por outros estudos, a linhagem B.1.1 foi introduzida no Estado por meio de amostras da Europa ou vindas de outros estados brasileiros com ligação com a Europa, se fixando no Brasil e Pernambuco através da transmissão comunitária (NASCIMENTO et al., 2020). Essa informação foi corroborada com a filogenia do gene spike realizada neste trabalho, onde as primeiras amostras são dessa linhagem e são geneticamente próximas da B.1 (20A) que é sua ancestral.

Após essa introdução, com passar dos meses surgiu a linhagem derivada B.1.1.28, que partilha as mutações K484E, Y655H, I1027T com B.1 mas que apresenta, também, outras mutações na spike: P812S, S939F e V1176F, com a B.1.1.28 apresentando uma alta frequência no Brasil em relação ao mundo (NAVECA et al., 2021), as mutações citadas estão no Apêndice A. Algumas amostras presentes no nosso conjunto de dados são tão antigas quanto a primeira notificação dela que é de 03/05/2020, podendo ser visto também na filogenia, que ela teve seu aparecimento, em Pernambuco, no início de maio de 2020, demonstrando que o estado de Pernambuco foi um grande foco da introdução dessa linhagem em alguns estados do nordeste como Rio Grande do Norte (PEREIRA et al., 2022). Por conseguinte, a linhagem P.1 (Clado 20J) e P.2 (Clado 20B) são derivadas a partir da linhagem B.1.1.28 (FARIA et al., 2021), a presença de P.2 no

estado de Pernambuco, aconteceu por volta de outubro de 2020, com introdução provável decorrendo dos estados como Rio de Janeiro (SILVEIRA et al., 2022).

Já a P.1 teve uma alta taxa de disseminação (Figura 4) sendo predominante no Estado ao longo de 2021. Na nossa filogenia construída com a Spike, a linhagem P.2 (clado 20B) presente em Pernambuco apresenta as mutações Y655H e I1027T, que são mutações que foram detectadas em amostras do Estado desde outubro de 2020 (SILVEIRA et al., 2022) mas que não estão presentes no ramo do clado 20J, que porta as amostras de spike da linhagem P.1. Esse ramo tem as mutações únicas T20N e D138Y. A mutação L18F foi encontrada nos nossos dados e está ligada ao escape imunológico dando as amostras que apresentam essa mutação tem uma valor adaptativo maior em relação aos que não apresentam essa variação (HARVEY et al., 2021).

A linhagem AY. (21J) apresenta forte presença no Brasil (GULARTE et al., 2022) a sua variação, AY.99.2 em Pernambuco começa a ser dominante entre agosto e setembro de 2021, apesar da primeira notificação internacional ter sido no dia 28/02/2021 (LATIF et al., 2021). A primeira amostra de AY.99.2 foi detectada por volta do fim de julho de 2021, mas a filogenia aponta uma possível presença desde junho de 2021. Essa linhagem compartilha as mutações na spike Y655H e I1027T com a linhagem P.2. O ramo ancestral de AY.99.2 em Pernambuco encontrado no sub-ramo 21J-A não apresenta a mutação da spike L452R, o que o separa do sub-ramo 21J-B e 21J-C. Esses últimos sub-ramos, por sua vez, são separados principalmente pela mutação D950N. Das mutações que possuem relação com maior gravidade da infecção, foi encontrado nos nossos dados a mutação D614G (FLORES-ALANIS et al., 2021)

Quando os resultados provenientes da análise de diversidade genética foram compilados, observou-se que o clado 20J, mesmo com a maior quantidade de amostras, obteve valores de diversidade nucleotídica inferior quando comparado com outros clados com quantidade significativa de sequências, como 20B e 21J. Nos nossos resultados isso aconteceu já que dentro de 20J foi encontrado majoritariamente a linhagem P.1. A diversidade de P.1 é caracterizada por alguns fatores, como as mutação características dessa linhagem ser as mutações no gene spike, como N501Y, K417T e E484K, essas mutações aumentando o nível de escape imunológico e transmissibilidade (FARIA et al., 2021), assim como quantidade de países que ela circulou que foram quase 40, logo acumulando mais diferenças ao infectar pessoas de diferentes regiões (O'TOOLE et al., 2021). Além disso, como já foi dito, pelo fato do P1 ser a linhagem mais representa no nosso conjunto de linhagens, já apontava uma possibilidade de apresentar alta diversidade.

A quantidade de haplótipos encontrada no Estado quase supera o apontado no segundo ano de pandemia para todo o território brasileiro, correspondente a 300 haplótipos, demonstrando como a Spike vai adquirindo mudanças em poucos meses (BUITRAGO et al., 2021). O número médio de diferenças nucleotídicas (K) assim como a quantidade de mutações compartilhadas entre os três maiores clados 21J, 20J e 20B apontam essa diferença encontrada na Spike (Tabela 5), observado na reconstrução filogenética (Figura 13), onde os clados 20J e 20B apresentam uma K

de 6,376 mas são os dois grupos que mais partilham mutações, equivalente a 27 mutações. Isso acontece já que dentro de 20B, como já foi mencionado antes, se encontram as linhagens B.1.B.1.1.28 e P.2 que tem sua ancestralidade muito bem definida com P.1 presente no grupo 20J (Tabela 1). Assim como já vinha sendo observado por estudos anteriores, uma maior presença de sítios parcimoniosamente informativos, que são sítios variantes encontrados em mais de uma sequências estão em maior quantidade em comparação aos singletons, que são encontrados em uma única sequência, demonstrando que existe maior variabilidade partilhada do que em sequências únicas, contrastando com o que foi relatado para toda América do Sul no ano de 2021 (BUITRAGO et al., 2021).

O clado 21J obteve um K maior que 8 em comparação a 20J, o que se reflete na filogenia (Figura 13) mas mesmo assim compartilha mais de 20 mutações, confirmada na análise filogenética onde foi encontrada algumas mutações específicas que eles partilham, como L425R. Quando comparada com 20B, o valor de K é menor, próximo de 3, compartilham 15 mutações presentes na spike.

O clado 20A em comparação com 20B, tem um dos valores mais baixos para média de diferença nucleotídica, sendo 1,515, onde a spike presente em 20A e nas suas linhagens B.1 e B.1.212, demonstrou na filogenia (Figura 15) uma similaridade filogenética muito grande compartilhando uma mutação com o restante do clado 20B. Quando comparado 20A com outros clados, o valor de K, diferença média nucleotídica chegando no valor de 8,000, demonstrando uma grande distância na diversidade nucleotídica do gene em relação aos outros clados.

De acordo com os resultados dos testes de neutralidade (Tabela 6), foi observado que os clados com maior representatividade 21J, 20B e 20J estão com valores muito próximos tanto no teste de Tajima quanto no Fu's FS, assim demonstrando que estão ou estiveram em expansão ao longo dos quase três anos de pandemia, como já foi apontado por outros estudos (LIU et al., 2020; YU et al., 2020; PANDIT et al., 2021), o restante dos clados não obtiveram valores ou muito próximos de 0, demonstrando que não se estão se expandindo ou não se expandiram nos anos anteriores.

Quando se analisou os resultados de Fst para observar a diferença genética do gene spike entre os clados, o 20J apresentou os maiores valores quando comparado aos outros (Figura 5), o que ficou mais claro na filogenia, quando observado que a spike, presente no clado, é considerada a mais basal do conjunto de dados. A 21J obteve valores baixos de FST, no caso maior similaridade genética, com o clado 21I, que dentro da filogenia se apresenta com alta similaridade da spike, já o clado 20B mostrou valores próximos às de 0, sendo altamente próxima geneticamente de 19A e 20A, a similaridade evolutiva foi demonstrada também na filogenia. Os resultados da diferença entre os clados também foi obtido nos teste da diferença de Nei's (Figura 6), que foi observado resultados similares ao FST, com por exemplo 20J sendo bem distante geneticamente dos outros clados e tendo um pouco de semelhança com 20B, os dados corrigidos de Nei's presentes em azul (Figura 6), asseguram também essas diferenças que ficaram evidentes nos testes de diversidade da spike e demonstrado na filogenia gerada.

Dentro da análise de Nei's, a linha diagonal apontou uma diferença significativa da spike dentro de alguns clados, Isso já era esperado, pois sabe-se que os clados 21J e 20B apresentam uma quantidade significativa de linhagens. Os valores do diagrama de Nei's ficou um pouco mais claro em 20J, pois mesmo tendo uma grande quantidade de amostras e boa quantidade de linhagens dentro, a dominância de P.1 dentro do clado, fez esse valor ser menor.

Com a estruturação dos clados, foi possível observar de que forma o gene Spike está estruturado geneticamente nas linhagens. O software utilizado apresentou dificuldade em analisar os dados com a quantidade de SNPs presentes, principalmente para entender a hibridização do conjunto de dados, e por isso não se pode tirar conclusões sobre as linhagens que apresentaram uma grande quantidade de dados.

O clado 20A, apresentou uma estruturação em que a linhagem B.1 é diferente da população de B.1.212, mesmo ocorrendo um pouco de hibridização em um indivíduo, o que é esperado já que B.1 é o ancestral direto dela, porém tem uma composição genética de SNPs bem distinta do seu ancestral. Entretanto, por falta de amostras para representar as linhagens dentro de 20A, não é possível ter um alto grau de certeza sobre esse resultado. A 20B, que apresenta dez linhagens, se mostrou dividida em cinco grupos genéticos, com P.2 e B.1.1 bem estruturado em uma único grupo distinto. As outras linhagens apresentam mais de dois grupos genéticos dentro da linhagem, demonstrando a necessidade de mais dados para ter certeza que as linhagens já se tornaram duas populações distintas.

O 20I apresentou nenhuma diferença genética entre as amostras, demonstrando que todos os indivíduos são do mesmo grupo genético, respaldado por um valor muito baixo de divergência dentro do clado visto na Figura 6, e quando observado como as amostras que estão na filogenia, há uma similaridade muito alta, como visto na Figura 16. A 20J, foi caracterizado como portando duas populações distintas, sendo formado pela linhagem P.1 que foi a única pertencendo seus indivíduos a uma única população. O restante se apresentou hibridizado entre duas e a linhagem P.1.8 não tinha amostras o suficiente para indicar alguma diferenciação, enquanto P.1.2 está caracterizado como duas populações diferentes assim como P.1.14. Dessa forma, tem-se a suspeita, como visto na separação dos sub-ramos 20J-A e 20J-B, que estas estão se diferenciando em populações distintas dentro das linhagens (Figura 19).

Dentro do ramo de 21J presente na filogenia, foi apontado que as amostras da linhagem AY.124.1 não são filogeneticamente próximas pois se apresentam no sub-ramos 21J-B e 21J-C, e podem estar se tornando tão distantes que não possam ser mais consideradas o mesmo clado. Essa situação em específico é respaldada por uma das amostras dessas linhagens apresentar as mutações G142D e D950N, enquanto a outra não apresenta. Além disso, a estruturação populacional também apontou essa situação (Figura 12). Já a linhagem AY.46.3 tem indivíduos do sub-ramo 21J-C, que tem a mutação V1264L, enquanto as amostras que pertencem ao sub-ramo 21J-B não apresentam. Por fim, a linhagem AY.34.1.1 que se agrupou

de forma monofilética na filogenia (Figura 18) se comporta como parte da mesma população genética.

Dentro do estudo presente foram obtidos diversos resultados que acompanham como foi a diversidade dos clados e a relação evolutiva deles e das linhagens dentro do estado de Pernambuco, mas pelos filtros utilizados ficaram de fora algumas linhagens de grande interesse epidemiológico, fazendo assim um necessidade de um estudo futuro abordando esses novos pontos para corroborar ainda mais ou alterar os resultados que foram discutidos.

6. CONCLUSÃO

O presente trabalho mostrou que os principais clados e suas linhagens, que tiveram maior impacto nos três anos de pandemia em Pernambuco, foram os 20B, portando a linhagem B.1.1 mais representativa no seu clado, 20J composta majoritariamente por P.1 e 21J que apresenta a linhagem AY.99.2 como maioria presente no clado.

Os clados e linhagens que são constituídas por um número maior de amostras, normalmente apresentam dominância em épocas distintas e com baixa sobreposição, como o que foi encontrado neste estudo em P.1 (20J), AY.99.2 (21J) e B.1.1 (20B). Os valores do teste de neutralidade são o reflexo dessa dominância ao apontar que os clados que portam essas linhagens estavam em expansão. Em conjunto com o resultados do Fst, que apontam um elevado valor de diferença genética dos clados que comportam essas linhagens quando comparados entre si, a diversidade dessas linhagens e suas diferenças é demonstrada na quantidade de mutações que são encontradas. O isolamento na dominância pode ser apontado pela quantidade de mutações que essas linhagens não partilham entre si, que na P.1 chega próximo a 35 substituições, na B.1.1 com 24 modificações e AY.99.2 com 20 dessas mutações, essas características gerais elucidam como ocorreu essas dominâncias em faixa de tempo distintas.

A estruturação populacional apontou que algumas linhagens como N.9, B.1.1.28, P.1.2, P.1.14 entre outras, estão apresentando um estruturação genética de mais de uma população possível dentro da linhagem em relação aos SNPs encontrados, apontando uma diferença individual nos indivíduos encontrados que podem estar se tornando outras linhagens ou que a classificação dessas linhagens está ficando inconsistente pelos algoritmos utilizados.

REFERÊNCIAS

ALM, Erik et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Eurosurveillance, v. 25, n. 32, p. 2001410, 2020.

BONI, Maciej F. et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nature microbiology, v. 5, n. 11, p. 1408-1417, 2020.

BRASIL. Ministério da Saúde. Infecção Humana pelo Novo Coronavírus (2019-nCoV). COVID-19: boletim epidemiológico, Brasília, n. 1, jan. 2020. Disponível em:https://www.saude.gov.br/images/pdf/2020/fevereiro/04/Boletim-epidemiologico-SVS-04fev20.pdf. Acesso em: 06 de jun. 2022. (BRASIL, 2020b).

BROWN, Joseph; PIRRUNG, Meg; MCCUE, Lee Ann. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics, v. 33, n. 19, p. 3137-3139, 2017.

BUITRAGO, Sindy P.; GARZÓN-OSPINA, Diego. Genetic diversity of SARS-CoV-2 in South America: demographic history and structuration signals. Archives of virology, v. 166, n. 12, p. 3357-3371, 2021.

CALLAWAY, Ewen. Time to use the p-word? Coronavirus enters dangerous new phase. Nature (Lond.), 2020.

CHAU, Steven WH et al. History for some or lesson for all? A systematic review and meta-analysis on the immediate and long-term mental health impact of the 2002–2003 Severe Acute Respiratory Syndrome (SARS) outbreak. BMC Public Health, v. 21, n. 1, p. 1-23, 2021.

CHERIAN, Sarah et al. SARS-CoV-2 spike mutations, L452R, T478K, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. Microorganisms, v. 9, n. 7, p. 1542, 2021.

COVID, EDIÇÃO ESPECIAL. Principais variantes do SARS-CoV-2 notificadas no Brasil. RBAC, v. 53, n. 2, p. 109-116, 2021.

CORMAN, Victor M. et al. Hosts and sources of endemic human coronaviruses. Advances in virus research, v. 100, p. 163-188, 2018.

DECARO, Nicola; BUONAVOGLIA, Canio. An update on canine coronaviruses: viral evolution and pathobiology. Veterinary microbiology, v. 132, n. 3-4, p. 221-234, 2008.

ELBE, S.; BUCKLAND-MERRETT, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall. 2017; 1 (1): 33-46.

EXCOFFIER, Laurent; LISCHER, Heidi EL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular ecology resources, v. 10, n. 3, p. 564-567, 2010.

FANG, Bin et al. Genome-wide data inferring the evolution and population demography of the novel pneumonia coronavirus (SARS-CoV-2). BioRxiv, 2020.

FARIA, Nuno R. et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. Virological, v. 372, p. 815-821, 2021.

FLORES-ALANIS, Alejandro et al. Molecular epidemiology surveillance of SARS-CoV-2: mutations and genetic diversity one year after emerging. Pathogens, v. 10, n. 2, p. 184, 2021.

FRANCIS, Roy M. pophelper: an R package and web app to analyse and visualize population structure. Molecular ecology resources, v. 17, n. 1, p. 27-32, 2017.

GIOVANETTI, M. et al. Genomic epidemiology reveals how restriction measures shaped the SARS-CoV-2 epidemic in Brazil. medRxiv, 2021.

GULARTE, Juliana Schons et al. Early introduction, dispersal and evolution of Delta SARS-CoV-2 in Southern Brazil, late predominance of AY. 99.2 and AY. 101 related lineages. Virus Research, v. 311, p. 198702, 2022.

HADFIELD, James et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, v. 34, n. 23, p. 4121-4123, 2018.

HARVEY, William T. et al. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology, v. 19, n. 7, p. 409-424, 2021.

HUDDLESTON, John et al. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. Journal of open source software, v. 6, n. 57, 2021.

KATOH, Kazutaka; ROZEWICKI, John; YAMADA, Kazunori D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in bioinformatics, v. 20, n. 4, p. 1160-1166, 2019.

KIRTIPAL, Nikhil; BHARADWAJ, Shiv; KANG, Sang Gu. From SARS to SARS-CoV-2, insights on structure, pathogenicity and immunity aspects of pandemic human coronaviruses. Infection, Genetics and Evolution, v. 85, p. 104502, 2020.

LATIF, A.A., Mullen, J.L., Alkuzweny, M., Tsueng, G., Cano, M., Haag, E., Zhou, J., Zeller, M., Hufbauer, E., Matteson, N., Wu, C., Andersen, K.G., Su, A.I., Gangavarapu, K., Hughes, L.D., Biology, C. for V.S., 2021b. AY.99.2 Lineage Report. https://outbreak.info/situation-reports?pango=AY.99.2&loc=USA&loc=BRA&selected=BRA&overlay=false (accessed 12.09.22).

LARSSON, Anders. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics, v. 30, n. 22, p. 3276-3278, 2014.

LIBRADO, P. and Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451-1452.

LIU, Q. et al. Population Genetics of SARS-CoV-2: Disentangling Effects of Sampling Bias and Infection Clusters. Genomics, proteomics & bioinformatics, 2020.

LOKMAN, Syed Mohammad et al. Explorando as variações genômicas e proteômicas da glicoproteína de pico SARS-CoV-2: uma abordagem de biologia computacional. Infection, Genetics and Evolution , v. 84, p. 104389, 2020.

LU, Roujian et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The lancet, v. 395, n. 10224, p. 565-574, 2020.

MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, v. 17, n. 1, p. 10-12, 2011.

MICHALAKIS, Y. and Excoffier, L. , 1996 A generic estimation of population subdivision using distances between alleles with special reference to microsatellite loci. Genetics 142:1061-1064.

MISHRA, Jitendra; MISHRA, Priya; ARORA, Naveen Kumar. Linkages between environmental issues and zoonotic diseases: with reference to COVID-19 pandemic. Environmental Sustainability, v. 4, n. 3, p. 455-467, 2021.

MOREIRA, L. M. Ciências genômicas: fundamentos e aplicações. Moreira, LM & Varani, AM Plasticidade e fluxo genômico. Ribeirão Preto: Sociedade Brasileira de Genética, v. 1, p. 101-116, 2015.

NAVECA, Felipe et al. SARS-CoV-2 reinfection by the new Variant of Concern (VOC) P. 1 in Amazonas, Brazil. Virological. org, 2021.

NAVECA, Felipe et al. Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein. Virological. org, v. 1, p. 1-8, 2021.

NASCIMENTO, V. A. et al. Genomic and phylogenetic characterisation of an imported case of SARS-CoV-2 in Amazonas State, Brazil. Memórias do Instituto Oswaldo Cruz, v. 115, 2020.

O'TOOLE, Áine et al. Tracking the international spread of SARS-CoV-2 lineages B. 1.1. 7 and B. 1.351/501Y-V2 with grinch. Wellcome Open Research, v. 6, 2021.

PANDIT, B. et al. Association of clade-G SARS-CoV-2 viruses and age with increased mortality rates across 57 countries and India. Infection, Genetics and Evolution, v. 90, p. 104734, 2021.

PEREIRA, Raíssa Liane do Nascimento. Análise filogenética da variante Gamma (p. 1) do SARS-COV-2 circulante no estado do Rio Grande do Norte. 2022. Trabalho de Conclusão de Curso. Universidade Federal do Rio Grande do Norte.

PRITCHARD, Jonathan K.; WEN, Xiaoquan; FALUSH, Daniel. Documentation for structure software: Version 2.3. University of Chicago, Chicago, IL, p. 1-37, 2010.

RAHMAN, Mohammad Mahmudur; HASAN, Maruf; AHMED, Asif. Potential detrimental role of soluble ACE2 in severe COVID‐19 comorbid patients. Reviews in Medical Virology, v. 31, n. 5, p. 1-12, 2021.

RAMBAUT, Andrew et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature microbiology, v. 5, n. 11, p. 1403-1407, 2020.

RAY, Manisha et al. Essential interpretations of bioinformatics in COVID-19 pandemic. Meta Gene, v. 27, p. 100844, 2021.

SILVEIRA, Zildene de Sousa. Diversidade genômica de cepas de SARS-CoV-2 durante a primeira onda da Covid-19 no Estado de Pernambuco, Brasil. 2022. Dissertação de Mestrado. Universidade Federal de Pernambuco.

SINGH, Devika; YI, Soojin V. On the origin and evolution of SARS-CoV-2. Experimental & Molecular Medicine, v. 53, n. 4, p. 537-547, 2021.

TAN, Anthony T. et al. Early induction of functional SARS-CoV-2-specific T cells associates with rapid viral clearance and mild disease in COVID-19 patients. Cell reports, v. 34, n. 6, p. 108728, 2021.

TEAM, R. Core et al. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2020.

UNEP (2016) UNEP frontiers 2016 report: emerging issues of environmental concern united nations environment programme, Nairobi.

VERLI, Hugo. Bioinformática: da biologia à flexibilidade molecular. 2014.

WEN, Feng et al. Identification of the hyper-variable genomic hotspot for the novel coronavirus SARS-CoV-2. Journal of Infection, v. 80, n. 6, p. 671-693, 2020.

WEIR, B.S. and Cockerham, C.C. 1984 Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

XU, Hao et al. High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa. International journal of oral science, v. 12, n. 1, p. 1-5, 2020.

YOUNT, Boyd et al. Severe acute respiratory syndrome coronavirus group-specific open reading frames encode nonessential functions for replication in cell cultures and mice. Journal of virology, v. 79, n. 23, p. 14909-14922, 2005.

YU, W. B. et al. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2/HCoV-19) using whole genomic data. Zoological research, v. 41, n. 3, p. 247, 2020.

APÊNDICE A - MUTAÇÕES ENCONTRADAS NO ESTUDO

Compilação das principais mutações encontradas nos clados e linhagens presentes no presente estudo.

Tabela 6 - Principais mutações encontradas no conjunto de dados trabalhados para o gene Spike apresentando sua presença em relação às linhagens e clados..

Clado	Linhagem	Mutação
20A	B.1/B.1.212	D614G
20B	B.1.1	D614G/ A1078S/ A846V/ D1084Y/ D1260Y/ K1191N/ P681H/ P812S/ Q677H/ V1176F/ S640F/ S810A/ S939F/ V1068D/ V1104L/ K528R/ L938V/ I1216V/ V1264L/ L452Q/ V1228L/ P26L/ P9L/ Q14H/ S12F/ S151T/ S477N/ T470A/ Y145C
20B	B.1.1.117	D614G/ A688V
20B	B.1.1.192	D614G
20B	B.1.1.28	D614G/ D138H/ V1176F/ A1020S/ A694S/ E1144V/ P812L/ Q675H/ P812S/ S939F/ C1235F/ G1219V/ I1216V/ E583D/ V1133I/ L5F/ Q14H/ S151I/ G1219V/ S254F/ S255F/ A1070V/ S98F/ W258L
20B	B.1.1.33	D614G/ A1020S/ A845S/ D1153Y/ P812L/ D80Y/ D88H/ G142V/ V1128I/ H245Y/ I203V/ Q677H/ L5F/ S640F/ P25L/ R408S/ S939F/ T250A/ N1173D
20B	B.1.1.348	D614G/V1176F
20B	B.1.1.371	D614G/ S221L/ L452R

20B	B.1.1.398	D614G/ E1258D
20B	N.9	A344S/ E484K/ D614G/ Q121K/ L176F/ D796Y
20B	P.2	E484K/ D614G/ P1162S/ V1176F/ H207Q/ T323I
20I (Alpha, V1)	B.1.1.7	N501Y/ A570D/ D614G/ P681H/ T716I/ S982A/ D1118H
20J (Gamma, V3)	P.1	L18F/ P26S/ D138Y/ R190S/ K417T/ E484K/ N501Y/ D614G/ H655Y/ Q675H/ T1027I/ I1114T/ V11176F/ T20N/ P25L/ G181A/ A222S/ A243S/ D1139Y/ H245R/ P681H/ A623S/ L1244F/ A688V/ L938F/ H1159Y/ K1038E/ G1219C/ M1237I/ P1263S/ T678S/ T732A/ T859I/ C1247F/ Q580H/ D1163Y/
20J (Gamma, V3)	P.1.14	L18F/ T20N/ P26S/ D138Y/ L179R/ R190S/ K417T/ E484K/ N501Y/ D614G/ H655Y/ Q675H/ T1027I/ V1176F
20J (Gamma, V3)	P.1.2	L18F/ T20N/ P26S/ D138Y/ L179R/ R190S/ K417T/ E484K/ N501Y/ D614G/ H655Y/ Q675H/ T1027I/ V1176F/ D1163Y/
20J (Gamma, V3)	P.1.4	L18F/ T20N/ P26S/ D138Y/ L179R/ R190S/ K417T/ E484K/ N501Y/ D614G/ H655Y/ Q675H/ T1027I/ V1176F
20J (Gamma, V3)	P.1.7	L18F/ T20N/ P26S/ D138Y/ L179R/ R190S/ K417T/ E484K/ N501Y/ D614G/ H655Y/ Q675H/ T1027I/ V1176F/ P681H/ E654Q/ I666V/
20J (Gamma, V3)	P.1.8	L18F/ T20N/ P26S/

		D138Y/ I203V/ K417T/ T470N/ E484K/ N501Y/ D614G/ H655Y/ P681R/ T1027I/ V1176F/ C1235F
21I (Delta)	AY.47	T19R/ A222V/ V289I/ L452R/ T478K/ Y508H/ D614G/ P681R/ D950N
21J (Delta)	AY.100	T19R/ L452R/ T478K/ D614G/ P681R/ D950N/ N354K/ T95I/ G142D
21J (Delta)	AY.101	T19R/ T95I/ L452R/ T478K/ D614G/ P681R/ N440K/ D950N/
21J (Delta)	AY.113	T19R/ T95I/ L452R/ T478K/ D614G/ P681R/ D950N
21J (Delta)	AY.122	T19R/ L452R/ T478K/ D614G/ P681R
21J (Delta)	AY.124.1	T19R/ T95I/ G142D/ L452R/ T478K/ D614G/ P681R/ D950N
21J (Delta)	AY.34.1.1	T19R/ T95I/ G142D/ L452R/ T478K/ D614G/ Q677H/ P681R/ D950N/ L1265F
21J (Delta)	AY.36	T19R/ L452R/ T478K/ D614G/ P681R/ D950N/ V1104L
21J (Delta)	AY.42	T19R/ T95I/ I128V/ L452R/ T478K/ D614G/ P681R/ D950N
21J (Delta)	AY.43	T19R/ L452R/ T478K/ D614G/ P681R/ D950N/ P812L
21J (Delta)	AY.46.3	T19R/ L452R/ T478K/ D614G/ P681R/ S704A/ D950N/ V1264L/ P621S/
21J (Delta)	AY.6	T19R/ S247I/ L452R/ S459F/ T478K/ D614G/ P681R/ D950N/ A701S/

21J (Delta)	AY.99.1	T19R/ L452R/ T478K/ D614G/ P681R/ A845S/ D950N/
21J (Delta)	AY.99.2	L5F/ T19R/ L452R/ T478K/ D614G/ P681R/ D950N/ D867N/ D936Y/ C1243F/ T747N/ A626S/ A688T/ D1163Y/ A1078S/ T1117I/ V1176F/ G1219C/ M1229I/ S1030A/ K1191N/ N1187T/ T240I/ T1231I/ V3I

Fonte: O autor (2022).