



Universidade Federal de Pernambuco
Centro de Biociências

JUSSARA MOURA DE LIMA

**CARACTERIZAÇÃO GENÔMICA DE ISOLADOS DE SARS-CoV-2
CIRCULANTES EM PERNAMBUCO, BRASIL**

Recife
2022

JUSSARA MOURA DE LIMA

**CARACTERIZAÇÃO GENÔMICA DE ISOLADOS DE SARS-CoV-2
CIRCULANTES EM PERNAMBUCO, BRASIL**

Trabalho de Conclusão de Curso
apresentado ao Curso de Graduação em
Biomedicina da Universidade Federal de
Pernambuco, como pré-requisito à
obtenção do título de Bacharel em
Biomedicina.

Orientador: Prof. Dr. Wilson José da Silva
Junior

Recife
2022

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

Lima, Jussara Moura de .

Caracterização genômica de isolados de SARS-CoV-2 circulantes em
Pernambuco, Brasil / Jussara Moura de Lima. - Recife, 2022.
40 : il.

Orientador(a): Wilson José da Silva Junior

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de
Pernambuco, Centro de Biociências, Biomedicina, 2022.

Inclui referências, apêndices.

1. COVID-19. 2. Vigilância genômica. 3. Mutações. 4. Genoma. I. Silva Junior
, Wilson José da . (Orientação). II. Título.

570 CDD (22.ed.)

JUSSARA MOURA DE LIMA

**CARACTERIZAÇÃO GENÔMICA DE ISOLADOS DE SARS-CoV-2
CIRCULANTES EM PERNAMBUCO, BRASIL**

Trabalho de Conclusão de Curso
apresentado ao Curso de
Graduação em Biomedicina da
Universidade Federal de
Pernambuco, como pré-requisito à
obtenção do título de Bacharel em
Biomedicina.

Aprovada em: 01/12/2022

BANCA EXAMINADORA

Orientador: Dr. Wilson José da Silva Junior
Hospital Israelita Albert Einstein/ Varsomics

Profa. Dra. Vladia Maria Assis Costa
Universidade Federal de Pernambuco

Ms. Keyla Vitória Marques Xavier
FIOCRUZ Fundação Oswaldo Cruz - Instituto Aggeu Magalhães

Dedico este trabalho a minha família,
que nunca mediu esforços para me
apoiar em todos os momentos. E a
minha falecida mãe, Neide Moura de
Lucena que estaria muito orgulhosa.

AGRADECIMENTOS

Gostaria de agradecer primeiramente a Deus por ter me sustentado até aqui e por me ajudar a realizar mais este sonho.

Gostaria de agradecer ao meu orientador Dr. Wilson José da Silva Júnior pela orientação e por todos os ensinamentos durante a minha formação acadêmica.

Gostaria de agradecer imensamente ao Prof. Dr. Valdir de Queiroz Balbino por ter me orientado e me recebido no Laboratório de Bioinformática e Biologia evolutiva da Universidade Federal de Pernambuco (LABBE-UFPE), além de todas as oportunidades a mim concedidas.

Agradeço a todos os integrantes do LABBE-UFPE por todos os momentos e conhecimentos compartilhados. Especialmente a Bandeira, Heidi, Marcos, Dayana, Léo, Roberto, Karol, Keyla, Sérgio, Bruno e Abigail.

Gostaria de agradecer imensamente a Caio e a Zildene por todo apoio, troca de conhecimentos e madrugadas juntos. Sem vocês esse trabalho não seria possível.

Gostaria de agradecer a minha família por todo apoio, amor e paciência comigo. Em especial a minha mãe, Dalvaniz e ao meu pai, Marcos Emílio. Ao meu namorado, Gabriel Eduardo que sempre esteve presente em todos os momentos me apoiando e incentivando.

Agradeço aos meus amigos, Alex, Karolaine e Manoela por todo apoio, amor, conhecimento e momentos partilhados durante toda a graduação. Vocês tornaram a graduação especial.

Gostaria de agradecer às minhas colegas de estágio Kevelly, Vitória e Suellen, por todo o apoio e momentos partilhados.

DE LIMA, Jussara Moura . **Caracterização genômica de isolados de SARS-CoV-2 circulantes em Pernambuco, Brasil.** 40. Trabalho de Conclusão de Curso (Graduação em Biomedicina) – Universidade Federal de Pernambuco, Recife, 2022.

RESUMO

Síndrome Respiratória Aguda Grave Coronavírus 2 (SARS-CoV-2) é o mais recente coronavírus (CoV) identificado, responsável por causar a doença do coronavírus 2019 (COVID-19). Os primeiros relatos da doença foram em dezembro de 2019 na cidade de Wuhan, China. Posteriormente, o vírus se disseminou por diversos países ocasionando mais de 600 milhões de casos e 6 milhões de mortes em todo o mundo. As iniciativas de sequenciamento e vigilância genômica desenvolveram um papel fundamental frente à atual pandemia, permitindo a identificação e compreensão da evolução do SARS-CoV-2. Esse projeto teve como objetivo analisar os genomas das cepas de SARS-CoV-2 circulantes em Pernambuco durante os dois primeiros anos da pandemia da COVID-19 no estado, a fim de identificar as principais mutações e linhagens virais presentes. Para este fim, realizamos a montagem de 63 isolados sequenciados previamente pelo Laboratório de Bioinformática e Biologia Evolutiva da Universidade de Pernambuco (LABBE-UFPE), oriundos de diferentes municípios de Pernambuco. Além disso, também recuperamos genomas completos e com alta qualidade disponíveis no GISAID coletadas em Pernambuco entre março de 2020 a março de 2022, resultando em um conjunto de dados com 1898 sequências. Observamos a circulação de 44 linhagens, sendo 3 variantes de preocupação, P.1 (gama), B.1.617.2 (delta) e a B.1.1.7 (alfa). As variantes P.1 e AY.99.2 (sublinhagem Delta) apresentaram maior frequência e foram encontradas em todas as mesorregiões do estado de Pernambuco, demonstrando uma disseminação sustentada por todo o estado. Além disso, identificamos 3885 SNPs, sendo a maior parte de sua ocorrência em regiões codificantes do genoma viral e foram classificados como alterações missense e sinônimas de impacto baixo a moderado. Observamos as mutações C241T, C3037T (F105F), C14408T (P323L), A23403G (D614G) apresentando frequência acima de 99% e as G28881A (R203K), G28882A (R203R) e G28883C (G204R) presente e mais de 75% das cepas. A partir das análises evidenciamos a dinâmica das variantes circulantes no período de quase dois anos da pandemia no estado de Pernambuco. As diferentes variantes surgem conforme o vírus evolui e traz impactos à saúde pública, conforme observamos o aumento no número de casos e mortes associados à disseminação da P.1 no estado. Destacamos o importante papel da vigilância genômica no rastreamento das variantes e compreensão da evolução viral.

Palavras-chave: COVID-19. Vigilância genômica. Mutações. Genoma.

DE LIMA, Jussara Moura . **Genomic characterization of SARS-CoV-2 isolates from Pernambuco, Brazil**. 2022. 40. Trabalho de Conclusão de Curso (Graduação em Biomedicina) – Universidade Federal de Pernambuco, Cidade, ano.

ABSTRACT

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the latest identified coronavirus (CoV) responsible for causing coronavirus disease 2019 (COVID-19). The first reports of the disease were in December 2019 in the city of Wuhan, China. Subsequently, the virus spread to several countries, causing more than 600 million cases and 6 million deaths worldwide. Sequencing and genomic surveillance initiatives have played a key role in the face of the current pandemic, allowing the identification and understanding of the evolution of SARS-CoV-2. This project aimed to analyze the genomes of SARS-CoV-2 strains circulating in Pernambuco during the first two years of the COVID-19 pandemic in the state, in order to identify the main mutations and viral lineages present. For this purpose, we performed the assembly of 63 isolates previously sequenced by the Laboratório de Bioinformática e Biologia Evolutiva da Universidade de Pernambuco (LABBE-UFPE), from different municipalities in Pernambuco. In addition, we also recovered high quality sequences available at GISAID collected in Pernambuco between March 2020 and March 2022, resulting in a dataset with 1898 genomes. We observed the circulation of 44 lineages, with 3 variants of concern, P.1 (gamma), B.1.617.2 (delta) and B.1.1.7 (alpha). The P.1 and AY.99.2 variants (Delta sublineage) were more frequent and were found in all mesoregions of the state of Pernambuco, demonstrating a sustained spread throughout the state. In addition, we identified 3885 SNPs, most of which occur in coding regions of the viral genome and were classified as missense alterations and synonymous with low to moderate impact. We observed a predominance of GISAID clades GR and mutations C241T, C3037T (F105F), C14408T (P323L), A23403G (D614G) presenting a frequency above 99% and G28881A (R203K), G28882A (R203R) and G28883C (G204R) present in more than 75% of the strains. From the analyzes we evidenced the dynamics of the circulating variants in the period of almost two years of the pandemic in the state of Pernambuco. The different variants emerge as the virus evolves and impacts public health, as we observe an increase in the number of cases and deaths associated with the spread of P.1 in the state. We highlight the important role of genomic surveillance in screening for variants and understanding viral evolution.

Key words: COVID-19. genomic surveillance. Mutations. Genome.

LISTA DE ILUSTRAÇÕES

- Figura 1** – Taxonomia da família *Coronaviridae* de acordo com a classificação do Comitê Internacional de Taxonomia de Vírus (ICTV). 13
- Figura 2** – Esquema da estrutura da genômica e dos componentes estruturais do SARS-CoV-2. 17
- Figura 3** – Distribuição das 1898 amostras nas mesorregiões de Pernambuco. 21
- Figura 4** – A: Distribuição dos SNPs por região genômica dos genomas de SARS-CoV-2 isolados em Pernambuco. B: Distribuição por gene dos SNPs classificados de acordo com seu impacto nos genomas de SARS-CoV-2 circulantes em Pernambuco. 23
- Figura 5** – Distribuição das linhagens de SARS-CoV-2 circulantes em Pernambuco por mês de amostragem (Março de 2020 a Janeiro de 2022) com frequência maior que 10%. 25
- Figura 6** – Dados epidemiológicos mensais de Pernambuco datados entre Março de 2020 a Janeiro de 2022 de acordo com os boletins epidemiológicos da COVID-19 divulgados pela Secretaria de Saúde de Pernambuco. Em azul observamos os números de casos novos, em vermelho o número de mortes, em amarelo o número de casos de Síndrome Respiratória Aguda Grave (Srag) e em verde observamos o número de casos leves. 27
- Figura 7** – Árvore filogenética de máxima verossimilhança gerada pelo pipeline do Nexstrain, composta por 1989 cepas de SARS-CoV-2 de diferentes regiões de Pernambuco. Cada amostra está colorida de acordo com a linhagem a qual foi classificada pelo PANGOLIN. 28
- Figura 8** – Árvore filogenética de máxima verossimilhança gerada pelo pipeline do Nexstrain, composta por 1989 cepas de SARS-CoV-2 de diferentes regiões de Pernambuco. Cada amostra está colorida de acordo com o clado GISAID. 29

LISTA DE ABREVIATURAS E SIGLAS

CoVs	Coronavírus
GISAID	Iniciativa Global para o Compartilhamento de Todos os Dados sobre Influenza
MERS-CoV	Middle East respiratory syndrome coronavirus
SNP	Single Nucleotide Polymorphism – Polimorfismo de Nucleotídeo Único
PANGOLIN	Phylogenetic Assignment of Named Global Outbreak Lineages – Atribuição filogenética de linhagens de surto globais nomeada
RBD	Receptor Binding Domain – Domínio de ligação ao receptor
RNA	Ácido ribonucleico
SARS	Síndrome respiratória aguda grave
SARS-CoV	Severe acute respiratory syndrome coronavirus
VOC	Variant Of Concern - Variante de preocupação

SUMÁRIO

1	REVISÃO BIBLIOGRÁFICA	
1.1	CLASSIFICAÇÃO TAXONÔMICA DA FAMÍLIA <i>CORONAVIRIDAE</i> E SARS-CoV-2	13
1.2	ORIGEM DO SARS-COV-2	15
1.3	ESTRUTURA VIRAL E GENÔMICA DO SARS-COV-2	16
2	OBJETIVOS	
2.1	OBJETIVO GERAL	18
2.2	OBJETIVOS ESPECÍFICOS	18
2	METODOLOGIA	
3.1	OBTENÇÃO DAS AMOSTRAS	19
3.2	MONTAGEM DOS GENOMAS	19
3.3	ANOTAÇÃO DOS GENOMAS	20
3.4	CLASSIFICAÇÃO DAS LINHAGENS	20
3.5	ÁRVORE FILOGENÉTICA	20
4	RESULTADOS E DISCUSSÃO	21
5	CONCLUSÃO	30
	REFERÊNCIAS	31
	APÊNDICES	
	APÊNDICE A – Frequência das linhagens classificadas pelo PANGOLIN para os 1898 genomas de SARS-CoV-2 circulantes em Pernambuco.	37
	APÊNDICE B – Mutações identificadas com frequência maior que 50% entre os 1898 genomas analisados neste estudo.	39

Introdução

A doença do coronavírus 2019 (COVID-19) causada pelo SARS-CoV-2 teve seu primeiro relato em dezembro de 2019 na cidade de Wuhan, China. Até o momento, a doença ocasionou o registro de 632.953.782 casos e o acometimento de 6.593.715 mortes em todo o mundo (até 16 de novembro de 2022, <https://covid19.who.int/>). O Brasil teve o primeiro caso de COVID-19 confirmado em fevereiro de 2020 no estado de São Paulo (RODRIGUEZ-MORALES et al., 2020) e atualmente é o 5º país com mais casos notificados do mundo, totalizando 34.921.054 casos e 688.694 mortes. Em Pernambuco, o primeiro caso notificado foi em 12 de março de 2020, de acordo com um boletim publicado pela secretaria de saúde do estado de Pernambuco. Atualmente o estado contabiliza 1.068.886 casos notificados e 22.427 óbitos (16 de Novembro de 2022, <https://covid.saude.gov.br/>).

Síndrome Respiratória Aguda Grave Coronavírus 2 (SARS-CoV-2) é um vírus envelopado, com RNA de fita simples e sentido positivo pertencente a ordem Nidovirales, família Coronaviridae e ao gênero *Betacoronavirus* (RAHIMI; MIRZAZADEH; TAVAKOLPOUR, 2021). Seu genoma possui aproximadamente 29,9 kb de comprimento, os quais são distribuídos em 12 ORFs, que codificam proteínas não estruturais, acessórias e proteínas estruturais.

O Brasil, até então, possui 189.092 genomas completos depositados no GISAID (<https://gisaid.org/>), um banco de dados público de compartilhamento de dados de vírus influenza e do coronavírus causador da COVID-19. Enquanto Pernambuco possui 6.292 genomas completos depositados no GISAID, representando 3,34% dos genomas depositados (19 de Novembro de 2022). A partir do sequenciamento e análise desses genomas foi possível detectar as entradas, as linhagens predominantes e a dinâmica da disseminação do vírus no país (CANDIDO et., 2020).

Devido a alta taxa de mutação desse vírus (KHATAMZAS et al., 2021) diversas variantes surgiram, dentre elas, variantes de preocupação que ocasionaram aumento no número de casos em diversas regiões do Brasil e do mundo (ROMERO et al., 2021; DEMOLINER et al., 2021; GULARTE et al., 2022). Os estudos genômicos são importantes no rastreamento de novas variantes, fornecer informações que possibilitem a adoção de medidas de contenção da propagação e para o desenvolvimento de vacinas.

1 REVISÃO BIBLIOGRÁFICA

1.1 CLASSIFICAÇÃO TAXONÔMICA DA FAMÍLIA *CORONAVIRIDAE* E DO SARS-CoV-2

Os Coronavírus (CoVs) são vírus de RNA de fita simples e sentido positivo com tamanho variando entre 26 a 32 kilobases (kb), envelopados e com peplômeros na superfície viral (GRAHAM; DONALDSON; BARIC, 2013; CHANNAPPANAVAR,; PERLMAN , 2017) Esse grupo de vírus pertencem à subfamília *Orthocoronavirinae* e juntamente com a subfamília *Torovirinae* compõe a família *Coronaviridae* da ordem *Nidovirales* e reino *Riboviria* (figura 1). A subfamília *Orthocoronavirinae* é composta por quatro gêneros: *alphacoronavirus* (α -CoV), *betacoronavirus* (β -CoV), *gammacoronavirus*, (γ -CoV) e *deltacoronavirus* (d-CoV) (WOO et al., 2012) .

Figura 1. Taxonomia da família *Coronaviridae* de acordo com a classificação do Comitê Internacional de Taxonomia de Vírus (ICTV)



Fonte: Silveira (2022)

Os CoVs podem infectar humanos e vertebrados e ocasionar doenças no sistema respiratório, gastrointestinal, hepático e nervoso central (CHEN; LIU; GUO, 2020). Dentre os coronavírus, até o momento, são conhecidas sete espécies patogênicas para humanos (HCoVs), pertencentes aos gêneros *alphacoronavirus* (α -CoV) e *betacoronavirus* (β -CoV) que ocasionam desde de doenças leves a quadros fatais (DHAMA et al., 2020). O HCoVs 229E (HCoV-229E) e o NL63 (HCoV-NL63) pertencem ao gênero *alphacoronavirus*. Já os HCoV-OC43 e HCoV-HKU1 compõem o gênero *betacoronavirus* e subgênero *Embecovirus*. Esses quatro HCoVs são associados a resfriados comuns e representam de 10 a 30% das infecções do trato respiratório superior em adultos (ABDUL-RASOOL; FIELDING, 2010; PAULES; MARSTON; FAUCI, 2020).

Diferentemente destes, os três últimos HCoVs que emergiram nas últimas duas décadas podem atingir o trato respiratório inferior e ocasionar quadros respiratórios graves podendo ser fatais. O primeiro foi o SARS-CoV membro do gênero *betacoronavirus* e subgênero *Sarbecovirus*, identificado na China em 2002 ocasionando a Síndrome Respiratória Aguda Grave (SARS, Severe Acute Respiratory Syndrome) (AL-ROHAIMI; AL OTAIBI, 2020). O vírus se disseminou por diferentes países e levou a notificação de mais de 8 mil casos e 812 mortes (KSIAZEK et al., 2003). Em setembro de 2012, ocorreu o primeiro caso da síndrome respiratória do Oriente Médio (MERS) na Arábia Saudita, ocasionado pelo MERS-CoV, o qual levou a notificação de 2.220 casos e 790 mortes apresentando uma taxa de mortalidade acima de 35% (Zaki et al., 2012). Esse HCoV também pertence ao gênero *betacoronavirus*, contudo é integrante do subgênero *Merbecovirus*.

O mais recente HCoV foi identificado na cidade de Wuhan, na China, em dezembro de 2019 associado a um surto de pneumonia. Posteriormente, denominado de SARS-CoV-2 (Síndrome Respiratória Aguda Grave Coronavírus 2) pelo Comitê Internacional de Taxonomia de Vírus (ICTV) (ZHOU, Peng et al., 2020). A doença foi nomeada pela Organização Mundial da Saúde como a doença do coronavírus 2019 (COVID-19). O vírus rapidamente se espalhou para vários países e em 30 de janeiro de 2020 a Organização Mundial da Saúde (OMS) declarou Emergência de Saúde Pública de Importância Internacional (ESPII) (SHANG et al., 2020). Assim como o SARS-CoV, o SARS-CoV-2 também integra o gênero *betacoronavirus* e subgênero *Sarbecovirus* (GORBALENYA et al., 2020).

1.2 ORIGEM DO SARS-CoV-2

A origem do SARS-CoV-2 ainda não foi completamente elucidada. Após o surgimento do SARS-CoV e do MERS-CoV, diversos estudos buscaram compreender a origem e foi constatado que ambos eram próximos filogeneticamente de CoV de morcegos (LI et al., 2005; ITHETE et al., 2013). Posteriormente, foi identificado que esses vírus foram transmitidos para humanos por meio de hospedeiros intermediários, o SARS-CoV por meio das civetas e o MERS-CoV através dos camelos dromedários (GUAN et al., 2003; CUI; LI; SHI, 2019).

A partir da comparação do genoma do SARS-CoV-2 com outros coronavírus, CHAN et al, (2020) relatou 82% de identidade nucleotídica do novo coronavírus com o SARS-CoV e 89% de similaridade com o coronavírus de morcego. Além disso, CHAN et al, (2020) também comparou as relações filogenéticas da sequência de aminoácidos da ORF1ab e dos 4 genes estruturais (S, E, M e N), e para todos os genes o SARS-CoV-2 estava mais relacionado à SARS de morcego CoVs (ZXC21 e ZC45). ZHOU et al, (2020) também relataram similaridades de 96% do SARS-CoV-2 com o genoma do coronavírus de morcego de *Rhinolophus affinis* (BatCoV RaTG13). Esses achados indicam que os morcegos são os prováveis hospedeiros reservatórios do SARS-CoV-2. Entretanto, ainda não foi comprovado um salto entre espécies de morcegos para infectar humanos ou a existência de um hospedeiro intermediário atuando, semelhante ao observado com o SARS-CoV e o MERS-CoV (PEKAR et al., 2021).

Os coronavírus geralmente infectam de forma espécie-específica (MASTERS, 2006). Contudo, os CoVs apresentam uma alta taxa mutacional e de recombinação, o que favorece a adaptação e transferência entre hospedeiros, representando um risco para o surgimento de outros CoVs zoonóticos (PACHETTI et al., 2020). Estudos voltados para a compreensão da transmissão entre espécies são de extrema importância para prevenir e fornecer informações importantes para o enfrentamento a possíveis HCoV emergentes (FORNI et al., 2017).

1.3 ESTRUTURA VIRAL E GENÔMICA DO SARS-CoV-2

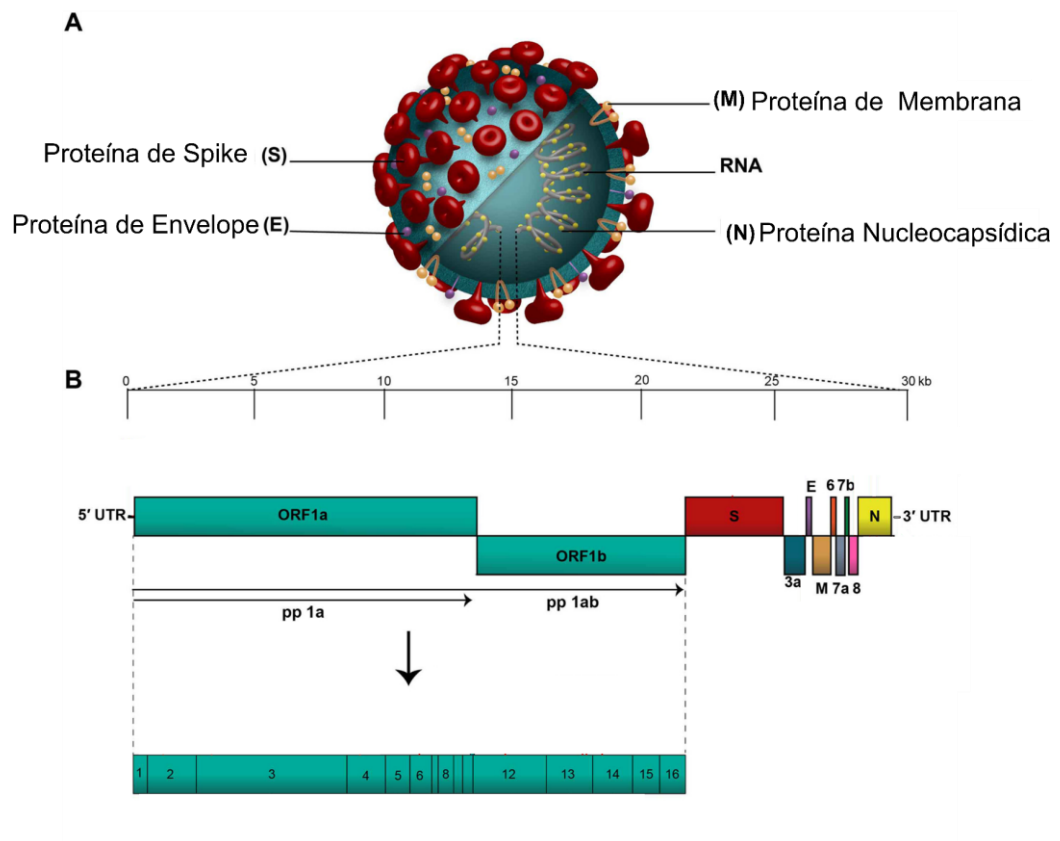
O SARS-CoV-2 é um vírus de RNA de fita simples e sentido positivo, não segmentado, envelopado, esférico, com diâmetro variando entre 70 a 120 nanômetros (CHEN; LIU; GUO, 2020). O vírion apresenta 4 proteínas estruturais: a glicoproteína spike (S), a proteína de membrana (M), envelope (E) e nucleocapsídeo (N) (Figura 2A). A glicoproteína spike (S) é dividida em duas subunidades a S1 e S2, onde a S1 possui o domínio de ligação ao receptor (*RBD-receptor-binding domain*) que se liga ao receptor ACE2 (Enzima Conversora de Angiotensina 2), a S2 é responsável pela fusão das membranas (MASTERS, 2006). A proteína S do SARS-CoV-2 apresenta uma afinidade à ACE2 humana 10 a 20 vezes maior que à proteína S do SARS-CoV (GHEBLAWI et al., 2020).

A proteína de membrana (M) interage com outras proteínas estruturais, participa do empacotamento viral, confere forma ao envelope viral e apresenta-se em maior número comparado às outras proteínas estruturais (DE HAAN et al., 2000; BAI; ZHONG; GAO, 2021). A proteína N se liga ao RNA viral, formando o nucleocapsídeo no interior do envelope, desempenhando um papel importante na manutenção do RNA ordenado para replicação e transcrição (CHANG et al., 2016; LIANG et al., 2020). A proteína E é a menor das proteínas e participa da montagem viral, brotamento, formação de envelope e patogênese (RAHMAN et al., 2021).

A organização do genoma do SARS-CoV-2 é semelhante a outros coronavírus. O tamanho é de aproximadamente 29,9 kb distribuídos em 12 ORFs (*Open reading frame* - quadros de leitura aberto) com regiões não codificantes em ambas extremidades, a 5' UTR e 3' UTR (Figura 2B) e apresenta uma estrutura cap 5' e cauda poli-A na extremidade 3' conferindo capacidade de atuar como um RNA mensageiro para a tradução das poliproteínas replicase (CHAN et al., 2020). Na porção 5' encontra-se ocupando dois terços do genoma duas ORFs, ORF1a e a ORF1b, traduzidas em duas poliproteínas, a pp1a e pp1ab, posteriormente processadas e subdivididas em 16 proteínas não estruturais associadas a transcrição e replicação viral (Figura 2B) (RAHIMI; MIRZAZADEH; TAVAKOLPOUR, 2021). Além dessas proteínas, o vírus também apresenta proteínas acessórias codificadas pelas ORF3a, ORF6, ORF7a, ORF7b, ORF8 e ORF10, as quais podem estar associadas a supressão das respostas imunes, virulência e adaptação do vírus a hospedeiros específicos (FORNI et al., 2017). Nas regiões intergênicas das

proteínas estruturais e acessórias encontram-se sequências regulatórias para a expressão de cada gene (FEHR, Anthony; PERLMAN, 2015)

Figura 2. Esquema da estrutura da genômica e dos componentes estruturais do SARS-CoV-2.



Fonte: Adaptado de RAHIMI (2021)

2 OBJETIVOS

2.1 OBJETIVO GERAL

Esse projeto teve como objetivo realizar a caracterização genômica das cepas de SARS-CoV-2 circulantes durante os dois primeiros anos da pandemia da COVID-19 em Pernambuco, Brasil.

2.1 OBJETIVOS ESPECÍFICOS

- Realizar a montagem dos genomas de SARS-CoV-2 obtidos neste estudo;
- Avaliar as principais mutações presentes nos genomas dos isolados pernambucanos;
- Classificar as linhagens circulantes.

3 METODOLOGIA

3.1 OBTENÇÃO DAS AMOSTRAS

Para o presente estudo foram utilizadas 63 cepas de SARS-CoV-2 sequenciadas pelo Laboratório de Bioinformática e Biologia Evolutiva da Universidade Federal de Pernambuco (LABBE-UFPE), coletadas entre 23 de março a 07 de agosto de 2021. As amostras foram obtidas através de swabs nasofaríngeos e orofaríngeos de pacientes positivos para COVID-19 de 24 municípios do estado de Pernambuco, Brasil. As amostras foram encaminhadas ao Núcleo de Pesquisa em Inovação Terapêutica - Suely Galdino (NUPIT-SG), na Universidade Federal de Pernambuco (UFPE), pelo Laboratório Central de Saúde Pública do Estado de Pernambuco (LACEN-PE) para o diagnóstico molecular da COVID-19. As amostras positivas foram encaminhadas ao LABBE para o sequenciamento *paired-end* através da plataforma MiSeq (Illumina).

. Para compor as análises, foi realizado o download de sequências de SARS-CoV-2 disponíveis no GISAID - *Global Initiative on sharing all influenza data* (<https://www.gisaid.org/>) com data de coleta entre 01 de Março de 2020 a 31 de Março de 2022. Os filtros utilizados para a filtragem dos genomas foram os de completo (>29.000 pb) e alta cobertura (< 1% de Ns e <0,05% de mutações únicas de aminoácido não vistas nas sequências do banco de dados).

3.2 MONTAGEM DOS GENOMAS

Os dados brutos obtidos do sequenciamento foram submetidos ao software FastQC versão 0.11.9 (ANDREWS, 2010) a fim de avaliar a qualidade, comprimento e presença de adaptadores. Posteriormente, os reads que apresentaram valores de comprimento e qualidade do fragmento inferior a 20 foram filtrados pelo software Sickel versão 1.33 (Joshi & Fass, 2011). A montagem foi realizada através do software BWA versão 0.7.17-r1188 (Li & Durbin, 2010), usando a sequência de referência Wuhan-Hu-1, disponível no NCBI (ID: NC_045512.2, <https://www.ncbi.nlm.nih.gov/>). O software Samtools versão 1.10 (Li et al., 2009) foi utilizado para converter os dados, remover duplicatas, indexar e obter o genoma consenso para cada amostra. Posteriormente, foi avaliada a qualidade da montagem pelo QUAST v5.0.2 (GUREVICH et al., 2013).

3.3 ANOTAÇÃO DOS GENOMAS

Os genomas no formato fasta oriundos do GISAID foram alinhados pelo Bowtie2 versão 2.3.5.1 (LANGMEAD; SALZBERG, 2012) ao genoma de referência NC_045512.2. Posteriormente, os arquivos obtidos foram convertidos, ordenados e indexados pelo Samtools versão 1.10 (Li et al., 2009). Em seguida, todos os genomas obtidos neste estudo (sequenciados + obtidos do GISAID) foram submetidos à chamada de variantes realizada por meio do Freebayes versão v1.3.0 (GARRISON & MARTH, 2012), para detectar Polimorfismos de Nucleotídeo Único (SNPs) e obter os arquivos VCFs. O VCFtools versão 0.1.15 (DANECEK et al., 2011) foi utilizado para unir os arquivos VCFs em um único arquivo e, posteriormente, esse arquivo foi anotado pelo SnpEff versão 5.1 (CINGOLANI et al., 2012).

3.4 CLASSIFICAÇÃO DAS LINHAGENS

Os genomas que compõem esse estudo foram submetidos a um alinhamento múltiplo de sequências usando o MAFFT versão 7.310 (KATO et al., 2019). Em seguida, as linhagens foram definidas usando o método de classificação de linhagem dinâmica proposto por Rambaut et al. (2020), por meio do pacote de software *Phylogenetic Assignment of Named Global Outbreak Lineages* (PANGOLIN, versão 1.2.127) (<https://pangolin.cog-uk.io/>). O objetivo foi identificar as linhagens de SARS-CoV-2 circulantes em Pernambuco.

3.5 ÁRVORE FILOGENÉTICA

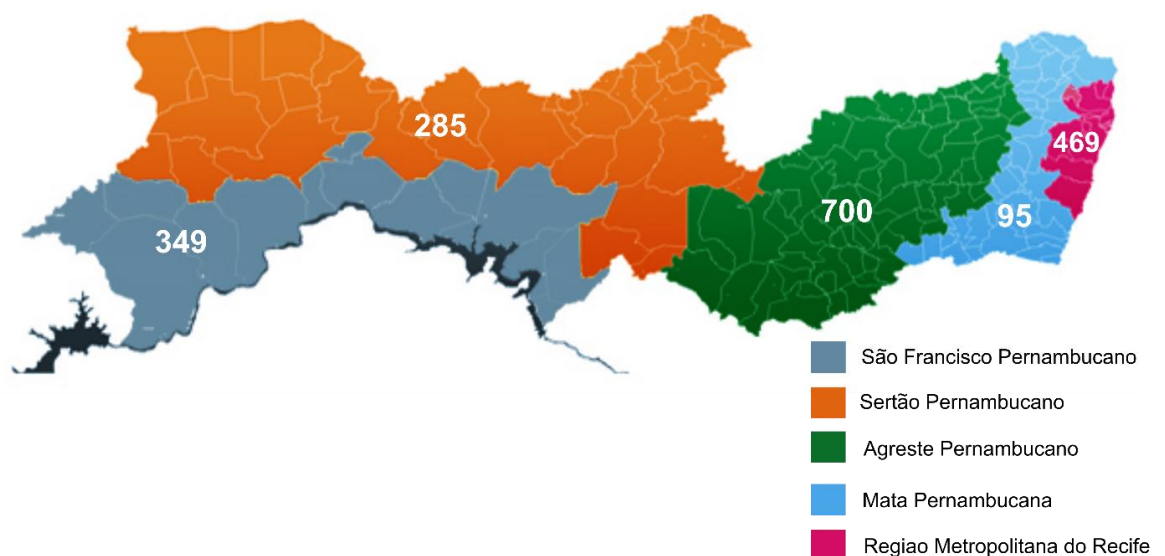
A árvore filogenética de máxima verossimilhança foi estimada utilizando o pipeline do Nextstrain (HADFIELD et al., 2018). O modelo evolutivo utilizado pelo IQ-TREE (NGUYEN et al., 2015) para construir a árvore foi *General Time Reversible Model* GTR. A árvore gerada foi resolvida no tempo pelo TreeTime (V.0.8.2) (SAGULENKO; PULLER & NEHER, 2018). Posteriormente, os arquivos gerados (.json) foram visualizados pelo Auspice (<https://auspice.us/>).

4 RESULTADOS E DISCUSSÃO

4.1 AMOSTRAS

Para o presente estudo avaliamos 1.898 genomas de cepas do SARS-CoV-2 circulantes no estado de Pernambuco (63 montadas + 1.835 obtidos do GISAID), com data de coleta entre o período de 01 de março de 2020 a 17 de janeiro de 2022. Essas amostras são oriundas de 143 municípios do estado, distribuídas nas 5 mesorregiões, sendo o Agreste Pernambucano composto por 700 amostras, a região Metropolitana do Recife por 469 amostras, São Francisco Pernambucano por 349, Sertão Pernambucano por 285 e Mata pernambucana com 95 (Figura 1).

Figura 3: Distribuição das 1898 amostras nas mesorregiões de Pernambuco.



Fonte: Adaptado IBGE (2022)

4.2 CHAMADA DE VARIANTES

O SARS-CoV-2 atingiu mais de 190 países desde a sua primeira notificação em Wuhan, China (<https://covid19.who.int/table>). Conforme o vírus se difunde pelas diversas regiões do mundo, ocorre o acúmulo de mutações que podem conferir qualidades adaptativas, como aumento da transmissibilidade e redução na neutralização por anticorpos (WALENSKY et al, 2021; SYED et al., 2022). Para avaliar as mutações presentes nas cepas de SARS-CoV-2 circulantes em Pernambuco, realizamos uma chamada de variantes. Essa análise foi feita através da comparação entre o genoma completo das amostras estudadas e o genoma da

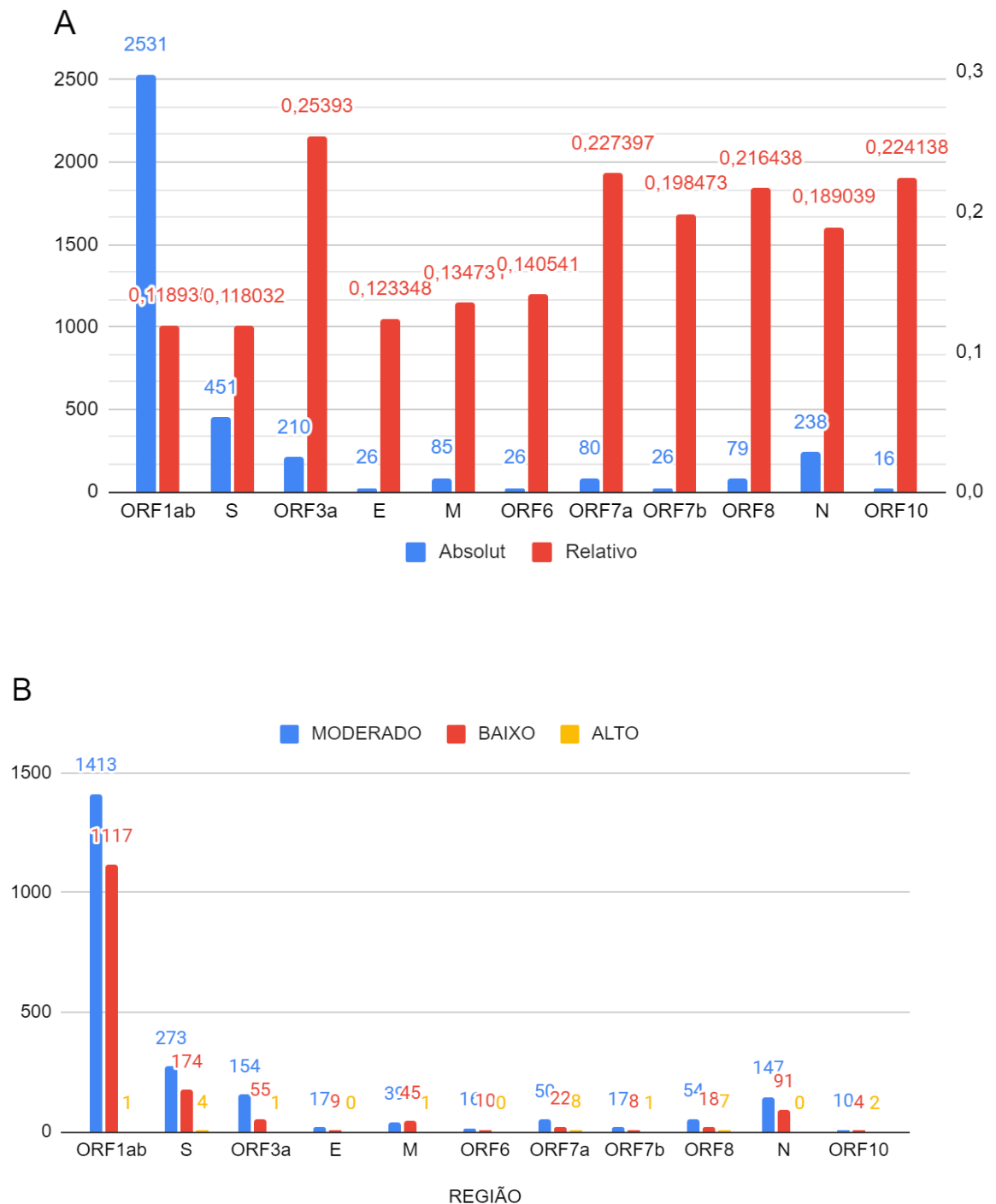
cepa de referência Wuhan-Hu-1 (ID: NC_045512.2). A partir dessa análise, observamos a presença de 3885 SNPs (*Single nucleotide polymorphism*), sendo 96,99% (n=3768) das mutações em regiões codificantes e 3,01% (n=117) em regiões não codificantes.

Dentre as variantes em regiões codificantes, 2.190 (58,12%) correspondem a mutações missense, no qual o códon mutado resulta na alteração do aminoácido, 1553 (41,22%) correspondem a mutações sinônimas, que correspondem a alterações onde o códon alterado produz o mesmo aminoácido, 20 (0,53%) mutações são de ganho de códon de parada (*stop gained*) e 5 (0,13%) são de perda do códon inicial (*start lost*). Esse achado condiz com relatórios anteriores, onde a maior parte das mutações encontradas nos genomas são alterações missense e sinônimas (LAAMARTI et al., 2020; RAHIMI; MIRZAZADEH; TAVAKOLPOUR, 2021)

A ORF1ab apresentou 2.531 (67,17%) variantes, sendo a maior em número de mutações, seguida pelo gene da proteína spike com 451 (11,97%) mutações, o gene N com 238 (6,32%) e a ORF3a com 210 (5,57%) variantes (figura 1A). Porém, quando os dados foram normalizados em relação ao tamanho de cada gene, observamos que a ORF3a apresentou uma maior proporção de mutações por região gênica. A ORF3a está associada à replicação e liberação do vírus, mutações nessa região podem conferir ao vírus capacidade de evasão do sistema imune do hospedeiro e causar um aumento na produção de citocinas (BIANCHI; BORSETTI; CICCOCCHI, 2021; MAJUMDAR; NIYOGI, 2020).

Os SNPs também foram classificados em relação ao impacto causado nos genomas do SARS-CoV-2. As categorias atribuídas são as de baixo impacto (variante que presume-se não causar alteração na proteína), impacto moderado (a variante pode vir ocasionar uma alteração na funcionalidade da proteína) e alto impacto (é uma variante que provavelmente pode ocasionar o truncamento da proteína, perda da função ou alterar sua expressão) (CINGOLANI, et al., 2012). A ORF1ab apresentou os maiores valores nas categorias de moderado e baixo impacto, seguido pela proteína Spike (figura 1B). A ORF7a apresentou maior número de mutações de alto impacto. Após a normalização dos dados de acordo com o tamanho da região gênica, a ORF3a se destacou na categoria de impacto moderado, o gene N em baixo impacto e a ORF7a em alto impacto.

Figura 4. A: Distribuição dos SNPs por região genômica dos genomas de SARS-CoV-2 isolados em Pernambuco. B: Distribuição por gene dos SNPs classificados de acordo com seu impacto nos genomas de SARS-CoV-2 circulantes em Pernambuco.



Fonte: Lima (2022)

Para compreender melhor a diversidade e expansão do vírus foi criado a nomenclatura de clados e linhagens baseado na ocorrência simultânea de mutações e relações filogenéticas pelo GISAID e Rambaut et al. (2020), respectivamente. Dentre as 3885 mutações, identificamos 31 que obtiveram frequência superior a 50% das cepas estudadas (apêndice B), sendo 1 em região não codificante e as outras em regiões gênicas. Das mutações em regiões codificantes, 8 são do tipo sinônima e 22 do tipo missense com troca de aminoácido.

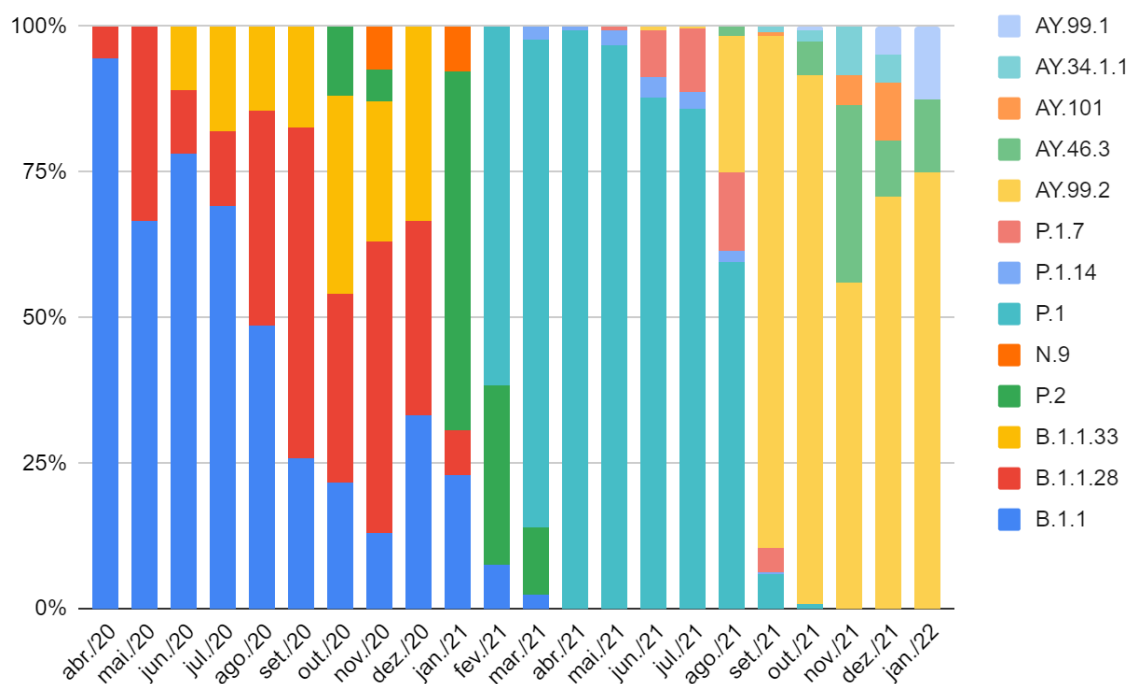
A ORF1ab e a Spike (S) apresentaram 11 mutações cada, e todas localizadas no gene da spike foram classificadas como missense. O gene N apresentou 6 mutações, a ORF3a e a ORF8 com 1 mutação cada. Dentre as mutações identificadas, a C241T, C3037T (F105F), C14408T (P323L), A23403G (D614G) apresentaram frequências acima de 99% e são observadas no clado GISAID G. Além dessas mutações, mais de 75% das cepas também apresentaram as mutações trinucleotídicas G28881A (R203K), G28882A (R203R) e G28883C (G204R) na proteína N, associadas ao clado GISAID GR. Todas as alterações anteriormente citadas estão entre as mutações mais frequentes relatadas na literatura (RAHIMI; MIRZAZADEHBC; TAVAKOLPOUR, 2021).

4.3 CLASSIFICAÇÃO DAS LINHAGENS

Através da atribuição de linhagens realizada pelo software PANGOLIN, observamos 4 linhagens atribuídas para as 63 amostras montadas neste estudo. A P.1 foi a mais frequente com 54 sequências, seguida pela P.1.14 (n=5), P.1.7 (n=3) e P.1.15 (n=1). Quando avaliamos o dataset completo, identificamos 44 linhagens circulantes em Pernambuco entre o período de março de 2020 a janeiro de 2022 (apêndice A). Sendo a variante de preocupação P.1 (gama) a mais prevalente com 852 sequências, seguida pela sublinhagem da variante de preocupação Delta AY.99.2 (n=356), B.1.1 (n=169), B.1.1.28 (n=162), e a linhagem B.1.1.33 (n=84). Em relação a distribuição das linhagens por cada mesorregião, observamos a região metropolitana do Recife apresentando a maior diversidade de linhagens (n=29), seguida da mesorregião do Agreste Pernambucano (n=25), São Francisco Pernambucano (n=22), Sertão Pernambucano (n=13) e a mesorregião da Mata Pernambucana (n=10).

A disseminação do vírus resultou no surgimento de diversas mutações. Quando as mutações que conferem vantagens adaptativas são mantidas nas variantes, observa-se aumento de suas frequências, como observado com a mutação D614G na proteína spike gerando diferentes padrões de dominância entre as variantes (KORBER et al., 2020; VOLZ et al., 2021). Assim como em outras regiões, Pernambuco apresentou uma variação na dominância das linhagens ao longo do tempo (Figura 5). Nos momentos iniciais da epidemia em Pernambuco a linhagem com maior prevalência foi a B.1.1 (PAIVA et al., 2020), diferentemente dos estados do sudeste que apresentaram as linhagens B.1.1.33 e B.1.1.28 em maior frequência na fase inicial (RESENDE et al., 2020; CAMARGO et al., 2021). Posteriormente, essas linhagens apresentaram um aumento na frequência até o surgimento da P.2 em outubro de 2020 (SILVEIRA, 2020), apresentando maior frequência em janeiro de 2021 (Figura 5). No mês seguinte à detecção da P.2 em Pernambuco, observamos um aumento no número de casos possivelmente impulsionado por essa variante (Figura 6).

Figura 5. Distribuição das linhagens de SARS-CoV-2 circulantes em Pernambuco por mês de amostragem (Março de 2020 a Janeiro de 2022) com frequência maior que 10%.



Fonte: Lima (2022)

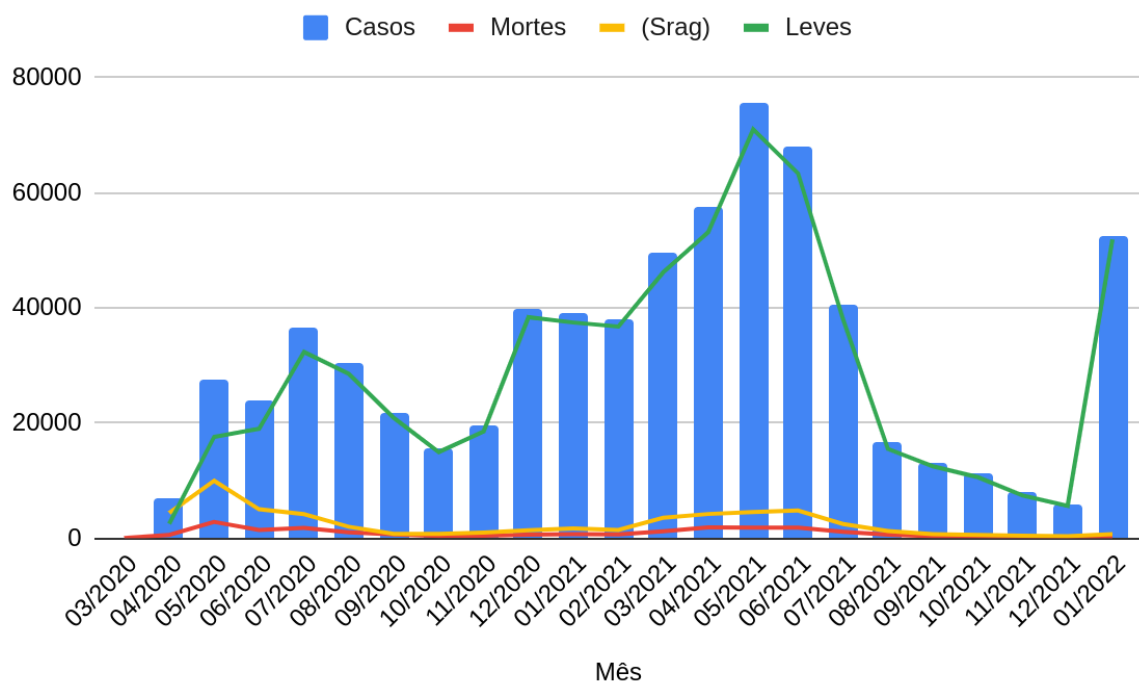
Dentre as diferentes linhagens do SARS-CoV-2 que emergiram, as linhagens de preocupação (VOC) se destacam por apresentar mutações que conferem uma maior aptidão viral e consequente aumento na disseminação em um curto espaço de tempo (KONINGS et al., 2021). A partir da análise dos genomas utilizados nesse estudo, identificamos a circulação de 3 variantes de preocupação, a B.1.1.7 (alfa), a B.1.617.2 (Delta) e P.1 (gama) com predominância entre fevereiro a agosto de 2021, posteriormente sendo substituída pela sublinhagem Delta AY.99.2, cuja frequência apresentou-se maior entre de setembro de 2021 a Janeiro de 2022 (figura 5).

A linhagem de preocupação P.1 do SARS-CoV-2 é descendente da linhagem B.1.1.28 e foi detectada pela primeira vez em dezembro de 2020 em Manaus, estado do Amazonas, Brasil (FARIA et al., 2021). Essa variante apresenta mutações importantes no domínio de ligação ao receptor como a K417T, E484K, e N501Y associadas a uma maior afinidade com o receptor ACE-2 e escape imunológico (SANCHES et al., 2021; YANG et al., 2022). Em Pernambuco, a P.1 foi detectada pela primeira vez em fevereiro de 2021 e durante o predomínio dessa variante observamos o maior pico no número de casos em Pernambuco no período estudado, além do aumento no número de mortes e casos mais graves (Figura 6). Conforme constatado por FunK et al, (2021), a linhagem P.1 está associada a um risco aumentado de hospitalização e internação em UTI.

A segunda linhagem mais prevalente em Pernambuco foi a AY.99.2, uma sublinhagem da VOC Delta, a qual teve seu primeiro caso detectado na Índia em outubro de 2020 (<https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>). A variante Delta surgiu e substituiu a Gama em diversas regiões do Brasil e do mundo (ALIZON et al., 2021; GULARTE et al., 2022). Porém, semelhante ao constatado por Gualarte et al. (2022), no Rio Grande do Sul, o aumento da sua frequência não foi associado ao aumento no número de casos em Pernambuco. A primeira amostra classificada como sub-linhagem Delta observada nos nossos dados, em junho de 2021, foi a AY.99.2, apresentando maior incidência entre setembro de 2021 a janeiro de 2022 (Figura 5). Durante esse período foi evidenciado uma redução no número de casos no estado (Figura 6).

A linhagem de preocupação B.1.1.7 foi primeiramente identificada em setembro de 2020 no Reino Unido (FUNK et al, 2021). No Brasil, a transmissão local da linhagem foi relatada em São Paulo em dezembro de 2020 (CLARO et al., 2021). A primeira detecção em Pernambuco ocorreu em abril de 2021 (BEZERRA JANUÁRIO, 2022). A frequência dessa linhagem foi baixa, apresentando-se em apenas 4 sequências. Contudo, esse achado pode ser em decorrência de uma subamostragem.

Figura 6. Dados epidemiológicos mensais de Pernambuco datados entre Março de 2020 a Janeiro de 2022 de acordo com os boletins epidemiológicos da COVID-19 divulgados pela Secretaria de Saúde de Pernambuco. Em azul observamos os números de casos novos, em vermelho o número de mortes, em amarelo o número de casos de Síndrome Respiratória Aguda Grave (Srag) e em verde observamos o número de casos leves.

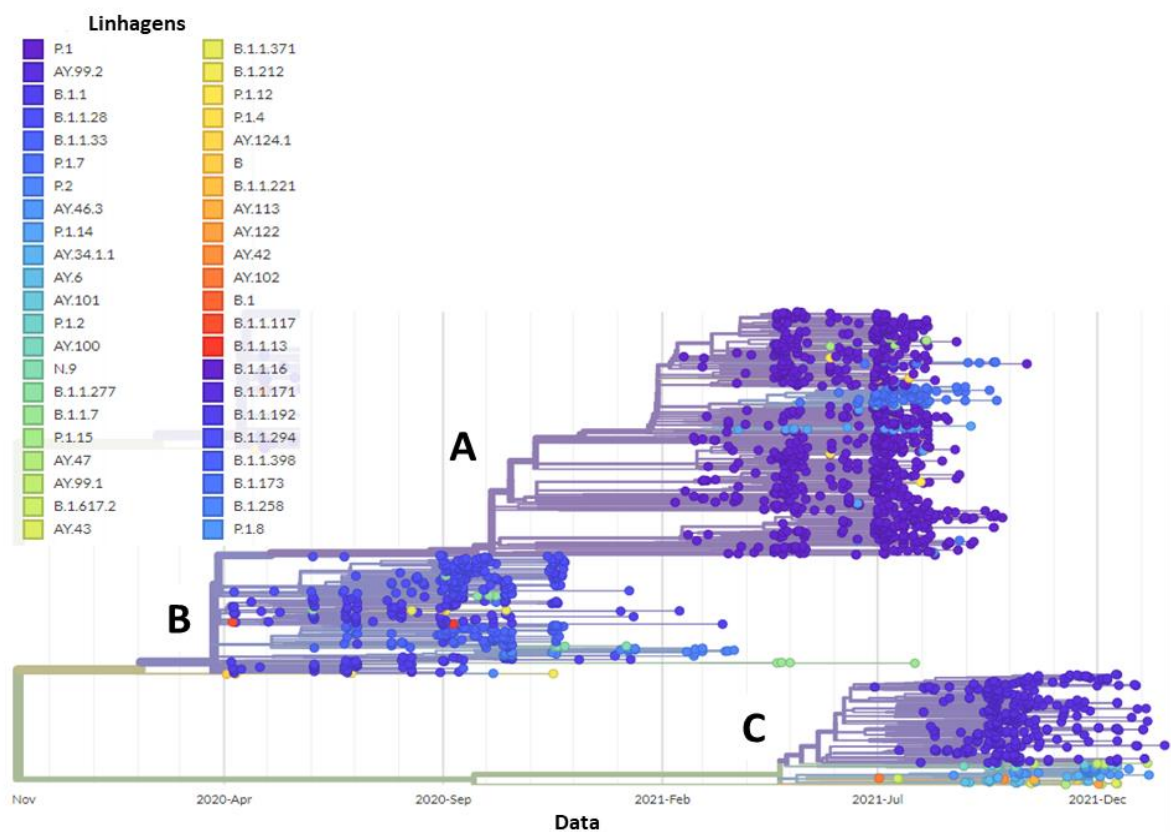


Fonte: Lima (2022)

4.4 ÁRVORE FILOGENÉTICA

A partir da análise da árvore filogenética obtida no nosso estudo, observamos a divisão das 1898 amostras em 3 principais clados (Figura 7). Esse achado encontra-se de acordo com estudos anteriores para amostras de SARS-CoV-2 oriundas de Pernambuco (BEZERRA JANUÁRIO, 2022). O clado A possui 978 cepas, formado majoritariamente pela linhagem P.1 e apresenta cepas pertencentes a 7 sublinhagens de P.1 (Figura 7A). Observamos o agrupamento das 63 cepas obtidas neste estudo no clado A, uma vez que foram atribuídas a linhagens P.1 e suas sublinhagens (P.1.7; P.1.14 e P.1.15). Além das principais mutações encontradas na Spike para linhagem P.1 (K417T, H655Y, T1027I, E484K, N501Y) (SANCHES et al., 2021), também identificamos no clado as mutações K1795Q na ORF1a, E92K na ORF8 e P80R na N.

Figura 7. Árvore filogenética de máxima verossimilhança gerada pelo pipeline do Nexstrain, composta por 1989 cepas de SARS-CoV-2 de diferentes regiões de Pernambuco. Cada amostra está colorida de acordo com a linhagem a qual foi classificada pelo PANGOLIN.

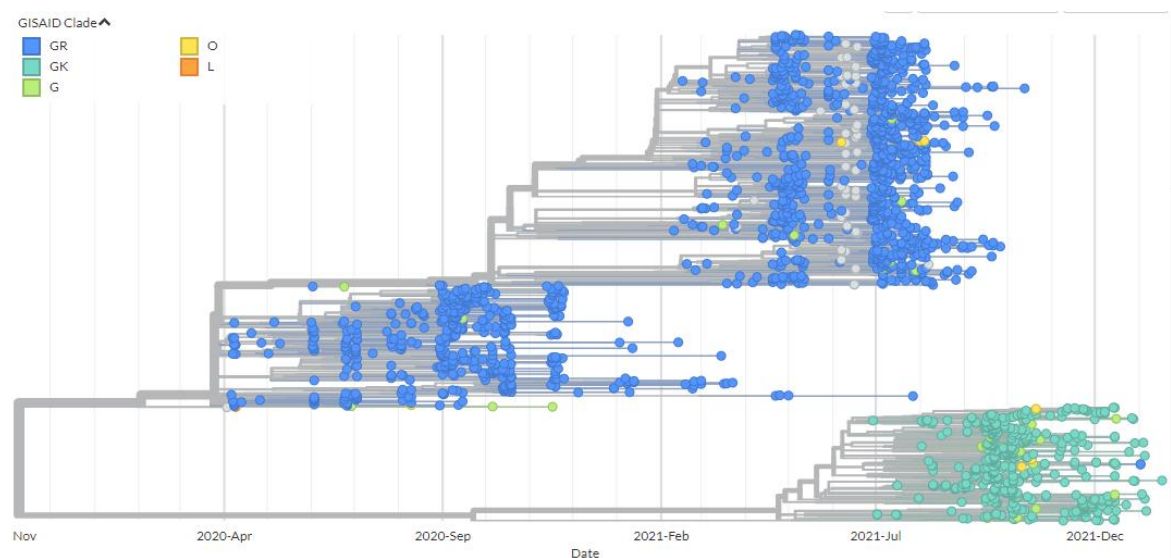


Fonte: Lima (2022)

O clado B é composto por 477 sequências e as linhagens observadas nesse clado são as encontradas com prevalência maior no início da epidemia em Pernambuco, dentre elas encontram-se a B.1.1, B.1.1.28 e a B.1.1.33. Também visualizamos as linhagens N.9, B.1.1.7 e a P.2 agrupadas nesse clado (Figura 7B). O clado C é composto por 443 cepas pertencentes a linhagem Delta (B.1.617.2) e 14 sublinhagens da mesma. A sublinhagem AY.99.2 foi a mais frequente e apresentou-se como um grupo monofilético (Figura 7C).

Quando avaliamos o agrupamento das sequências em relação aos clados definidos pelo GISAID, observamos 4 clados (excluindo o 5 clado, denominado de L pertencente à linhagem de referência) (Figura 8). O clado GR foi o mais frequente em Pernambuco, seguido pelo GK, G e o O. Esse achado é semelhante ao observamos em outros estudos, onde o clado GR foi o mais frequente dentre os observados (MERCATELLI; GIORGI, 2020; DE SOUZA et al., 2021)

Figura 8. Árvore filogenética de máxima verossimilhança gerada pelo pipeline do Nexstrain, composta por 1989 cepas de SARS-CoV-2 de diferentes regiões de Pernambuco. Cada amostra está colorida de acordo com o clado GISAID.



Fonte: Lima (2022)

5 CONCLUSÃO

De acordo com os dados obtidos em um período de quase dois anos de pandemia da COVID-19 em Pernambuco, observamos a circulação de 44 linhagens, sendo 3 variantes de preocupação. As variantes P.1 (Gama) e AY.99.2 (sublinhagem Delta) apresentaram um maior número de sequências e foram encontradas em todas as mesorregiões do estado de Pernambuco, demonstrando uma disseminação sustentada por todo o estado. Dentre as mesorregiões, a região metropolitana do Recife apresentou a maior diversidade de linhagens. Além disso, identificamos 3885 SNPs, sendo a maior parte de sua ocorrência em regiões codificantes do genoma viral e foram classificados como alterações missense e sinônimas de impacto baixo a moderado. As mutações ocorrem durante o processo de replicação viral e podem ou não conferir uma maior aptidão ao vírus. Por isso, reafirmamos a necessidade da vigilância genômica contínua a fim de detectar novas variantes e fornecer informações importantes no desenvolvimento de vacinas compatíveis com as cepas circulantes.

REFERÊNCIAS

ABDUL-RASOOL, Sahar; FIELDING, Burtram C. Entendendo o coronavírus humano HCoV-NL63. **The open virology journal** , v. 4, p. 76, 2010.

AL-ROHAIMI, Abdulmohsen H.; AL OTAIBI, Faisal. Novel SARS-CoV-2 outbreak and COVID19 disease; a systemic review on the global pandemic. **Genes & Diseases**, v. 7, n. 4, p. 491-501, 2020.

ALIZON, S., HAIM-BOUKOBZA, S., FOULONGNE, V., VERDURME, L., TROMBERT- PAOLANTONI, S., LECORCHE, E et al. Rapid spread of the SARS-CoV-2 Delta variant in some French regions, June 2021. **Euro Surveill.** v. 26, n. 28, p. 2100573, 2021.

ANDREWS, Simon et al. FastQC. **A quality control tool for high throughput sequence data**, v. 370, 2010.

BAI, Chongzhi; ZHONG, Qiming; GAO, George Fu. Overview of SARS-CoV-2 genome-encoded proteins. **Science China Life Sciences**, p. 1-15, 2021.

BEZERRA JANUÁRIO, C. A. Diversidade e estruturação genética do gene spike de SARS-COV-2 circulante no estado de Pernambuco / Caio Andrey Bezerra Januário - Recife, 2022.

BIANCHI, Martina et al. SARS-Cov-2 ORF3a: mutability and function. **International journal of biological macromolecules**, v. 170, p. 820-826, 2021.

CAMARGO, Carlos Henrique et al. Um ano de pandemia da COVID-19: diversidade genética do SARS-CoV-2 no Brasil. **BEPA. Boletim Epidemiológico Paulista**, v. 18, n. 207, p. 12-33, 2021.

CANDIDO, Darlan S. et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. **Science**, v. 369, n. 6508, p. 1255-1260, 2020.

CINGOLANI P, PLATTS A, WANG LE L, COON M, NGUYEN T, WANG L, LAND SJ, LU X, RUDEN DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. **Fly**, v. 6, n. 2, p. 80-92, 2012.

CHANG, Chung-ke et al. Recent insights into the development of therapeutics against coronavirus diseases by targeting N protein. **Drug discovery today**, v. 21, n. 4, p. 562-572, 2016.

CHAN, Jasper Fuk-Woo 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, 2020.

CHANNAPPANAVAR, Rudragouda; PERLMAN, Stanley. Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. In: **Seminars in immunopathology**, p. 529-539, 2017.

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

CLARO, Ingra Morales et al. Local Transmission of SARS-CoV-2 Lineage B. 1.1. 7, Brazil, December 2020. **Emerging infectious diseases**, v. 27, n. 3, p. 970, 2021.

CUI, Jie; LI, Fang; SHI, Zheng-Li. Origem e evolução dos coronavírus patogênicos. **Nature revisa microbiologia**, v. 17, n. 3, pág. 181-192, 2019.

DANECEK, P.; MCCARTHY, S A. BCFtools/csq: haplotype-aware variant consequences. **Bioinformatics**, v. 33, n. 13, p. 2037-2039, 2017.

DHAMA, K. et al. COVID-19, an emerging coronavirus infection: advances and prospects in designing and developing vaccines, immunotherapeutics, and therapeutics. **Human Vaccines & Immunotherapeutics**, v. 16, n. 6, p. 1232–1238, 2 jun. 2020.

DE HAAN, Cornelis AM; VENNEMA, Harry; ROTTIER, Peter JM. Assembly of the coronavirus envelope: homotypic interactions between the M proteins. **Journal of virology**, v. 74, n. 11, p. 4967-4978, 2000.

DEMOLINER, M., DA SILVA, M. S., GULARTE, J. S., HANSEN, A. W., DE ALMEIDA, P. R., WEBER, M. N. et al. Predominance of SARS-CoV-2 P. 1 (Gamma) lineage inducing the recent COVID-19 wave in southern Brazil and the finding of an additional S: D614A mutation. **Infection, Genetics and Evolution**, v. 96, p. 105134, 2021.

DE SOUZA, Ueric José Borges et al. High rate of mutational events in sars-cov-2 genomes across brazilian geographical regions, february 2020 to june 2021. **Viruses**, v. 13, n. 9, p. 1806, 2021.

FARIA, Nuno R. et al. Genomics and epidemiology of the P. 1 SARS-CoV-2 lineage in Manaus, Brazil. **Science**, v. 372, n. 6544, p. 815-821, 2021.

FEHR, Anthony R.; PERLMAN, Stanley. Coronaviruses: an overview of their replication and pathogenesis. **Coronaviruses**, p. 1-23, 2015.

FUNK, Tjede et al. Characteristics of SARS-CoV-2 variants of concern B. 1.1. 7, B. 1.351 or P. 1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. **Eurosurveillance**, v. 26, n. 16, p. 2100348, 2021.

FORNI, Diego et al. Molecular evolution of human coronavirus genomes. **Trends in microbiology**, v. 25, n. 1, p. 35-48, 2017.

GARRISON, Erik; MARTH, Gabor. Haplotype-based variant detection from short-read sequencing. **arXiv preprint arXiv:1207.3907**, 2012.

GHEBLAWI, Mahmoud et al. Angiotensin-converting enzyme 2: SARS-CoV-2

receptor and regulator of the renin-angiotensin system: celebrating the 20th anniversary of the discovery of ACE2. **Circulation research**, v. 126, n. 10, p. 1456-1474, 2020.

GORBALENYA, Alexander E. et al. Severe acute respiratory syndrome-related coronavirus: The species and its viruses—a statement of the Coronavirus Study Group. **BioRxiv**, 2020.

GRAHAM, R. L.; DONALDSON, E. F.; BARIC, R. S. A decade after SARS: strategies for controlling emerging coronaviruses. **Nature Reviews Microbiology**, v. 11, n. 12, p. 836– 848, dez. 2013.

GUAN, Yi et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. **Science**, v. 302, n. 5643, p. 276-278, 2003.

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.

GUREVICH, Alexey et al. QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072-1075, 2013.

HADFIELD J, MEGILL C, BELL SM, HUDDLESTON J, POTTER B, CALLENDER C, et al. Nextstrain: real-time tracking of pathogen evolution. **Bioinformatics**. 2018;34(23):4121-3.

ITHETE, Ndapewa Laudika et al. Close relative of human Middle East respiratory syndrome coronavirus in bat, South Africa. **Emerging infectious diseases**, v. 19, n. 10, p. 1697, 2013.

JOSHI, N.; FASS, J. Sickle-A windowed adaptive trimming tool for FASTQ files using quality. **Version 1.33**, 2011.

KANG, D., CHOI, H., KIM, J. H., & CHOI, J. Spatial epidemic dynamics of the COVID-19 outbreak in China. **International Journal of Infectious Diseases**, v. 94, p. 96-102, 2020.

KATOH K, ROZEWICKI J, YAMADA KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. **Briefings in bioinformatics**, v. 20, n. 4, p. 1160-1166, 2019.

KHATAMZAS, Elham et al. Emergence of multiple SARS-CoV-2 mutations in an immunocompromised host. **MedRxiv**, 2021.

KONINGS, F., PERKINS, M. D., KUHN, J. H., PALLAN, M. J., ALM, E. J., ARCHER, B. N., e VAN KERKHOVE, M. D. SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. **Nature microbiology**, v. 6, n. 7, p. 821-823, 2021.

KORBER, Bette et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. **Cell**, v. 182, n. 4, p. 812-827. e19, 2020.

KSIAZEK, Thomas G. et al. A novel coronavirus associated with severe acute respiratory syndrome. **New England journal of medicine**, v. 348, n. 20, p. 1953-1966, 2003.

LAAMARTI, Meriem et al. Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations. **PloS one**, v. 15, n. 11, p. e0240345, 2020.

LANGMEAD, Ben; SALZBERG, Steven L. Fast gapped-read alignment with Bowtie 2. **Nature methods**, v. 9, n. 4, p. 357-359, 2012.

LIANG, Yanwen et al. Highlight of immune pathogenic response and hematopathologic effect in SARS-CoV, MERS-CoV, and SARS-Cov-2 infection. **Frontiers in immunology**, v. 11, p. 1022, 2020.

LI, Wendong et al. Bats are natural reservoirs of SARS-like coronaviruses. **Science**, v. 310, n. 5748, p. 676-679, 2005.

LI, Heng; DURBIN, Richard. Fast and accurate long-read alignment with Burrows–Wheeler transform. **Bioinformatics**, v. 26, n. 5, p. 589-595, 2010.

LI H, HANDSAKER B, WYSOKER A, FENNEL T, RUAN J, HOMER N, et al. The sequence alignment/map format and SAMtools. **Bioinformatics**. 25:16, 2078–9. 2009.

MASTERS, Paul S. The molecular biology of coronaviruses. **Advances in virus research**, v. 66, p. 193-292, 2006.

MAJUMDAR, Parinita; NIYOGI, Sougata. ORF3a mutation associated with higher mortality rate in SARS-CoV-2 infection. **Epidemiology & Infection**, v. 148, 2020.

MERCATELLI, Daniele; GIORGI, Federico M. Geographic and genomic distribution of SARS-CoV-2 mutations. **Frontiers in microbiology**, v. 11, p. 1800, 2020.

NGUYEN, L. T., SCHMIDT, H. A., VON HAESELER, A., & MINH, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. **Molecular biology and evolution**, v. 32, n. 1, p. 268-274, 2015.

PAULES, Catharine I.; MARSTON, Hilary D.; FAUCI, Anthony S. Coronavirus infections—more than just the common cold. **Jama**, v. 323, n. 8, p. 707-708, 2020.

PACHETTI, Maria et al. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. **Journal of translational medicine**, v. 18, n. 1, p. 1-9, 2020.

PAIVA, M. H. S., GUEDES, D. R. D., DOCENA, C., BEZERRA, M. F., DEZORDI, F. Z., MACHADO, L. C. e WALLAU, G. L. Multiple introductions followed by ongoing

community spread of SARS-CoV-2 at one of the largest metropolitan areas of Northeast Brazil. **Viruses**, v. 12, n. 12, p. 1414, 2020.

PEKAR, J. et al. Timing the SARS-CoV-2 index case in Hubei province. **Science**, v. 372, n. 6540, p. 412-417, 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.

RAHIMI, Azadeh; MIRZAZADEH, Azin; TAVAKOLPOUR, Soheil. Genetics and genomics of SARS-CoV-2: A review of the literature with the special focus on genetic diversity and SARS-CoV-2 genome detection. **Genomics**, v. 113, n. 1, p. 1221-1232, 2021.

RESENDE, P. C. et al. Evolutionary dynamics and dissemination pattern of the SARS-CoV-2 lineage B. 1.1. 33 during the early pandemic phase in Brazil. **Frontiers in microbiology**, v. 11, 2020.

RODRIGUEZ-MORALES, A. J., Gallego, V., Escalera-Antezana, J. P., Mendez, C. A., Zambrano, L. I., Franco Paredes, C., Cimerman, S. COVID-19 in Latin America: The implications of the first confirmed case in Brazil. **Travel Medicine and Infectious Disease**, v. 35, p. 101613, 2020.

ROMERO, P. E., DÁVILA-BARCLAY, A., SALVATIERRA, G., GONZÁLEZ, L., CUICAPUZA, D., SOLÍS, L. et al. The emergence of SARS-CoV-2 variant lambda (C. 37) in South America. **Microbiology spectrum**, v. 9, n. 2, p. e00789-21, 2021.

SAGULENKO, P.; PULLER, V.; NEHER, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. **Virus evolution**, v. 4, n. 1, p. vex042, 2018.

SANCHES, Paulo RS et al. Recent advances in SARS-CoV-2 Spike protein and RBD mutations comparison between new variants Alpha (B. 1.1. 7, United Kingdom), Beta (B. 1.351, South Africa), Gamma (P. 1, Brazil) and Delta (B. 1.617. 2, India). **Journal of virus eradication**, v. 7, n. 3, p. 100054, 2021.

SHANG, Weilong et al. The outbreak of SARS-CoV-2 pneumonia calls for viral vaccines. **npj Vaccines**, v. 5, n. 1, p. 1-3, 2020.

SILVEIRA, Z. S. Diversidade genômica de cepas de SARS-CoV-2 durante a primeira onda da Covid19 no Estado de Pernambuco, Brasil / Zildene de Sousa Silveira.2022.

SYED, A. M., CILING, A., TAHA, T. Y., CHEN, I. P., KHALID, M. M., SREEKUMAR, B., ... & DOUDNA, J. A. Omicron mutations enhance infectivity and reduce antibody neutralization of SARS-CoV-2 virus-like particles. **Proceedings of the National Academy of Sciences**, v. 119, n. 31, p. e2200592119, 2022.

VOLZ, Erik et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. **Cell**, v. 184, n. 1, p. 64-75. e11, 2021.

WALENSKY, ROCHELLE P.; WALKE, HENRY T.; FAUCI, ANTHONY S. SARS-CoV-2 variants of concern in the United States—challenges and opportunities. **Jama**, v. 325, n. 11, p. 1037-1038, 2021.

WOO, P. C. et al. Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. **Journal of virology**, v. 86, n. 7, p.3995-4008. 2012.

YANG, Wan-Ting et al. SARS-CoV-2 E484K Mutation Narrative Review: Epidemiology, Immune Escape, Clinical Implications, and Future Considerations. **Infection and Drug Resistance**, v. 15, p. 373, 2022.

ZAKI, Ali M. et al. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. **New England Journal of Medicine**, v. 367, n. 19, p. 1814-1820, 2012.

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

APÊNDICE A

Frequência das linhagens classificadas pelo PANGOLIN para os 1898 genomas de
SARS-CoV-2 circulantes em Pernambuco.

Linhagens	Frequência
P.1	852
AY.99.2	356
B.1.1	169
B.1.1.28	162
B.1.1.33	84
P.1.7	79
AY.46.3	32
P.2	30
P.1.14	25
AY.6	12
AY.34.1.1	11
AY.101	8
P.1.15	7
P.1.2	6
AY.100	6
B.1.1.277	5
N.9	5
P.1.12	5
B.1.1.7	4
AY.99.1	4
B.1.1.371	3
P.1.4	3
B.1.617.2	3
AY.43	3
B.1.212	2
B.1.1.221	2
AY.124.1	2
AY.47	2

B.1.1.117	1
B	1
B.1	1
B.1.1.192	1
B.1.1.171	1
B.1.1.16	1
B.1.1.13	1
B.1.1.398	1
B.1.258	1
B.1.173	1
B.1.1.294	1
AY.102	1
P.1.8	1
AY.42	1
AY.113	1
AY.122	1

APÊNDICE B

Mutações identificadas com frequência maior que 50% entre os 1898 genomas analisados neste estudo.

Alteração nucleotídica	Frequência %	Tipo da mutação	Impacto	Troca de aminoácido	Gene
C241T	99,21	Não codificante	Modificadora	NA	5' UTR
C3037T	99,79	sinônima	Baixo	F105F	ORF1ab
C14408T	99,47	missense	Moderado	P323L	ORF1ab
T733C	51,42	sinônima	Baixo	D156D	ORF1ab
C2749T	51,26	sinônima	Baixo	D828D	ORF1ab
C3828T	50,68	missense	Moderado	S1188L	ORF1ab
A5648C	51,37	missense	Moderado	K1795Q	ORF1ab
A6319G	51,48	sinônima	Baixo	P2018P	ORF1ab
A6613G	51,32	sinônima	Baixo	V2116V	ORF1ab
C12778T	51,26	sinônima	Baixo	Y4171Y	ORF1ab
C13860T	51,42	sinônima	Baixo	T4532I	ORF1ab
G17259T	51,00	missense	Moderado	S5665I	ORF1ab
C21614T	51,90	missense	Moderado	L18F	S
C21621A	51,26	missense	Moderado	T20N	S
C21638T	51,48	missense	Moderado	P26S	S

G21974T	51,00	missense	Moderado	D138Y	S
G22132T	51,26	missense	Moderado	R190S	S
A22812C	51,32	missense	Moderado	K417T	S
G23012A	51,26	missense	Moderado	E484K	S
A23403G	99,89	missense	Moderado	D614G	S
C23525T	51,42	missense	Moderado	H655Y	S
C24642T	51,37	missense	Moderado	T1027I	S
G25088T	61,96	missense	Moderado	V1176F	S
T26149C	51,53	missense	Moderado	S253P	ORF3a
G28167A	51,37	missense	Moderado	E92K	ORF8
C28512G	51,11	missense	Moderado	P80R	N
A28877T	51,21	missense	Moderado	S202C	N
G28878C	51,21	missense	Moderado	S202T	N
G28881A	75,97	missense	Moderado	R203K	N
G28882A	76,08	sinônima	Baixo	R203R	N
G28883C	76,03	missense	Moderado	G204R	N