Universidade Federal do Rio Grande do Sul

Centro de Biotecnologia

Programa de Pós-Graduação em Biologia Celular e Molecular

Caracterização genômica do *Betacoronavirus* SARS-CoV-2 para compreensão da distribuição de linhagens e padrões de espalhamento geográfico no estado do Rio Grande do Sul e no território brasileiro

Dissertação de Mestrado

Vinícius Bonetti Franceschi

Porto Alegre, 07 de fevereiro de 2022

Universidade Federal do Rio Grande do Sul

Centro de Biotecnologia

Programa de Pós-Graduação em Biologia Celular e Molecular

Caracterização genômica do *Betacoronavirus* SARS-CoV-2 para compreensão da distribuição de linhagens e padrões de espalhamento geográfico no estado do Rio Grande do Sul e no território brasileiro

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS como requisito parcial para obtenção do grau de Mestre.

Vinícius Bonetti Franceschi

Profa. Dra. Claudia Elizabeth Thompson – Orientadora

Profa. Dra. Gabriela Bettella Cybis – Coorientadora

Porto Alegre, 07 de fevereiro de 2022

Vinícius Bonetti Franceschi

Caracterização genômica do *Betacoronavirus* SARS-CoV-2 para compreensão da distribuição de linhagens e padrões de espalhamento geográfico no estado do Rio Grande do Sul e no território brasileiro

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS como requisito parcial para obtenção do grau de Mestre.

Data da avaliação: __/__/____

Conceito: _____

BANCA EXAMINADORA

_____
Profa. Dra. Claudia Elizabeth Thompson
Orientadora — UFCSPA

_____
Prof. Dr. Augusto Schrank
Examinador Interno — UFRGS

_____
Prof. Dr. Luís Fernando Saraiva Macedo Timmers
Examinador Externo — UNIVATES

_____
Dra. Tatiana Schäffer Gregianini
Examinador Externo — Secretaria da Saúde do RS

Porto Alegre
2022

Este trabalho foi desenvolvido na modalidade remota devido às restrições impostas pela pandemia de COVID-19 e contou com o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Prefeitura Municipal de Esteio e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – Chamada CNPq/AWS Nº 032/2019 – Acesso às Plataformas de Computação em Nuvem da AWS – Cloud Credits for Research).

Dedico esta, bem como todas as minhas demais conquistas, aos meus amados pais e familiares, à minha namorada, às minhas orientadoras, colegas do grupo de pesquisa e a todos os meus professores, que foram primordiais para a conclusão desta importante etapa da minha vida. Também me solidarizo e deixo minhas profundas condolências a todas as pessoas que tiveram entes queridos perdidos durante a pandemia de COVID-19, objeto de estudo deste trabalho.

AGRADECIMENTOS

Agradeço especialmente às seguintes pessoas, que foram primordiais para que esta dissertação de mestrado fosse finalizada:

Aos meus pais, Valdecir e Jaqueline, que sempre batalharam para me oferecer uma educação qualificada, repassar seus valores corretos, dar amor e carinho incondicional.

À minha namorada Andressa, que como pesquisadora e biomédica contribuiu com seu conhecimento em pesquisa, além de me fazer muito feliz em todos os momentos e me apoiar incansavelmente nesses dois anos difíceis de pandemia.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da minha bolsa de estudos. Ao coordenador do Programa de Pós Graduação de Biologia Celular e Molecular (PPGBCM), Prof. Dr. Hugo Verli, à sempre querida e competente Silvia Centeno e ao corpo docente do PPGBCM, que não mediram esforços para esclarecer dúvidas, planejar as atividades e disciplinas à distância, sempre prezando pela excelência e qualidade do PPGBCM.

Aos membros da comissão de acompanhamento do PPGBCM, Prof. Dr. Arnaldo Zaha e Prof. Dr. Augusto Schrank, pelas sugestões apresentadas durante as reuniões de acompanhamento, além da participação na correção da atividade de redação científica e na revisão da presente dissertação, respectivamente.

Aos membros da banca examinadora desta dissertação, Prof. Dr. Augusto Schrank, Prof. Dr. Luís Fernando Saraiva Macedo Timmers, Dra. Tatiana Schäffer Gregianini e Dra. Gabriela Prado Paludo, pelo tempo dispendido na leitura desta dissertação e pelos pertinentes apontamentos e sugestões levantadas para melhoria do trabalho desenvolvido.

Aos colegas da Unidade de Biologia Teórica e Computacional (UBTEC) da UFRGS: Andrea Tavanti, Fabio Andreis, Meiski Vedovatto, Renato Corá e Rodrigo Streit, que me motivaram a ingressar no PPGBCM após meu estágio curricular pelo ótimo ambiente do laboratório. Infelizmente, nos vimos apenas virtualmente durante a pandemia, mas essas interações foram muito importantes para mim.

Aos meus colegas do grupo de pesquisa, especialmente Patrícia Ferrareze, Gabriel Caldana e Amanda Mayer, pelo auxílio na coleta de amostras e metadados

clínicos, nas análises de bioinformática e na escrita dos manuscritos integrantes desta dissertação. Importante destacar a resiliência, força de vontade e criatividade do grupo para produzir conhecimento sobre a pandemia mesmo com os poucos incentivos financeiros recebidos.

À prefeitura de Esteio e demais financiadores da iniciativa privada pelo financiamento da pesquisa. Aos hospitais e laboratórios que contribuíram com a coleta, processamento e sequenciamento das amostras utilizadas neste trabalho, bem como aos participantes que consentiram em participar dessa investigação genômica e epidemiológica, contribuindo para a compreensão da doença e para o avanço técnico-científico.

Aos pesquisadores brasileiros e internacionais que estabeleceram uma rede de colaboração sem precedentes para compartilhamento de dados genômicos em tempo real, permitindo o acompanhamento da evolução do vírus SARS-CoV-2 e a identificação de novas variantes, guiando ações de saúde pública.

À minha coorientadora, Profa. Dra. Gabriela Cybis, que, mesmo em período de licença maternidade, contribuiu ativamente para a realização das análises filodinâmicas Bayesianas, levantando as limitações e as possíveis interpretações dos resultados obtidos. Também me motivou a entender mais sobre os métodos utilizados e a expandir meu conhecimento nessa área antes desconhecida.

À minha orientadora, Profa. Dra. Claudia Thompson, que despertou meu interesse pela bioinformática ainda quando aluno de graduação em 2018 e ajudou a consolidar os conhecimentos teóricos (por meio de suas aulas) e práticos (durante meu Trabalho de Conclusão de Curso e Dissertação de Mestrado) na área até o presente momento. Durante a pandemia, dedicou-se diuturnamente para compreender a epidemiologia da COVID-19 no município de Esteio e na orientação das análises e escrita dos trabalhos genômicos aqui realizados. Agradeço imensamente pela sua contribuição para a minha formação como Informata Biomédico e como Mestre em Biologia Celular e Molecular.

“Quanto mais diversificados os descendentes de uma espécie se tornarem em estrutura, constituição e hábitos, na mesma medida eles estarão mais capacitados para aproveitar lugares numerosos e amplamente diversificados no estado de natureza, e assim mais capacitados para crescer em número.”

Charles Darwin, A Origem das Espécies, 1859

RESUMO

Em dezembro de 2019, um novo coronavírus foi detectado em pacientes com síndrome respiratória aguda grave em Wuhan, China. Este *Betacoronavirus*, denominado Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), espalhou-se rapidamente pelo mundo, de modo que a Organização Mundial da Saúde (OMS) declarou estado de pandemia em março de 2020. Após o sequenciamento do primeiro genoma completo do vírus, esforços internacionais sem precedentes foram estabelecidos por meio do banco de dados GISAID, permitindo o acompanhamento em tempo real da evolução viral, bem como seu espalhamento geográfico em níveis locais, regionais, nacionais e globais. Nesse sentido, este trabalho objetivou realizar análises genômicas de amostras isoladas de SARS-CoV-2 a fim de compreender a distribuição de mutações e linhagens virais em nível municipal (Esteio, Rio Grande do Sul [RS], Brasil), estadual (RS) e nacional (Brasil). Para este fim, sequenciamos 21 amostras do município de Esteio na primeira fase da epidemia (maio a outubro de 2020), demonstrando a presença principal das linhagens B.1.1.28 e B.1.1.33, a caracterização inicial da linhagem P.2 no estado e a contribuição principal da região Sudeste para a difusão destas linhagens para o sul do Brasil. Subsequentemente, analisamos 56 genomas de 13 municípios do RS em período de aumento de hospitalizações e mortes (março de 2021), demonstrando a rápida difusão da variante P.1 (Gama) para o estado a partir de múltiplas introduções vindas principalmente do Norte, bem como descrevendo as mutações e a difusão geográfica da sublinhagem P.1.2. Finalmente, utilizamos 2.732 genomas de todo o território brasileiro no primeiro ano da epidemia (entre fevereiro de 2020 e 2021), descrevendo esforços de sequenciamento heterogêneos temporal e espacialmente, a rápida substituição das linhagens B.1.1.28 e B.1.1.33 por P.1 e P.2 e complexos padrões filogeográficos, nos quais algumas linhagens se espalham principalmente de modo intra-estadual e, outras, interestadual. Portanto, ao utilizar dados epidemiológicos e genomas completos do SARS-CoV-2 dos pacientes locais e de um conjunto representativo da diversidade viral mundial, caracterizamos as mutações virais observadas, a abundância de linhagens, bem como compreendemos padrões de espalhamento geográfico no território Brasileiro.

ABSTRACT

In December 2019, a novel coronavirus was detected in patients with severe acute respiratory syndrome in Wuhan, China. This *Betacoronavirus*, named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has spread rapidly around the world leading the World Health Organization (WHO) to declare a pandemic state in March 2020. Following the sequencing of the virus' first complete genome, unprecedented international efforts have been established through the GISAID database, allowing real-time tracking of viral evolution as well as its geographic spread at local, regional, national, and global levels. In this context, this work aimed to perform genomic analyses of SARS-CoV-2 samples in order to understand the distribution of mutations and viral lineages at municipal (Esteio, Rio Grande do Sul [RS], Brazil), state (RS) and national (Brazil) levels. To this end, we sequenced 21 samples from the municipality of Esteio in the first epidemic phase (May to October 2020), demonstrating the prominent presence of the B.1.1.28 and B.1.1.33 lineages, the initial characterization of the P.2 lineage in the state, and the major contribution of the Southeast region to the spread of these strains to Southern Brazil. Subsequently, we analyzed 56 genomes from 13 municipalities in RS during a period of increased hospitalizations and deaths (March 2021), demonstrating the rapid diffusion of the P.1 (Gamma) variant into the state from multiple introductions mainly from the Northern region of Brazil, as well as describing the mutations and geographic diffusion of the P.1.2. Finally, we used 2732 genomes from across Brazil in the first year of the epidemic (between February 2020 and 2021), describing temporally and spatially heterogeneous sequencing efforts, the rapid replacement of the B.1.1.28 and B.1.1.33 lineages for P.1 and P.2, and complex phylogeographic patterns in which some lineages spread primarily intrastate and others interstate. Therefore, by using epidemiological data and complete SARS-CoV-2 genomes from local patients and a set of genomes representative of worldwide viral diversity, we characterized viral mutations, lineage abundance, as well as understood patterns of geographic spread in the Brazilian territory.

LISTA DE ABREVIATURAS, SÍMBOLOS E UNIDADES

CoVs	Coronavírus
DNA	Ácido desoxirribonucleico
dN / dS	Razão das taxas de substituições não-sinônimas por sinônimas
E	Proteína do envelope
ERGIC	Complexo intermediário Retículo Endoplasmático-Golgi
GISAID	Iniciativa Global para o Compartilhamento de Todos os Dados sobre Influenza
GTR	Modelo geral reversível no tempo
hACE2	Enzima conversora de Angiotensina 2 humana
HKY	Modelo de Hasegawa-Kishino-Yano
IFN	Interferon
kb	kilobases
M	Proteína da membrana
MERS-CoV	Middle East respiratory syndrome coronavirus
ML	Máxima Verossimilhança
MP	Máxima Parcimônia
MRCA	Ancestral comum mais recente
mRNAs	RNAs mensageiros
N	Proteína do nucleocapsídeo
NGS	Sequenciamento de nova geração
nsp	Proteínas não estruturais
NTD	Domínio N-terminal
OMS	Organização Mundial da Saúde
ORFs	Quadros de leitura abertos
R_0	Número básico de reprodução

RBD	Domínio de ligação ao receptor
RE	Retículo Endoplasmático
RNA	Ácido ribonucleico
RS	Rio Grande do Sul
RTC	Complexo replicase-transcriptase
S	Proteína *Spike*
SARS	Síndrome respiratória aguda grave
SARS-CoV	Severe acute respiratory syndrome coronavirus
+ ssRNA	RNA de sentido positivo de fita simples
TMPRRS2	Serinoprotease transmembranar 2
TRS	Sequências reguladoras da transcrição
UTR	Região não traduzida
VOC	Variante de preocupação
VOI	Variante de interesse

LISTA DE FIGURAS

Figura 6. Princípios dos métodos de relógio molecular e filogeografia. (a) Filogenias moleculares enraizadas podem ser estimadas a partir de sequências de genes ou genomas virais. Esta filogenia não tem escala de tempo, portanto o comprimento do ramo representa a divergência genética do ancestral em substituições por sítio (círculo preto). (b) A mesma filogenia também pode ser reconstruída usando um modelo de relógio molecular, que define uma relação entre distância genética e tempo. Nesse caso, as sequências foram amostradas em pontos de tempo conhecidos e os ramos da filogenia têm comprimentos em unidades de anos, permitindo a estimativa da idade dos eventos de ramificação. (c) Os dados filodinâmicos também podem demonstrar a evolução das mutações ao longo do tempo. (d) Sequências virais também podem ser analisadas utilizando filogeografia temporal. No exemplo, as nove sequências foram amostradas na França (verde, A), no Reino Unido (azul, B) e em dois locais na Espanha (vermelho, C1 e C2). Métodos estatísticos podem ser usados para reconstruir o histórico de propagação de patógenos, de modo que cada ramo seja rotulado com sua posição geográfica estimada. (e) Os princípios de análises coalescentes, que incorporam um modelo explícito da população amostrada. Cada círculo representa uma infecção e os círculos na mesma linha ocorrem durante o mesmo período de tempo. A largura crescente de cada fileira reflete o crescimento da epidemia ao longo do tempo. A partir das infecções amostradas (vermelho), as linhagens amostradas (linhas pretas) podem ser rastreadas por meio de infecções não amostradas (cinza) até o ancestral comum (círculo preto). A taxa na qual as linhagens amostradas coalescem depende de processos populacionais como dinâmica e estrutura populacional e seleção natural.

Figura 7. Análise filogeográfica e filodinâmica da epidemia do vírus Ebola na África Ocidental (2013-2016), abrangendo a estimativa simultânea dos dados de sequência e informações geográficas. Os gráficos mostram uma fotografia do espalhamento geográfico e uma árvore de credibilidade máxima de clado (MCC) que sumariza os resultados de maior probabilidade posterior da inferência Bayesiana.

Figura 8. Mapa do Brasil, evidenciando as cinco regiões brasileiras e os 26 estados brasileiros somados ao Distrito Federal.

Figura 9. Mutações na proteína *Spike* do SARS-CoV-2 compartilhadas entre as VOCs (rotuladas em vermelho) e VOIs (em amarelo). Linhas representam as diferentes linhagens e colunas mostram as substituições. O gradiente de cores iniciando em rosa claro (0%) até roxo escuro (100%) indica a frequência da substituição dentro da linhagem.

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1. CLASSIFICAÇÃO E ORGANIZAÇÃO GENÔMICA DOS CORONAVÍRUS

Os coronavírus (CoVs) são vírus envelopados[1], não segmentados[2], cujo genoma é composto por um RNA de sentido positivo de fita simples (+ ssRNA). Pertencem à ordem Nidovirales, família Coronaviridae e subfamília Orthocoronavirinae. Essa subfamília inclui quatro gêneros: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* e *Deltacoronavirus*, inicialmente classificados com base em dados sorológicos[3], mas atualmente divididos por inferências filogenéticas[4] (CHEN; LIU; GUO, 2020; FEHR; PERLMAN, 2015).

Em nível microscópico, a partícula viral pode ser considerada esférica (70 a 120 nanômetros de diâmetro), carregando proteínas estendidas na superfície da membrana (glicoproteínas[5] em forma de espigão) que fornecem a estrutura típica de coroa vista por microscopia eletrônica (GRAHAM; DONALDSON; BARIC, 2013; MASTERS, 2006; PYRC; BERKHOUT; HOEK, 2007) (Figura 1). Em nível molecular, os coronavírus empregam uma variedade de estratégias incomuns para realizar um processo complexo de expressão gênica (MASTERS, 2006).

Todos os vírus da ordem Nidovirales contêm genomas muito grandes em comparação aos demais vírus de RNA, sendo que os integrantes da família Coronavirinae possuem os maiores genomas de RNA identificados, variando de 26 a 32 kilobases (kb). Outras características comuns à ordem Nidovirales incluem: (i)

[1] Vírus recobertos por uma camada de lipídeos (gorduras).

[2] Genoma composto por apenas um fragmento de RNA.

[3] Estudos ou exames diagnósticos do soro sanguíneo, especialmente em relação à resposta do sistema imunológico a patógenos.

[4] Estudo das relações evolutivas históricas entre grupos de organismos.

[5] Proteínas que contêm cadeias de oligossacarídeos (açúcares) covalentemente ligadas a cadeias laterais de aminoácidos.

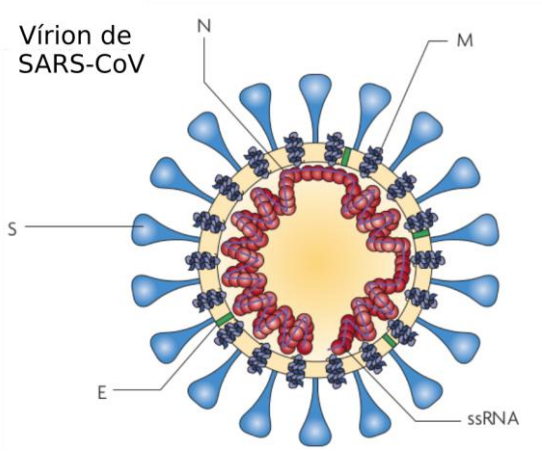

Figura 1. Esquema representando as principais estruturas do vírion (partícula viral infecciosa) do coronavírus SARS-CoV, envolvido na epidemia de 2002-2004.
E: Envelope; M: Membrana; N: Nucleocapsídeo; S: *Spike*; ssRNA: RNA de fita simples.
Fonte: PERLMAN & NETLAND (2009).

organização genômica altamente conservada, com um grande gene de replicase precedendo genes estruturais e acessórios; (ii) expressão de genes não estruturais por mudança de quadro ribossômico (*frameshifting*); (iii) atividades enzimáticas únicas codificadas na grande poliproteína[6] replicase-transcriptase; e (iv) expressão de genes a jusante por síntese de RNAs mensageiros (mRNAs) sub-genômicos aninhados na região 3' (FEHR; PERLMAN, 2015; MASTERS, 2006). As principais diferenças existentes entre as diversas espécies desta ordem estão no número, tipo e tamanho das proteínas estruturais, as quais causam alterações significativas na estrutura e morfologia dos nucleocapsídeos e vírions (FEHR; PERLMAN, 2015).

O genoma dos CoVs contém uma região 5'-cap e uma cauda 3'-poliA, permitindo que atue como um RNA mensageiro (mRNA) para a tradução das poliproteínas de replicação (replicases). O gene da replicase possui os grandes quadros de leitura abertos (ORFs) 1a e 1b, que codificam 16 proteínas não estruturais (nsp1- 16) necessárias para a replicação do RNA, ocupando dois terços

[6] Grande proteína que é clivada em várias proteínas menores com diferentes funções biológicas.

do genoma (cerca de 20 kb), enquanto as proteínas estruturais e acessórias correspondem a cerca de 10 kb do genoma viral (FEHR; PERLMAN, 2015; PEACOCK et al., 2021a; PYRC; BERKHOUT; HOEK, 2007). Na extremidade 5' do genoma há uma sequência líder e uma região não traduzida (UTR) que contém *stem-loops* necessários para replicação e transcrição de RNA. Além disso, no início de cada gene estrutural ou acessório, existem sequências reguladoras da transcrição (TRSs) necessárias para a expressão de cada um desses genes. A região 3' UTR também contém estruturas de RNA necessárias para replicação e síntese de RNA viral. Sendo assim, a organização geral do genoma dos CoVs é: 5'—sequência líder—5'UTR—replicase— proteínas S (*Spike*)–E (Envelope)—M (Membrana)—N (Nucleocapsídeo)— 3'UTR—cauda poliA, com genes acessórios intercalados dentro dos genes estruturais na extremidade 3' do genoma (FEHR; PERLMAN, 2015; PERLMAN; NETLAND, 2009) (Figura 2).

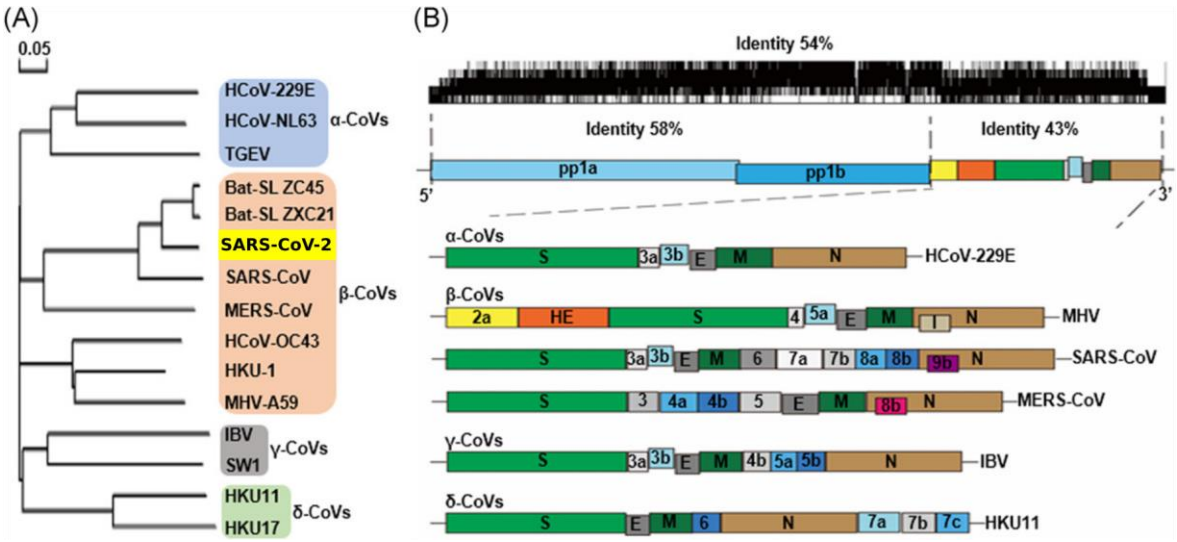

Figura 2. Organização genômica e árvore filogenética dos coronavírus. (A) Árvore filogenética dos CoVs representativos, com o novo coronavírus SARS-CoV-2 destacado em amarelo e os quatro diferentes gêneros evidenciados. (B) Estrutura do genoma de quatro gêneros de coronavírus. Pp1a e pp1b representam os dois polipeptídeos longos que são processados em 16 proteínas não estruturais. S, E, M e N indicam as quatro proteínas estruturais *Spike*, envelope, membrana e nucleocapsídeo.

Fonte: Adaptado de CHEN, LIU & GUO (2020).

Muitos dos nsps parecem ter múltiplas funções na síntese ou processamento de RNA viral ou em interações patógeno-hospedeiro, com o intuito de criar um ambiente ideal para sua replicação por meio da facilitação da entrada viral, da expressão gênica, da síntese de RNA ou da liberação do vírus (CHEN; LIU; GUO, 2020; DE WILDE et al., 2018).

Apesar da ordem dos genes estruturais essenciais ser notavelmente bem conservada, o mesmo não é verdadeiro para a proteína S, responsável pela ligação e entrada das células hospedeiras utilizando receptores celulares específicos. O domínio de ligação ao receptor (RBD) desta glicoproteína é pouco conservado entre os vírus e, consequentemente, o uso do receptor do hospedeiro varia entre gêneros e espécies virais (FEHR; PERLMAN, 2015; GRAHAM; DONALDSON; BARIC, 2013; PERLMAN; NETLAND, 2009) (Tabela 1). A proteína N é importante para a encapsulamento do RNA viral, desempenhando um papel fundamental durante a auto-montagem viral (CHANG et al., 2014) e atuando como um antagonista do interferon (IFN) (PERLMAN; NETLAND, 2009). Também está dinamicamente associado a complexos de replicação-transcrição, facilitando a síntese do mRNA subgenômico viral (VERHEIJE et al., 2010). A proteína E apresenta papel na morfogênese[7], montagem e brotamento, sendo sua ausência condicionante para a inibição completa — no caso do vírus da gastroenterite transmissível (ORTEGO et al., 2007) — ou parcial — no caso do SARS-CoV (DEDIEGO et al., 2007) — da liberação do vírus. Adicionalmente, possui atividade de canal iônico, que é necessária para uma replicação otimizada do vírus (PERLMAN; NETLAND, 2009;

[7] Processo biológico que permite o desenvolvimento de forma para um organismo.

Tabela 1. Receptores celulares de hospedeiros utilizados por diferentes coronavírus

Vírus	Gênero	Receptor
Human coronavirus 229E (HCoV-229E)	*Alphacoronavirus*	APN
Feline coronavirus (FCoV)	*Alphacoronavirus*	APN
Transmissible gastroenteritis virus (TGEV)	*Alphacoronavirus*	APN
Canine coronavirus (CCoV)	*Alphacoronavirus*	APN
Bat coronaviruses (BCoVs)	*Alphacoronavirus*	Desconhecido
Human coronavirus NL63 (HCoV-NL63)	*Alphacoronavirus*	ACE2
Murine hepatitis virus (MHV)	*Betacoronavirus*	CEACAM1a
Severe acute respiratory syndrome coronavirus (SARS-CoV)	*Betacoronavirus*	ACE2
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)	*Betacoronavirus*	ACE2
Bat SARS-related coronavirus (Bat-SCoV)	*Betacoronavirus*	ACE2?
Middle East respiratory syndrome coronavirus (MERS-CoV)	*Betacoronavirus*	DPP4
Human coronavirus OC43 (HCoV-OC43)	*Betacoronavirus*	Desconhecido
Avian infectious bronchitis virus (IBV)	*Gammacoronavirus*	Desconhecido
Bird coronaviruses	*Deltacoronavirus*	Desconhecido

Abreviações. APN: Aminopeptidase N; ACE2: Enzima Conversora de Angiotensina 2; CEACAM1a: *Carcinoembryonic antigen-related cell adhesion molecule 1*; DPP4: *Dipeptidyl peptidase 4*.
Fonte: Adaptado de FEHR & PERLMAN (2015); GRAHAM, DONALDSON & BARIC (2013); PERLMAN & NETLAND (2009).

WILSON et al., 2004).

O alinhamento das sequências dos genomas de CoVs mostra 58% de identidade na região codificadora de nsps e 43% de identidade na região codificadora de proteínas estruturais entre diferentes CoVs, com 54% em todo genoma, sugerindo que as nsps são mais conservadas e as proteínas estruturais

são mais diversificadas em relação à necessidade de adaptação a novos hospedeiros (CHEN; LIU; GUO, 2020).

1.2. CICLO DE VIDA

Por se tratarem de parasitas intracelulares obrigatórios, todos os vírus dependem do mecanismo de tradução da célula hospedeira para a produção de suas proteínas e progênie[8] infecciosa. Como a síntese de proteínas também é essencial para a resposta da célula hospedeira à infecção (resposta imune antiviral inata), não é surpreendente que muitos vírus de RNA, como os coronavírus, modulem a síntese de proteínas do hospedeiro, a fim de limitar a tradução de mRNAs celulares e favorecer a síntese de proteínas virais (DE WILDE et al., 2018; WALSH; MOHR, 2011).

O ciclo de vida dos CoVs envolve quatro etapas essenciais: (i) ligação e entrada, (ii) expressão da replicase, (iii) replicação e transcrição e (iv) montagem e liberação (Figura 3). A ligação eficiente do vírion à célula hospedeira é iniciada por interações entre a proteína S e um receptor de proteína na superfície celular (geralmente exo ou aminopeptidase[9], no caso do SARS-CoV-2 é a Enzima Conversora de Angiotensina 2 Humana [hACE2] [Tabela 1]), cujos RBDs presentes na região S1 variam dependendo do vírus. Essa interação, portanto, é o principal determinante para a infecção e também governa o tropismo tecidual[10] do vírus

[8] Descendência.

[9] Peptidases são um tipo de protease (ver [12]) que quebram ligações peptídicas nos aminoácidos terminais. Exopeptidase catalisa quebra de ligação peptídica terminal. Amonopeptidase catalisa quebra de ligação na extremidade N-terminal.

[10] Fenômeno pelo qual certos tecidos do hospedeiro têm preferência ao crescimento e a proliferação de patógenos.

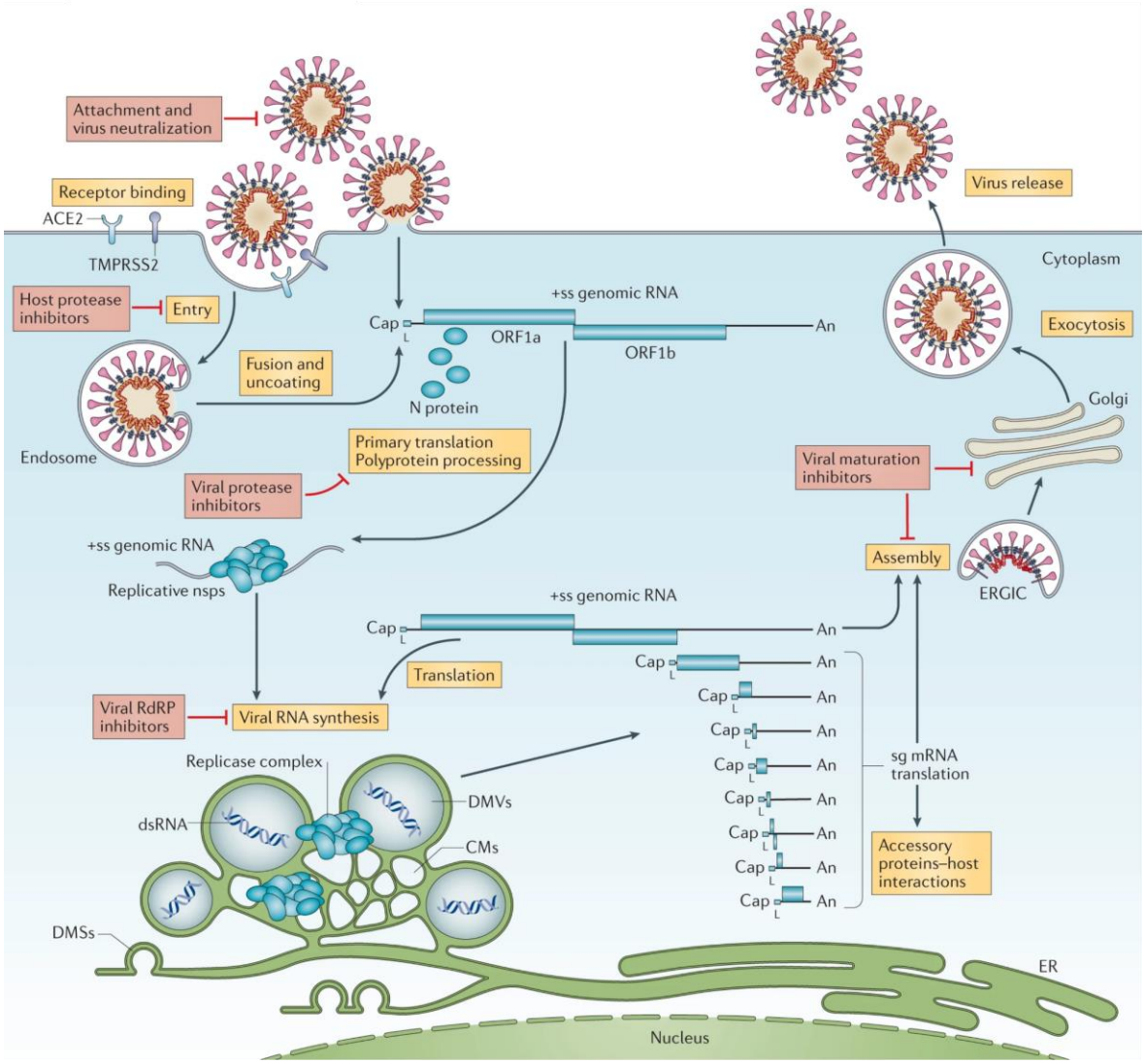

Figura 3. Ciclo de vida de CoVs, evidenciando desde a ligação ao receptor celular até a liberação dos vírus replicados.

Fonte: V'KOVSKI et al. (2021)

(DE WILDE et al., 2018; FEHR; PERLMAN, 2015; PEACOCK et al., 2021a; V'KOVSKI et al., 2021). Após a ligação ao receptor, o vírus deve obter acesso ao citoplasma da célula hospedeira, o que geralmente é realizado por clivagem proteolítica[11] da proteína S por uma serinoprotease (em geral TMPRRS2), seguida pela fusão das membranas virais e celulares, com a consequente liberação do

[11] Processo de quebra das ligações peptídicas entre aminoácidos em proteínas.

genoma viral no citoplasma (FEHR; PERLMAN, 2015; GRAHAM; DONALDSON; BARIC, 2013; V'KOVSKI et al., 2021).

O próximo passo no ciclo de vida do CoV é a tradução do gene da replicase contido em seu genoma. Este gene codifica dois ORFs grandes, rep1a e rep1b, que expressam duas poliproteínas, pp1a e pp1ab. Para expressar ambas as poliproteínas, o vírus utiliza sequências específicas (*slippery* e *pseudoknot*) que causam um sinal de mudança de quadro do ORF rep1a para ORF rep1b. As poliproteínas pp1a e pp1ab são clivadas principalmente pelas proteases[12] do tipo papaína (PLpro) e pela protease principal (3CLpro) para formar os nsps 1-16, os quais se reúnem no complexo replicase-transcriptase (RTC) para criar um ambiente adequado para a síntese de RNA (DE WIT et al., 2016; FEHR; PERLMAN, 2015; PEACOCK et al., 2021a; PERLMAN; NETLAND, 2009; V'KOVSKI et al., 2021).

O RTC direciona a produção de RNAs de sentido negativo por meio de replicação e transcrição. Durante a replicação, cópias de RNA (-) completas do genoma são produzidas e utilizadas como *template* para a geração de genomas de RNAs (+). Durante a transcrição, um conjunto de 7 a 9 RNAs subgenômicos (sgRNAs), incluindo aqueles que codificam todas as proteínas estruturais (S, E, M e N) e proteínas acessórias, são produzidos por transcrição descontínua[13]. Nesse processo, o RTC se liga ao final da extremidade 3′ (+) e prossegue ao longo do genoma na direção 3′ a 5′ sintetizando uma fita negativa. Quando o RTC atinge um TRS, a fita negativa recém-sintetizada pode ser translocada para a sequência líder 5' do genoma, onde é copiada. Isto forma um sgRNA (-) que é copiado para o

[12] Enzimas que quebram ligações peptídicas entre os aminoácidos das proteínas utilizando moléculas de água (hidrólise).

[13] Processo único aos vírus de RNA que ocorre durante a síntese de RNAs de fita negativa.

sgRNA (+) (CHEN; LIU; GUO, 2020; DE WIT et al., 2016; HUSSAIN et al., 2005; PEACOCK et al., 2021a; V'KOVSKI et al., 2021). Esta natureza descontínua tem como consequência um alto grau de recombinação[14] resultante da inserção de sequências virais e não virais ou frequentes deleções no genoma, o que pode levar tanto à formação de genomas viáveis como de RNAs interferentes defeituosos (PEACOCK et al., 2021a).

Uma vez que o processo de expressão e replicação viral esteja em andamento, os vírus da progênie podem começar a se reunir. Primeiramente, as proteínas estruturais são traduzidas e inseridas no retículo endoplasmático (RE). Essas proteínas se movem ao longo da via secretora para o compartimento intermediário ER-Golgi (ERGIC). Os genomas virais encapsulados pela proteína N brotam nas membranas do ERGIC contendo proteínas estruturais virais e formando vírions maduros. A proteína M direciona então a maioria das interações proteína-proteína necessárias para a montagem. Contudo, apesar de seu papel dominante, não é auto-suficiente para a formação dos vírions, necessitando sua expressão junto à proteína E para a produção dos envelopes virais. Adicionalmente, a capacidade da proteína S de trafegar para o ERGIC e interagir com a proteína M é crítica para sua incorporação nos vírions. A proteína M também se liga ao nucleocapsídeo e essa interação promove a conclusão da montagem do vírion (FEHR; PERLMAN, 2015; MASTERS, 2006).

Após a montagem, os vírions são transportados para a superfície celular em vesículas e liberados por exocitose. Ainda é investigado se neste processo é

[14] Processo que ocorre quando pelo menos duas linhagens virais distintas co-infectam a mesma célula hospedeira e trocam segmentos genéticos.

utilizada a via tradicional para o transporte de grandes cargas do Complexo de Golgi ou uma via separada única para sua própria saída. Em vários coronavírus, proteínas S que não são montadas em vírions transitam para a superfície celular, onde atua na fusão célula-célula entre células infectadas e células adjacentes não infectadas. Isso leva à formação de células gigantes multinucleadas, que permitem que o vírus se espalhe dentro de um organismo infectado sem ser detectado ou neutralizado por anticorpos neutralizantes[15] específicos ao vírus (FEHR; PERLMAN, 2015). Com relação ao SARS-CoV-2, evidências recentes mostram que as células infectadas egressam pela via de tráfico lisossômico em detrimento das vias secretoras biossintéticas, as quais são mais comumente utilizadas por outros vírus envelopados (GHOSH et al., 2020).

1.3. SÍNDROMES RESPIRATÓRIAS CAUSADAS POR CORONAVÍRUS

Os coronavírus são patógenos importantes para humanos e vertebrados. Podem infectar o sistema respiratório, gastrointestinal, hepático e nervoso central de humanos, pássaros, morcegos, camundongos e outros animais selvagens (CHEN; LIU; GUO, 2020; VIJAYKRISHNA et al., 2007). Esses vírus geralmente infectam seus hospedeiros de maneira espécie-específica e as infecções podem ser agudas ou persistentes, sendo transmitidas principalmente pelas vias respiratória e fecal-oral (MASTERS, 2006).

Os coronavírus humanos (HCoVs), compostos pelos gêneros *Alphacoronavirus* e *Betacoronavirus*, foram por muito tempo considerados patógenos pouco relevantes por serem associados ao resfriado comum em pessoas saudáveis. No entanto, no século XXI, 2 HCoVs altamente patogênicos —

[15] Anticorpos capazes de impedir e neutralizar a ligação do vírus ao receptor da célula humana.

coronavírus da síndrome respiratória aguda grave (SARS-CoV) e coronavírus da síndrome respiratória do Oriente Médio (MERS-CoV) — emergiram de reservatórios animais para causar epidemias globais de doenças respiratórias com taxas de morbidade[16] alarmantes (PAULES; MARSTON; FAUCI, 2020). Em dezembro de 2019, outro HCoV patogênico, denominado posteriormente SARS-CoV-2, foi identificado em Wuhan (China) como agente etiológico de uma pandemia[17] global (COVID-19) que, com dois anos de duração, já atingiu aproximadamente 270 milhões de pessoas em cerca de 200 países e levou a óbito cerca de 5,3 milhões de pessoas (JOHNS HOPKINS CORONAVIRUS RESOURCE CENTER, 2021; WORLD HEALTH ORGANIZATION, 2021a), provocando o colapso de diversos sistemas de saúde nacionais e regionais.

Além destes três novos coronavírus altamente patogênicos, pelo menos outros quatro HCoVs (229E, NL63, OC43 e HKU1) são endêmicos em todo o mundo e representam 10% a 30% das infecções do trato respiratório superior em adultos (FEHR; PERLMAN, 2015; PAULES; MARSTON; FAUCI, 2020).

1.3.1. SÍNDROME RESPIRATÓRIA AGUDA GRAVE (SARS)

Em novembro de 2002, um surto de uma pneumonia atípica infecciosa, posteriormente chamado de síndrome respiratória aguda grave (SARS), foi relatado no sudeste da China (Guangdong) e em Hong Kong. Acelerada pelas viagens aéreas e aglomerações em hospitais e regiões urbanas, a doença se espalhou rapidamente para várias partes do mundo. Como os surtos de SARS ocorreram no

[16] Taxa de portadores de determinada doença em relação à população total estudada, em determinado local e em determinado momento.

[17] Quando há disseminação de uma nova doença em diferentes continentes com transmissão sustentada de pessoa para pessoa.

sudeste da Ásia, América do Norte e Europa, pode-se dizer que deram origem à primeira pandemia do século XXI (ZHONG et al., 2003). Descobriu-se que o vírus, um *Betacoronavirus* da linhagem B, denominado SARS-CoV, foi transmitido para humanos a partir de reservatórios zoonóticos[18], sendo os morcegos os prováveis reservatórios do vírus, enquanto as civetas de palma mascarada (*Paguma larvata*) e cães-guaxinim (*Nyctereutes procyonoides*) serviram como hospedeiros intermediários (GUAN et al., 2003; LI et al., 2005).

Durante o surto, que se estendeu de 2002 a 2003, 8.098 casos foram confirmados, com 774 mortes, representando uma alta taxa de morbidade de 9.5% (WORLD HEALTH ORGANIZATION, 2003). Essa taxa foi muito maior em idosos, alcançando cerca de 50% em indivíduos acima de 60 anos. Estima-se que entre 20 e 30% dos indivíduos com SARS necessitaram de tratamento em Unidades de Terapia Intensiva (UTIs) (FEHR; PERLMAN, 2015). Embora em termos de número de mortes não seja comparável à influenza, HIV ou vírus da hepatite C, a pandemia causou preocupação mundial e afetou seriamente a economia global, cujas perdas foram estimadas em 30 a 100 bilhões de dólares (KEOGH-BROWN; SMITH, 2008).

Os sintomas da infecção por SARS-CoV são correspondentes àqueles das doenças do trato respiratório inferior. Além da febre, mal-estar e linfopenia[19], os indivíduos afetados apresentam contagens plaquetárias ligeiramente diminuídas, perfis prolongados de coagulação e enzimas hepáticas séricas[20] levemente elevadas. Além destes sintomas relacionados a doenças respiratórias graves, o SARS-CoV também pode causar infecção em outros órgãos, como intestino e rins,

[18] Animais não humanos que carregam patógenos cuja doença pode romper a barreira entre espécies e infectar humanos.

[19] Baixo nível de linfócitos (células de defesa) no sangue.

[20] Enzimas encontradas no fígado.

uma vez que alguns indivíduos afetados tiveram diarreia aquosa e o vírus também pode ser detectado nas fezes e na urina (PEIRIS et al., 2003).

1.3.2. SÍNDROME RESPIRATÓRIA DO ORIENTE MÉDIO (MERS)

Quase uma década após o surto controlado de SARS-CoV, emergiu o próximo coronavírus zoonótico: o coronavírus da Síndrome Respiratória do Oriente Médio (MERS-CoV) (DE GROOT et al., 2013). O vírus foi isolado pela primeira vez em junho de 2012 em um homem da Arábia Saudita de 60 anos de idade que morreu de SARS, falência múltipla de órgãos e insuficiência renal (ZAKI et al., 2012). Embora a maioria dos casos de MERS tenha ocorrido na Arábia Saudita e nos Emirados Árabes Unidos, foram relatados casos na Europa, Estados Unidos da América (EUA) e Ásia em viajantes do Oriente Médio ou de seus contatos (ZUMLA; HUI; PERLMAN, 2015).

Estudos sorológicos identificaram anticorpos de MERS-CoV em camelos dromedários e verificou-se que suas linhagens celulares eram permissivas para a replicação deste vírus, fornecendo evidências de que estes animais possivelmente sejam o hospedeiro natural ou intermediário, uma vez que eles abrigavam vírus similares a MERS-CoV há décadas (MEYER et al., 2014).

Estima-se que houve 1.728 casos confirmados durante o surto (2012-2015) de MERS em 27 países, incluindo 624 mortes, o que implica uma altíssima taxa de morbidade de 36%, quatro vezes maior em relação à SARS (WORLD HEALTH ORGANIZATION, 2016). Ainda é incerto quantos casos de MERS-CoV podem ser atribuídos a um hospedeiro intermediário em comparação com a transmissão de humano para humano, embora tenha sido demonstrado que a transmissão de camelos para humanos contribuiu para o surto (FEHR; PERLMAN, 2015). Até o

presente, continuam sendo reportados casos de MERS, tanto que em 2019 foram notificados 212 casos no Oriente Médio, incluindo 57 mortes. Na Arábia Saudita, dos 198 casos, 118 eram primários (51 relataram contato com camelos), 41 foram adquiridos em hospitais ou serviços de saúde e 32 por contatos domiciliares. Sendo assim, desde 2012 até 2019 foram relatados 2.494 casos, incluindo 912 óbitos (EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL, 2019).

As características clínicas do MERS variam de doença assintomática a SARS e falência múltipla de órgãos, resultando, nesses casos, em morte, especialmente em indivíduos com comorbidades subjacentes (WHO MERS-COV RESEARCH GROUP, 2013). Apesar dos sintomas similares ao SARS, os pacientes com MERS têm um tempo menor para a manifestação do início dos sintomas. Adicionalmente, os indivíduos acometidos apresentam maior necessidade de suporte ventilatório e cargas virais no trato respiratório mais altas durante a primeira semana da doença em relação aos pacientes com SARS (MEMISH et al., 2014).

1.3.3. COVID-19

Os coronavírus voltaram a assombrar a humanidade em dezembro de 2019, quando um grupo de pacientes foi internado em hospitais chineses com diagnóstico inicial de pneumonia com etiologia desconhecida. Posteriormente, essas infecções foram associadas à possível transmissão zoonótica em um mercado atacadista de frutos do mar e animais selvagens em Wuhan (província de Hubei, China), cidade que possui cerca de 11 milhões de habitantes (LI et al., 2020; LU; STRATTON; TANG, 2020). O subsequente isolamento do vírus e a caracterização molecular mostraram que o patógeno era um novo HCoV, o 2019-nCoV, posteriormente chamado de SARS-CoV-2 (ZHOU et al., 2020b; ZHU et al., 2020).

Na primeira fase, de dezembro de 2019 a meados de janeiro de 2020, 41 casos foram confirmados. A segunda fase começou em 13 de janeiro, marcada pela rápida disseminação do vírus nos hospitais (infecção hospitalar) e pela transmissão familiar por contato próximo, de modo que, em 23 de janeiro, 29 províncias da China e seis outros países já somavam 846 casos. Apesar do decreto de isolamento social, 5 milhões de pessoas já haviam deixado Wuhan em virtude do término das comemorações do Ano Novo chinês. A terceira fase começou em 26 de janeiro, marcada pelo rápido aumento de casos agrupados resultantes de transmissão comunitária, tanto que em 30 de janeiro, atingiram-se 9.826 casos confirmados, e a Organização Mundial da Saúde (OMS) declarou esta epidemia uma Emergência de Saúde Pública de Âmbito Internacional (SUN et al., 2020; WORLD HEALTH ORGANIZATION, 2020b). A partir disso, a doença atingiu todos os continentes do mundo, sendo considerada uma pandemia pela Organização Mundial da Saúde (OMS) em março de 2020 (WORLD HEALTH ORGANIZATION, 2020a). Essa pandemia perdura até a presente data, provocando centenas de milhões de casos e milhões de mortes principalmente nos EUA, Brasil, Índia e em países europeus (JOHNS HOPKINS CORONAVIRUS RESOURCE CENTER, 2021; WORLD HEALTH ORGANIZATION, 2021a).

Os sintomas da infecção por SARS-CoV-2 aparecem após um período de incubação de aproximadamente 5,2 dias (LI et al., 2020). Tais sintomas são semelhantes aos relacionados a SARS e MERS, como febre, tosse seca, dispneia[21] e opacidade bilateral em vidro fosco[22] nas tomografias computadorizadas de tórax.

[21] Falta de ar ou dificuldade de respirar.

[22] Padrão frequentemente observado em tomografias computadorizadas de alta resolução do tórax, que indica que está havendo um processo inflamatório ou infeccioso nos pulmões.

Além destas, algumas características clínicas únicas que incluem o acometimento das vias aéreas inferiores, evidenciadas por sintomas como rinorreia[23], espirros e dor de garganta foram observadas (HUANG et al., 2020a). Após o surgimento das Variantes de Preocupação (VOCs) como Delta e Ômicron, o tempo de incubação[24] passou a ser menor (entre 3 e 4 dias) e os aspectos clínicos (especialmente sintomatologia) da doença sofreram algumas alterações (GRANT et al., 2021; JANSEN et al., 2021).

1.4. EPIDEMIOLOGIA GENÔMICA

"A vigilância em saúde pública consiste na coleta, análise e interpretação sistemática e contínua de dados relacionados à saúde, necessários para o planejamento, implementação e avaliação das práticas de saúde pública" (WORLD HEALTH ORGANIZATION, 2021b). A vigilância é realizada para (i) promover um melhor gerenciamento das doenças e levar a ações de saúde pública, como a detecção de surtos; (ii) medir a magnitude, o impacto e as tendências da doença; (iii) melhorar o conhecimento de causas, fontes, reservatórios, riscos e morbidade; (iv) orientar programas para medir a eficácia das intervenções; e (v) ajudar os formuladores de políticas a definir prioridades (DENG; DEN BAKKER; HENDRIKSEN, 2016).

Os avanços recentes nas tecnologias de sequenciamento de nova geração (NGS) e nas ferramentas de bioinformática tornaram o sequenciamento uma solução viável e avançada para a vigilância epidemiológica. O termo Epidemiologia Genômica tem sido cada vez mais utilizado para descrever a utilização de NGS

[23] Congestão nasal ou coriza.

[24] Intervalo entre a data de contato com o vírus até o início dos sintomas.

para comparar sequências de ácidos nucleicos de importância epidemiológica, tais como os elementos genômicos que apresentam taxas de mutação variáveis durante a evolução dos microrganismos (como bactérias e vírus). Estes, por sua vez, são alvos para investigações epidemiológicas em diferentes escalas temporais e geográficas (DENG; DEN BAKKER; HENDRIKSEN, 2016).

A epidemiologia genômica possui aplicações diversas, especialmente para o enfrentamento de patógenos bacterianos causadores de surtos (*e. g.*, intoxicações alimentares), fornecendo um alto poder discriminatório na diferenciação de isolados intimamente relacionados (DENG; DEN BAKKER; HENDRIKSEN, 2016), bem como auxiliando na possível identificação da origem, evolução e disseminação de genes de resistência a antimicrobianos (DOWNING, 2015).

O sequenciamento do genoma do agente etiológico da SARS (SARS-CoV) foi fundamental para permitir a inferência das relações evolutivas existentes entre diferentes isolados de pacientes por meio de análises filogenéticas (ROTA et al., 2003). Uma combinação de informações genômicas e epidemiológicas permitiu às autoridades chinesas rastrear as variações genotípicas[25] determinantes para disseminação viral (RUAN et al., 2003). Durante a atual pandemia (COVID-19), esforços internacionais colaborativos sem precedentes têm permitido a ampla disponibilização de genomas sequenciados em diferentes províncias e países, bem como: (i) a investigação da história evolutiva do vírus, ajudando a inferir sua origem natural zoonótica ou artificial a partir da comparação com isolados proximamente relacionados de outros animais selvagens (ANDERSEN et al., 2020; WORLD

[25] Variações na composição genética de um indivíduo.

HEALTH ORGANIZATION, 2021c); (ii) o estudo estrutural e funcional do receptor hACE2 responsável pela facilitação da entrada do vírus à célula hospedeira em humanos em relação a outros animais, a fim de verificar se o vírus é capaz de transpor a barreira e infectar diferentes espécies mais facilmente (KIM et al., 2020; MUNNINK et al., 2021; ORESHKOVA et al., 2020; SHI et al., 2020); (iii) a compreensão do espalhamento viral em nível local, regional, nacional e internacional (MULLEN et al., 2021a; RAMBAUT et al., 2020a); (iv) entre outros estudos para compreender mudanças na transmissibilidade, na virulência[26] e na resposta às vacinas pelas novas variantes de interesse (VOIs) e preocupação (VOCs) (ALTMANN; BOYTON; BEALE, 2021; DAVIES et al., 2021; GREANEY et al., 2021; MCCORMICK; JACOBS; MELLORS, 2021; WEISBLUM et al., 2020).

1.5. EVOLUÇÃO MOLECULAR

A evolução é uma teoria que postula que os organismos mudam ao longo do tempo, de modo que os descendentes diferem estrutural e funcionalmente de seus ancestrais, uma vez que herdam características morfológicas[27] e fisiológicas[28] destes (DARWIN, 1859). Existem três mecanismos principais pelos quais a evolução pode ocorrer: (i) a existência de condições de crescimento que afetam o desenvolvimento (*i.e.*, fatores ambientais e doenças infecciosas); (ii) o mecanismo de reprodução sexual que assegura a mudança de uma geração para outra, formando uma combinação única a partir dos cromossomos de dois genitores; e (iii)

[26] A relativa infecciosidade ou quantificação da patogenicidade de um microrganismo causador de doenças.

[27] Características que definem a forma do organismo.

[28] Referente às funções orgânicas de um organismo ou aos processos quem o mantêm vivo.

a mutação seguida de seleção e a deriva genética[29] que podem produzir mudanças nos genes e nos cromossomos (SIMPSON; SIMPSON, 1949).

A teoria da evolução de Darwin sugere que, em nível fenotípico[30], são selecionados traços em uma população que aumentam a sobrevivência (seleção positiva), enquanto traços que reduzem a aptidão (*fitness*) não são selecionados (seleção negativa ou purificadora) (DARWIN, 1859).

Um ponto de vista evolutivo convencional considera que as seleções positiva e negativa também operam sob as sequências de DNA. Portanto, para avaliar se a seleção ocorreu nas sequências analisadas, pressupõe-se que a porção de DNA que codifica uma proteína pode ter substituições tanto sinônimas quanto não-sinônimas. Para uma mudança de nucleotídeo em um determinado códon[31], uma substituição sinônima não altera o aminoácido que é codificado, enquanto uma substituição não-sinônima o modifica. Consequentemente, a razão das taxas de substituição não-sinônima (dN) *versus* substituição sinônima (dS) pode revelar evidências de seleção positiva ou negativa. Se dS for maior que dN, considera-se que a sequência está sob seleção negativa, do contrário está sob seleção positiva (NEI; GOJOBORI, 1986).

Por outro lado, Kimura propôs um modelo diferente para explicar a evolução em nível de DNA. Nesta teoria, denominada teoria neutralista, a maioria das substituições de DNA observadas deve ser neutra (ou quase neutra) e a principal causa da variabilidade em nível molecular é a deriva genética. Portanto, sob esse

---

[29] Mudança na frequência dos alelos que ocorre, diferentemente da seleção natural, de maneira aleatória.

[30] Manifestação visível ou detectável de um genótipo, ou seja, conjunto de características observáveis de um organismo, incluindo características morfológicas e fisiológicas.

[31] Sequência de três bases nitrogenadas consecutivas do RNA mensageiro que especifica o aminoácido a ser codificado durante a síntese de proteínas de uma célula.

modelo se considera que a maioria das mutações não-sinônimas são deletérias, não sendo observadas como substituições na população, bem como a seleção positiva tem um papel limitado (KIMURA, 1983).

Os métodos de máxima verossimilhança têm sido comumente usados para estimar dN e dS entre sequências. Nesse caso, um modelo de substituição de códon (semelhante ao modelo de substituição de nucleotídeos) que inclui um parâmetro (ω), especificando a proporção de substituições não-sinônimas em relação a sinônimas, permitindo estimar os parâmetros do modelo por meio da maximização da função de verossimilhança (GOLDMAN; YANG, 1994).

Contudo, tal modelo simples que assume uma pressão seletiva uniforme (dN/dS) sobre todas as linhagens da filogenia é biologicamente irrealista, uma vez que algumas pressões seletivas são exercidas apenas episodicamente sobre a população (*e. g.*, de vírus) durante um determinado período de sua história evolutiva. Para lidar com a complexidade crescente dos processos evolutivos, modelos de códon ligeiramente mais complexos foram projetados para acomodar as pressões seletivas variáveis que são evidentes em diferentes linhagens (LAM; HON; TANG, 2010). Estes foram implementados, por exemplo, nas ferramentas HyPhy (POND; FROST; MUSE, 2005) (Figura 4) e PAML (YANG, 2007).

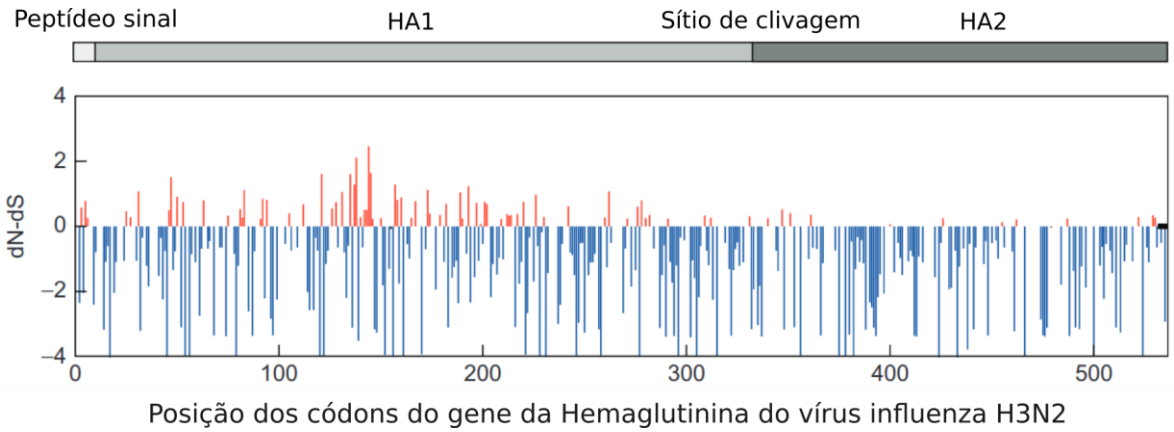

Figura 4. O dN/dS para cada códon do gene Hemaglutinina (HA) do vírus da influenza humana A (H3N2) (n = 100) estimado usando HYPHY. Embora a razão geral dN/dS do gene seja 0,20 (IC 95% = 0,18-0,23) indicando seleção purificadora, alguns códons dentro do gene HA, como mostrado pelas barras superiores (vermelhas), têm uma razão dN/dS > 1 (eixo y no gráfico). O painel superior mostra a estrutura do gene HA do vírus da influenza humana A (H3N2), de modo que apenas regiões parciais do peptídeo sinal e HA2 foram incluídas devido à remoção de algumas colunas com *gaps* no alinhamento.
Fonte: Adaptado de LAM, HON & TANG (2010).

1.6. GENÔMICA COMPARATIVA E FILOGENÉTICA

Nos últimos anos, pesquisadores têm usado vários métodos para alinhar e comparar os genomas sequenciados visando responder questões importantes sobre a função e evolução dos organismos, área que ficou conhecida como genômica comparativa (MILLER et al., 2004). Quando consideramos uma proteína (ou gene), uma das questões mais fundamentais é quais são as outras proteínas relacionadas. Essas sequências biológicas muitas vezes formam famílias de genes derivados de ancestral comum (homólogos), podendo ser genes relacionados dentro de um organismo resultantes de duplicação gênica (parálogos) ou genes em espécies diferentes gerados por eventos de especiação (ortólogos) (PEVSNER, 2015).

Acredita-se que genes homólogos são relativamente conservados ao longo da evolução, enquanto regiões não codificadoras tendem a mostrar graus variáveis de conservação. Desta forma, a análise comparativa de vários genomas filogeneticamente diversos pode fornecer pistas sobre as pressões seletivas que governam o agrupamento de genes e oferecer *insights* sobre seus mecanismos de evolução. Por outro lado, comparações genômicas de espécies intimamente relacionadas também podem ajudar a determinar a base genética para a variação

fenotípica observada, revelando regiões específicas (assinaturas) mais suscetíveis à variação molecular (CHAIN et al., 2003).

Com o objetivo de estudar essas mudanças nos genes e proteínas ao longo dos diferentes ramos da árvore da vida e reconstruir a história evolutiva das espécies, são realizadas análises filogenéticas. Historicamente, foram baseadas em características facilmente observáveis, como a presença ou ausência de asas ou de medula espinhal. Recentemente, essas análises também se baseiam em dados de sequências moleculares de DNA, RNA e proteínas. O fluxo de trabalho envolvido nas análises filogenéticas moleculares pode ser dividido em cinco etapas: (i) seleção de sequências homólogas para análise, (ii) alinhamento múltiplo das sequências de interesse, (iii) escolha do método e da ferramenta de inferência filogenética, (iv) especificação de um modelo evolutivo, e (v) avaliação das filogenias geradas (PEVSNER, 2015).

Após a escolha das sequências homólogas de interesse, pode ser realizado o alinhamento múltiplo de sequências, que consiste em parear uma coleção de três ou mais sequências parcial ou completamente. Para a realização do alinhamento, geralmente são consideradas cinco abordagens algorítmicas: (i) métodos exatos (NEEDLEMAN; WUNSCH, 1970), (ii) alinhamento progressivo, implementado no programa ClustalW (FENG; DOOLITTLE, 1987), (iii) abordagens iterativas, implantadas em MAFFT, PRALINE, IterAlign e MUSCLE (EDGAR, 2004), (iv) métodos baseados em consistência, implementados em MAFFT, ProbCons e T-Coffee (NOTREDAME; HIGGINS; HERINGA, 2000), e (v) métodos baseados em estrutura que incluem informações sobre uma ou mais estruturas proteicas tridimensionais conhecidas para facilitar a criação de um alinhamento de

sequências múltiplas (ARMOUGOM et al., 2006). Atualmente, as abordagens (iii) e (iv) são as mais utilizadas por sua maior velocidade e robustez metodológica. Em geral, MUSCLE e MAFFT são consideradas as ferramentas mais rápidas e precisas e, portanto, mais indicadas para o alinhamento de grandes números de sequências proximamente relacionadas (PEVSNER, 2015).

Há muitas maneiras de se construir uma árvore filogenética, mas consideraremos quatro métodos principais: (i) métodos baseados em distância: começam analisando alinhamentos em pares das sequências e simplesmente usam as distâncias genéticas para inferir a relação entre todos os *taxa* (FELSENSTEIN, 1984); (ii) máxima parcimônia (MP): método baseado em caracteres no qual colunas de resíduos são analisadas no alinhamento múltiplo para identificar a árvore com o menor comprimento total de ramos possível (CZELUSNIAK et al., 1990); (iii) máxima verossimilhança (ML): abordagem estatística baseadas em modelos, projetada para determinar a topologia da árvore e comprimentos de ramos que têm a maior probabilidade de produzir o conjunto de dados observados, a partir do cálculo da verossimilhança para cada resíduo em um alinhamento com a especificação de um modelo evolutivo (FELSENSTEIN, 1981); (iv) inferência Bayesiana: abordagem estatística para modelagem da incerteza em modelos complexos, a qual calcula a probabilidade dos dados fornecidos pelo modelo (P(dados|modelo)), ou seja, busca a probabilidade condicional de uma árvore em relação aos dados fornecidos (HUELSENBECK et al., 2001). Cada um destes métodos possui inúmeras ferramentas que os implementam (Tabela 2).

As análises filogenéticas também dependem de modelos de substituição de nucleotídeos ou aminoácidos, os quais podem ser implícitos ou explícitos. Para

Tabela 2. Principais funcionalidades de algumas das ferramentas de inferência filogenética mais utilizadas

Ferramenta	Classificação e funcionalidade	*Link* do *software*
BEAST	Realiza inferência Bayesiana baseada em MCMC utilizando modelos de relógio molecular e permitindo a inferência de processos filogeográficos e filodinâmicos	http://beast.community/
GARLI	Utiliza algoritmos genéticos para buscar por árvores com ML	https://code.google.com/archive/p/garli/
HYPHY	Gera modelos de evolução molecular utilizando ML, em particular para inferências de seleção natural	http://www.hyphy.org/
IQ-TREE	Possui um algoritmo rápido e estocástico para inferir filogenias com ML, incluindo um método de seleção de modelos evolutivos que testa rapidamente > 200 modelos	http://www.iqtree.org/
MEGA	Inclui métodos de distância, MP e ML em interface gráfica	https://www.megasoftware.net/
MrBayes	Realiza inferência Bayesiana baseada em MCMC para estimar a distribuição *a posteriori* dos parâmetros do modelo	http://nbisweden.github.io/MrBayes/
PAUP	Inclui métodos de distância, MP e ML	https://paup.phylosolutions.com/
PHYLIP	Inclui métodos de distância, MP e ML	http://evolution.gs.washington.edu/phylip.html
PhyML	Programa rápido que realiza inferência por ML	http://www.atgc-montpellier.fr/phyml/
RAxML	Programa rápido que realiza inferência por ML utilizando modelo GTR	https://cme.h-its.org/exelixis/web/software/raxml/

Abreviações. GTR: *General time reversible*; MCMC: *Markov chain Monte Carlo*.
Fonte: Adaptado de YANG & RANNALA (2012).

métodos baseados em distância e para a máxima parcimônia, são utilizados modelos de substituição implícitos, baseados no número de alterações entre os pares de sequências e o menor comprimento de ramos possível, respectivamente. Para a máxima verossimilhança e abordagens Bayesianas, modelos estatísticos

explícitos são aplicados a caracteres individuais (resíduos) a fim de estimar a topologia[32] mais provável, bem como outros parâmetros, tais como taxas de substituição ao longo de ramos individuais (PEVSNER, 2015).

Os modelos evolutivos, portanto, descrevem a probabilidade de que cada nucleotídeo (ou aminoácido) mude para outro. Os mais conhecidos são: (i) Jukes-Cantor (JC69): possui um parâmetro, assumindo que todas as mutações têm as mesmas probabilidades (JUKES; CANTOR, 1969) (Figura 5a); (ii) Kimura-2P (K80): possui dois parâmetros, permitindo diferentes taxas para transições[33] e transversões[34] e contabilizando com maior fidedignidade a probabilidade de transversões causarem mutações não-sinônimas nas regiões codificadoras de proteínas (KIMURA, 1980) (Figura 5b); (iii) Tamura: estende o modelo de dois parâmetros de Kimura para ajustar o conteúdo de guanina e citosina (GC) das sequências de DNA (TAMURA, 1992) (Figura 5c). Contudo, tais modelos representam simplificações importantes do processo evolutivo, de modo que modelos como *General Time Reversible* (GTR) (TAVARE, 1986) com oito parâmetros livres e Hasegawa-Kishino-Yano (HKY) (HASEGAWA; KISHINO; YANO, 1985) com quatro parâmetros podem ser mais apropriados e flexíveis, uma vez que permitem desigualdades nas frequências de bases.

Para a avaliação das filogenias geradas, os principais critérios utilizados são: consistência, eficiência e robustez (HILLIS, 1995). A abordagem mais comum é a

[32] A forma como os nós (pontos de união entre os ramos ou pontos de divergência) se relacionam entre si em uma árvore filogenética.

[33] Tipo de substituição de base na qual uma base nitrogenada é alterada para a outra base da mesma classe. Ou seja, as purinas podem trocar umas com as outras (A → G e vice-versa). Por outro lado, as pirimidinas podem trocar entre si (C → T e vice-versa).

[34] Outro tipo de substituição em que há alteração para uma base de outra classe. Ou seja, as purinas convertem-se em pirimidinas e as pirimidinas convertem-se em purinas.

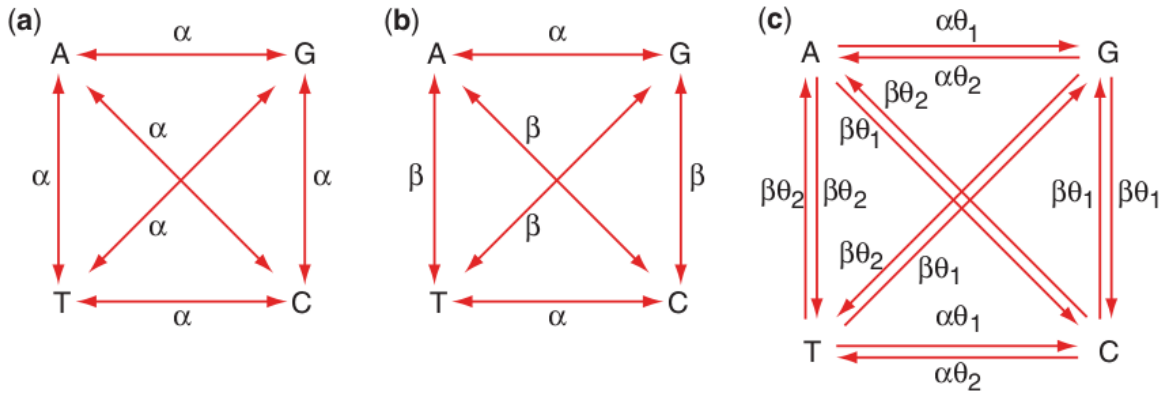

Figura 5. Exemplos de modelos de substituição de nucleotídeos. (a) O modelo JC69 assume que cada resíduo de nucleotídeo tem a mesma probabilidade de mudar para qualquer dos outros três resíduos e que as quatro bases estão presentes em proporções iguais. (b) O modelo K80 permite diferentes taxas para transições e transversões. (c) No modelo de Tamura, um modelo mais complexo, existem parâmetros distintos para diferentes substituições de nucleotídeos e tais parâmetros são direcionais (e.g., a taxa de mudança de T → C difere da taxa de C → T).

Fonte: PEVSNER (2009).

análise de *bootstrapping*, a qual avalia a robustez da topologia das árvores. Nessa técnica, o programa: (i) cria um conjunto de dados artificial do mesmo tamanho do conjunto de dados original (por reamostragem), escolhendo aleatoriamente colunas do alinhamento múltiplo; (ii) estima a árvore filogenética usando o mesmo método de inferência para cada uma das replicatas (*bootstrap*); (iii) as árvores *bootstrap* são então comparadas às árvores originais inferidas, de modo que se obtém a frequência com que cada clado[35] é observado na árvore original. Sendo assim, esta métrica permite inferir margens de erros com base em uma distribuição desconhecida da qual foram extraídos os dados (FELSENSTEIN, 1985).

A genômica comparativa e a filogenética têm sido utilizadas para a análise genética de vírus endêmicos, permitindo estudos que compreendessem a epidemiologia e dinâmica de transmissão do vírus da imunodeficiência humana

[35] Grupo de organismos originados a partir de um único ancestral comum exclusivo.

(HIV) (HEMELAAR et al., 2011; LEWIS et al., 2008), as origens e posterior evolução de síndromes respiratórias agudas graves (SARS e MERS) (ANDERSON et al., 2004; COTTEN et al., 2013; DUDAS et al., 2018; RUAN et al., 2003), bem como as origens, a evolução e a sazonalidade dos vírus da gripe (KOELLE et al., 2006; SMITH et al., 2009).

1.7. FILODINÂMICA

Os genomas virais podem ser estudados de várias maneiras, incluindo a análise das taxas evolutivas de vírus de RNA em rápida evolução, a identificação de mutações resistentes e vacinas e terapias, bem como a reconstrução da filogenia de linhagens de vírus estreitamente relacionadas para identificar e ligar epidemiologicamente os pacientes infectados em um surto ou evento de transmissão (LAM; HON; TANG, 2010).

Para a inferência das relações evolutivas, temporais e geográficas existentes entre os organismos, as análises filogenéticas devem considerar algumas teorias e pressupostos evolutivos importantes. O primeiro deles é o conceito de relógio molecular, o qual assume que para cada gene (ou proteína), a taxa de evolução molecular é aproximadamente constante (MARGOLIASH, 1963; ZUCKERKANDL; PAULING, 1962). Uma vez que as sequências de ácidos nucleicos e proteínas evoluem a taxas constantes, então elas podem ser utilizadas para estimar o tempo pelo qual tais sequências divergiram (PEVSNER, 2015). Evidentemente, algumas exceções a esta teoria existem, tal como a existência de variação na taxa de evolução de alguns organismos, como os vírus. Para estes casos particulares, existem métodos estatísticos que realizam a datação da árvore relaxando a

hipótese do relógio molecular e permitindo que cada ramo da filogenia tenha sua própria taxa de substituição (DRUMMOND et al., 2006).

A disponibilidade de dados genômicos em nível de espécie oferece oportunidades sem precedentes para responder questões evolutivas emergentes. Quando integradas com dados genômicos, informações demográficas e geográficas fornecem pistas sobre a evolução dos organismos, incluindo sua dispersão regional e o entendimento de eventos evolutivos como especiação, radiação adaptativa e extinção (BERMINGHAM; MORITZ, 1998; KNOWLES, 2009). Para tratar desta integração entre dados filogenéticos e geográficos, surgiu o conceito de filodinâmica, que consiste na análise estatística de dados populacionais de espécies intimamente relacionadas para inferir parâmetros e processos populacionais tais como tamanho populacional efetivo[36], padrões e taxas de crescimento, migração e transmissão (YANG; RANNALA, 2012). Portanto, é considerada uma ponte que liga o estudo dos processos micro e macroevolucionários (BERMINGHAM; MORITZ, 1998).

Embora estudos epidemiológicos prevejam que alguns organismos (*e.g.*, vírus de RNA) têm ancestrais que datam de milênios, as populações virais atualmente em circulação possuem um ancestral muito mais recente, indicando a contínua rotatividade de linhagens (HOLMES, 2008). Nesse sentido, o estudo da filodinâmica e evolução viral permite esclarecer tópicos importantes da epidemiologia de doenças infecciosas, tanto (i) na predição de quais novas infecções irão ocorrer; (ii) quais serão seus reservatórios; (iii) e em que locais

[36] Número de indivíduos em uma população que contribuem com descendentes para a próxima geração.

surgirão e se propagarão nas populações humanas (KILPATRICK et al., 2006); quanto no manejo de doenças em andamento, como a COVID-19.

Uma vez que os vírus se disseminam e infectam humanos, a demografia das populações humanas tem um papel importante na variação genética das populações virais. Sendo assim, padrões de variação genética viral são afetados tanto pela seleção natural atuando sobre os genomas virais, quanto pela rapidez com que a transmissão ocorre e pelas medidas adotadas para contê-la (VOLZ; KOELLE; BEDFORD, 2013). Portanto, tais padrões espaciais de dispersão podem ser recuperados por meio de análises filodinâmicas. O princípio básico da abordagem filodinâmica é que processos epidemiológicos como taxas de crescimento/declínio populacional e seleção natural estão implicitamente escritos em sequências genômicas e podem ser recuperados usando um conjunto de técnicas filogenéticas (HOLMES, 2008). Desde o surgimento do termo em 2004 (GRENFELL et al., 2004), a pesquisa relacionada à filodinâmica viral tem se concentrado na dinâmica de transmissão de epidemias e pandemias, visando esclarecer como essas dinâmicas impactam a variação genética viral e vice-versa (VOLZ; KOELLE; BEDFORD, 2013).

A unidade matemática mais fundamental da epidemiologia de doenças infecciosas é o número básico de reprodução (R_0), definido como o número de casos secundários produzidos quando um patógeno é introduzido em uma população suscetível (DIEKMANN; HEESTERBEEK; METZ, 1990). Outro conceito, relacionado à genética de populações[37], chamado de teoria coalescente, é muito

[37] Área da biologia que estuda a composição genética das populações e as mudanças nesta composição que resultam da influência de vários fatores (tamanho da população, mutação, deriva genética, seleção natural, diversidade ambiental, migração, etc.)

importante em estudos filodinâmicos. O objetivo da teoria coalescente é inferir processos evolutivos chave a partir da distribuição de eventos de ramificação em árvores filogenéticas de isolados de uma única espécie (KINGMAN, 1982), permitindo a estimativa de R_0 e a datação dos eventos de ramificação diretamente a partir de dados de sequências genéticas e fornecendo uma ligação natural entre a análise evolutiva das sequências genômicas e a epidemiologia das doenças infecciosas (GRENFELL et al., 2004; PYBUS et al., 2001). A rápida taxa de evolução dos vírus permite que modelos de relógio molecular sejam estimados a partir de sequências genéticas, fornecendo assim uma taxa anual de evolução do vírus, a qual permite a inferência da data do ancestral comum mais recente[38] (MRCA) para um conjunto de sequências (VOLZ; KOELLE; BEDFORD, 2013)

A maioria das análises filodinâmicas começa com a reconstrução de uma árvore filogenética. Como múltiplas sequências genéticas são frequentemente amostradas em múltiplos pontos no tempo, podem-se estimar as taxas de substituição e o tempo do MRCA utilizando um modelo de relógio molecular e algoritmos Bayesianos baseados em *Markov Chain Monte Carlo* (MCMC) que incorporam e quantificam as incertezas do modelo estimado (DRUMMOND et al., 2002).

A filogeografia se refere à estimativa de taxas de movimentação de linhagens virais entre localizações geográficas e a reconstrução das localidades das linhagens ancestrais. Convenientemente, estas variáveis podem ser estimadas utilizando modelos de cadeias de Markov semelhantes àqueles que modelam substituições de bases. Isso é possível porque a localização geográfica é tratada

[38] Indivíduo mais recente do qual todos os organismos de um grupo descendem diretamente.

como um estado de caracter (discreto ou contínuo), permitindo que a mesma possa ser usada para inferir matrizes de transição geográfica (LEMEY et al., 2009). O resultado final é usualmente uma taxa, medida em termos de anos ou de substituições de nucleotídeos por sítio, que uma linhagem de uma região se desloca para outra região ao longo do curso da árvore filogenética (Figura 6). Em uma rede de transmissão geográfica, algumas regiões/países podem apresentar *clusters* de transmissão, enquanto outras regiões podem estar mais isoladas (VOLZ; KOELLE; BEDFORD, 2013).

Por exemplo, em um estudo que caracterizou os quatro principais sorotipos circulantes do vírus da dengue de diferentes regiões das Américas, foi possível investigar fatores que influenciam a taxa e intensidade da transmissão deste vírus, elucidando sua dinâmica espaço-temporal. Foram realizadas análises Bayesianas incorporando um modelo de relógio relaxado com taxas de substituição ramo-específicas e um conjunto de sequências temporal e geograficamente diversas (ALLICOCK et al., 2012). Adicionalmente, baseado em um grande conjunto de dados de 1.610 genomas (mais de 5% dos casos conhecidos), foi realizada a reconstrução espaço-temporal da evolução e propagação do vírus Ebola durante a epidemia na África Ocidental de 2013-2016 (Figura 7) (DUDAS et al., 2017). Ambas análises foram implementadas utilizando o *software* BEAST (DRUMMOND; RAMBAUT, 2007; SUCHARD et al., 2018), uma ferramenta que incorpora a inferência da dinâmica espacial e temporal das linhagens virais para visualização e investigação de hipóteses filogeográficas utilizando inferência Bayesiana (LEMEY et al., 2009).

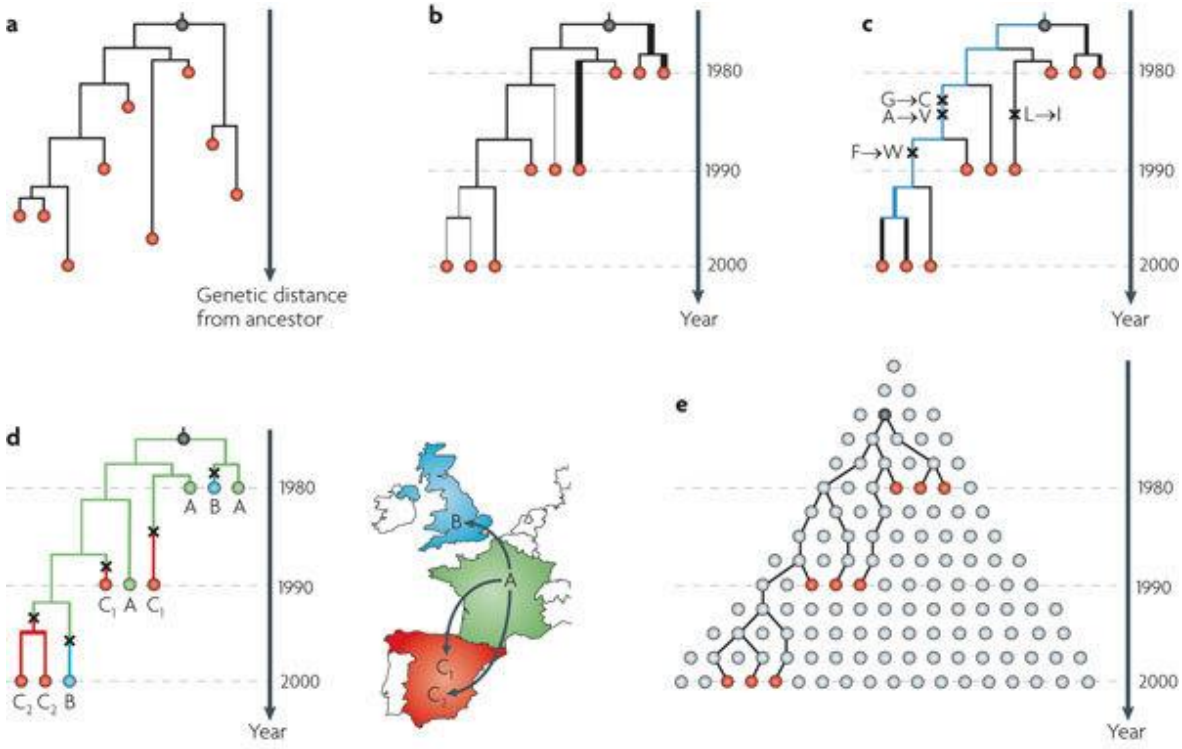

Figura 6. Princípios dos métodos de relógio molecular e filogeografia. (a) Filogenias moleculares enraizadas podem ser estimadas a partir de sequências de genes ou genomas virais. Esta filogenia não tem escala de tempo, portanto o comprimento do ramo representa a divergência genética do ancestral (círculo preto) em substituições por sítio . (b) A mesma filogenia também pode ser reconstruída usando um modelo de relógio molecular, que define uma relação entre distância genética e tempo. Nesse caso, as sequências foram amostradas em pontos de tempo conhecidos e os ramos da filogenia têm comprimentos em unidades de anos, permitindo a estimativa da idade dos eventos de ramificação. (c) Os dados filodinâmicos também podem demonstrar a evolução das mutações ao longo do tempo. (d) Sequências virais também podem ser analisadas utilizando filogeografia temporal. No exemplo, as nove sequências foram amostradas na França (verde, A), no Reino Unido (azul, B) e em dois locais na Espanha (vermelho, C1 e C2). Métodos estatísticos podem ser usados para reconstruir o histórico de propagação de patógenos, de modo que cada ramo seja rotulado com sua posição geográfica estimada. (e) Os princípios de análises coalescentes, que incorporam um modelo explícito da população amostrada. Cada círculo representa uma infecção, e os círculos na mesma linha ocorrem durante o mesmo período de tempo. A largura crescente de cada fileira reflete o crescimento da epidemia ao longo do tempo. A partir das infecções amostradas (vermelho), as linhagens amostradas (linhas pretas) podem ser rastreadas por meio de infecções não amostradas (cinza) até o ancestral comum (círculo preto). A taxa na qual as linhagens amostradas coalescem depende de processos populacionais como dinâmica e estrutura populacional e seleção natural.

Fonte: PYBUS & RAMBAUT (2009).

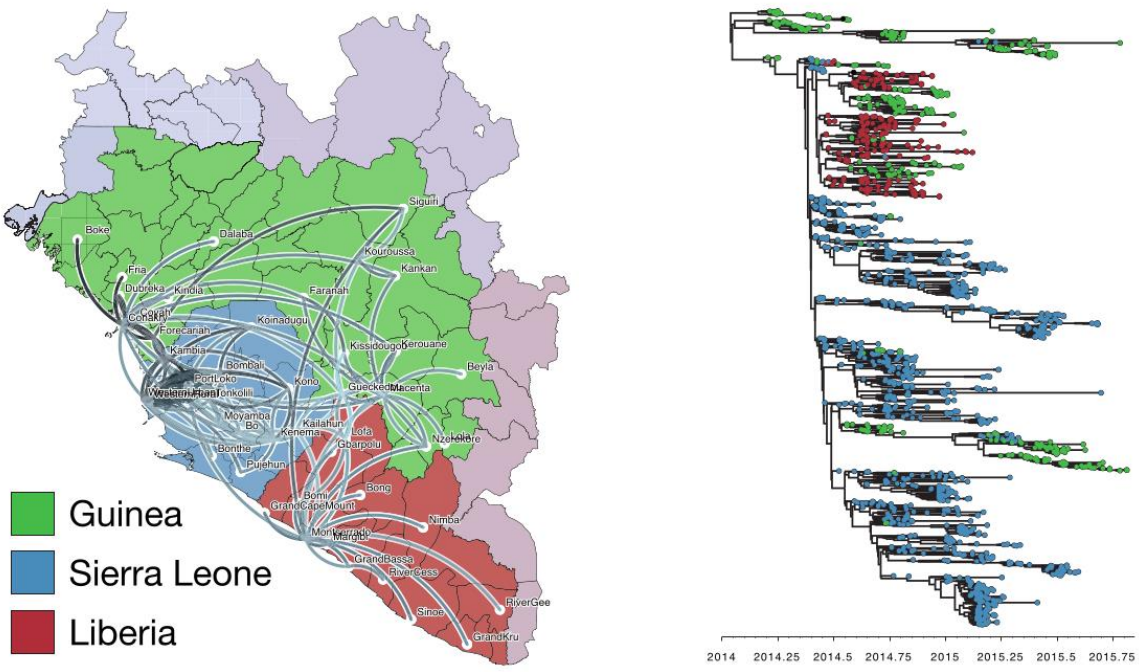

Figura 7. Análise filogeográfica e filodinâmica da epidemia do vírus Ebola na África Ocidental (2013-2016), abrangendo a estimativa simultânea dos dados de sequência e informações geográficas. Os gráficos mostram uma fotografia do espalhamento geográfico e uma árvore de credibilidade máxima de clado (MCC) que sumariza os resultados de maior probabilidade posterior da inferência Bayesiana.
Fonte: Adaptado de SUCHARD et al. (2018).

Portanto, o estudo da filodinâmica de uma doença infecciosa a partir de diferentes tipos de dados de uma série temporal de indivíduos infectados fornece um meio de quantificar e compreender os processos ecológicos, imunológicos e epidemiológicos que impulsionam sua dinâmica e evolução, o que é determinante para seu correto manejo e para a redução do número de mortes que ocorrem durante surtos, epidemias e pandemias (COBEY; KOELLE, 2008).

1.8. EVOLUÇÃO MOLECULAR DO SARS-COV-2

1.8.1. POSSÍVEIS ORIGENS DO SARS-COV-2

Apesar dos esforços colaborativos internacionais para a descoberta das origens do SARS-CoV-2, ainda não existem evidências suficientes para afirmar que

tenha ocorrido um *spillover*[39] direto de morcegos para humanos ou que algum hospedeiro intermediário tenha transmitido o vírus inicialmente para seres humanos cruzando a barreira das espécies, como ocorreu com o SARS-CoV e o MERS-CoV. A hipótese de escape de um laboratório de virologia também permanece sendo investigada, assim como a origem recombinante (LYTRAS; MACLEAN; ROBERTSON, 2020) e a possível transmissão crítica e adaptação em humanos antes da detecção dos primeiros casos na China (ZHANG; HOLMES, 2020).

Uma revisão sistemática publicada em 2001 identificou 1.415 espécies de organismos infecciosos capazes de se tornarem patogênicos em humanos, sendo 217 (15,3%) vírus ou príons. Importantemente, 132 (75%) das 175 espécies patogênicas consideradas emergentes se transmitem de animais para humanos (zoonoses) (TAYLOR; LATHAM; WOOLHOUSE, 2001).

Em um trabalho inicial, uma revisão e análise comparativa de dados genômicos de coronavírus forneceu uma visão geral sobre as características notáveis do genoma do SARS-CoV-2 e discutiu cenários pelos quais o vírus poderia ter emergido. Observou-se que (i) o SARS-CoV-2 parece ser otimizado para se ligar ao receptor hACE2, (ii) que a proteína *Spike* tem um local de clivagem polibásico funcional na junção S1/S2 com a inserção de 12 nucleotídeos, (iii) o RBD da proteína *Spike* é a parte mais variável do seu genoma. Dadas estas três características, os autores propuseram dois cenários: (1) seleção natural em um hospedeiro animal antes da transferência zoonótica; e (2) seleção natural em humanos após a transferência zoonótica (ANDERSEN et al., 2020).

[39] Transmissão de um patógeno de um animal vertebrado infectado para um humano rompendo a barreira das espécies.

Dada a semelhança do SARS-CoV-2 com os coronavírus circulantes em morcegos, é provável que os morcegos serviram como reservatório para seu progenitor. A diversidade de vírus SARS-*like* em morcegos foi evidenciada em um estudo no qual foram coletadas 411 amostras de morcegos na província de Yunnan (China) de maio de 2019 a novembro de 2020. Foram recuperados 24 genomas completos de coronavírus, incluindo quatro novos proximamente relacionados ao SARS-CoV-2 e três relacionados ao SARS-CoV. Destes, o RpYN06 exibiu 94,5% de identidade com o SARS-CoV-2 em todo o genoma e possuía a maior similaridade dentre todos os coronavírus já amostrados nos genes ORF1ab, ORF7a, ORF8, N e ORF10. Uma modelagem ecológica estimou a coexistência de até 23 espécies de morcegos *Rhinolophus* no Sudeste Asiático e sul da China, demonstrando que CoVs proximamente relacionados ao SARS-CoV-2 circulam em espécies selvagens em uma ampla região geográfica (ZHOU et al., 2021). Essa grande distribuição espacial e a necessidade de priorização do Sudeste Asiático para esforços de vigilância de coronavírus no futuro foram reforçados após a identificação de vírus relacionados ao SARS-CoV-2 em dois morcegos *Rhinolophus shameli* no Camboja em estudo retrospectivo com amostras de 2010. Os CoVs RshSTT182 e RshSTT200 apresentaram 92,6% de identidade nucleotídica ao longo do genoma, a maior identidade já registrada na região de nsp4 a nsp8 na ORF1ab, porém grande dissimilaridade[40] na proteína S (DELAUNE et al., 2021).

Em outro estudo que avaliou as possíveis origens evolutivas da linhagem que deu origem a pandemia de COVID-19, observou-se que os vírus pertencentes ao subgênero *Sarbecovirus* passam por frequentes recombinações e exibem ampla

[40] Diferença significativa nos resíduos de aminoácidos de ambos os organismos investigados.

diversidade genética e espacial na China. Utilizando três abordagens distintas, as datas de divergência (MRCA) entre o SARS-CoV-2 e o seu possível reservatório em morcegos (RaTG13) foram estimadas em 1948 (1879-1999), 1969 (1930-2000) e 1982 (1948-2009), indicando que a linhagem que deu origem ao SARS-CoV-2 tem circulado há décadas nos morcegos sem ser notada (BONI et al., 2020). Muito embora o isolado RaTG13 amostrado de *Rhinolophus affinis* tenha ~96% de identidade ao SARS-CoV-2 (ZHOU et al., 2020b), sua *Spike* diverge no RBD, o que sugere que pode não se ligar eficientemente ao hACE2 (WAN et al., 2020). Por estar associado a eventos de recombinação, outros três vírus (RmYN02, RpYN06 e PrC31) são mais similares ao longo do genoma (particularmente ORF1ab) e, portanto, compartilham um ancestral comum mais recente com o SARS-CoV-2 em relação ao RaTG13 (LYTRAS et al., 2021).

Em um esforço para compreender a diversidade de vírus SARS-*like* em Laos (península da Indochina) foram capturados 645 morcegos pertencentes a seis famílias e 46 espécies. O estudo, disponibilizado em setembro de 2021, demonstrou que três genomas (BANAL-52, BANAL-102 e BANAL-236) possuem alta similaridade ao SARS-CoV-2, com um a dois resíduos diferindo no RBD apenas. A sequência BANAL-52 possui a maior similaridade ao longo do genoma identificada até o momento, possuindo alta conservação no domínio S1 (TEMMAM et al., 2021).

A partir de uma análise metagenômica[41] de amostras de 227 morcegos coletadas na província de Yunnan (China), entre maio e outubro de 2019, um novo

[41] Abordagem que engloba o sequenciamento de todo o material genético de uma amostra, permitindo o estudo de genomas de microrganismos sem a necessidade de realizar cultivos celulares individuais.

CoV derivado de morcegos (RmYN02) foi identificado. Apesar de possuir somente 93,3% de identidade nucleotídica com o SARS-CoV-2 no genoma completo e 97,2% de identidade na ORF1ab, o RmYN02 mostrou baixa identidade (61,3%) no RBD, provavelmente não sendo capaz de se ligar à hACE2. Porém, apesar de não ser o progenitor imediato, esse CoV possui o sítio de clivagem de furinas[42] característico do SARS-CoV-2, indicando que tais eventos de inserção podem ocorrer naturalmente em *Betacoronavirus* que infectam animais (ZHOU et al., 2020a). Essa hipótese foi reforçada com a demonstração de que estes sítios de clivagem localizados na junção S1/S2 são comuns e emergiram de modo independente em múltiplos coronavírus (WU; ZHAO, 2021), incluindo *Alphacoronavirus* de felinos, MERS-CoV, HCoV-OC43, HCoV-HKU1 e HKU9-1 (HOLMES et al., 2021).

Em uma análise metatranscriptômica[43] na qual foram amostrados 1.725 animais de caça de 16 espécies e cinco ordens de mamíferos da China, foram identificados 71 vírus patogênicos em mamíferos, sendo 45 descritos pela primeira vez. Dezoito deles foram considerados de alto risco potencial para humanos e animais domésticos, os quais foram principalmente carregados pelas civetas (*Paguma larvata*). Adicionalmente, foram detectadas transmissões do coronavírus HKU8 de um morcego para uma civeta, assim como saltos de espécies de morcegos para ouriços e de aves para porcos-espinhos. Portanto, estes dados evidenciam o papel predominante dos animais selvagens para o surgimento de doenças infecciosas nas populações humanas (HE et al., 2021).

---

[42] Consiste em quatro aminoácidos PRRA codificados na proteína *Spike*, os quais são clivados por furinas, separando as porções S1 e S2 e mediando a fusão das membranas celulares e virais.

[43] Técnica que utiliza sequenciamento completo e estuda a expressão de genes de comunidades microbianas complexas.

Em março de 2021, a OMS publicou um documento após investigações na China de quatro possíveis fontes de emergência: (i) transmissão zoonótica direta para humanos (*spillover*); (ii) introdução por meio de um hospedeiro intermediário seguido de *spillover*; (iii) introdução por meio da cadeia de alimentos frios; e (iv) introdução por meio de um incidente de laboratório. Os pesquisadores consideraram a probabilidade de cada um desses eventos dadas as evidências prévias e levantadas no estudo, sendo (i) de possível a provável; (ii) muito provável; (iii) possível; e (v) extremamente improvável (WORLD HEALTH ORGANIZATION, 2021c). Contudo, alguns especialistas em evolução molecular argumentaram que os dados publicados neste documento não traziam evidências conclusivas e balanceadas para ambas as hipóteses (origem natural ou escape de laboratório), de modo que consideraram as afirmações demasiadamente definitivas (BLOOM et al., 2021).

Após tais argumentações, um dos autores identificou um conjunto de dados contendo 13 sequências do SARS-CoV-2 do início da epidemia de Wuhan que foi excluído do *Sequence Read Archive* (BLOOM, 2021) e verificou que os genomas do Mercado de Frutos do Mar de Huanan — uma das possíveis fontes de origem do vírus e o foco do relatório conjunto da OMS—, não são representativas de todos os SARS-CoV-2 circulantes em Wuhan no início da epidemia. Sendo assim, há duas identidades plausíveis para o progenitor de todos os SARS-CoV-2 conhecidos: uma é a proCoV2 (KUMAR et al., 2021) e outra é um dos novos genomas que carrega três mutações (C8782T, T28144C e C29095T) relativas ao Wuhan-Hu-1 (usada como referência atualmente e isolada no Mercado de Huanan). Mesmo ainda não respondendo às questões sobre origem natural ou escape de

laboratório, esses achados demonstram que tal remoção de dados, apesar de prática comum e permitida, pode ter sido propositalmente realizada para esconder pistas sobre o surgimento da doença.

Mais recentemente, valendo-se de registros das vendas de animais selvagens nos mercados de Wuhan entre maio de 2017 e novembro de 2019, observou-se a venda de 47.381 indivíduos de 38 espécies, incluindo 31 espécies protegidas. Além disso, demonstraram-se as más condições de higiene sob as quais estes animais foram mantidos antes da venda. Contudo, a ausência de representação de pangolins nesse conjunto de dados sugere que os pangolins dificilmente serviram como hospedeiros intermediários do SARS-CoV-2 em sua origem. Porém, a presença de outros animais suscetíveis ao SARS-CoV-2, como cães-guaxinim e *minks* não descarta a participação de algum hospedeiro intermediário e a origem zoonótica do vírus (XIAO et al., 2021). Além disso, uma análise detalhada da diversidade do SARS-CoV-2 durante o início da pandemia aponta para a presença de erros de sequenciamento ou montagem de genomas intermediários entre as linhagens A e B devido à observação repetida de pares de genomas com homoplasias[44] improváveis, o que põe em dúvida a presença de intermediários entre tais linhagens e torna plausível a transmissão do vírus para humanos em múltiplos mercados de Wuhan com cadeias de fornecimento compartilhadas para venda de animais selvagens (PEKAR et al., 2021b).

Em uma revisão crítica sobre as teorias de origem natural (zoonótica) e escape de laboratório escrita por alguns dos principais especialistas em biologia evolutiva do mundo, são levantadas todas as evidências epidemiológicas e

[44] Mutações que ocorrem de modo independente em diferentes ramos da árvore filogenética.

evolutivas disponíveis até o momento (HOLMES et al., 2021). Muito embora a hipótese de manipulação proposital ou escape acidental não possam ser completamente descartadas, a introdução de características complexas como patogenicidade[45] ou transmissibilidade em um vírus geralmente não é viável exceto em sistemas muito limitados e bem caracterizados (GRONVALL, 2021). Além disso, não há associação dos primeiros casos a laboratórios de pesquisa de Wuhan, nem evidências de que estes possuíam ou trabalhavam em um CoV semelhante ao SARS-CoV-2 antes da pandemia que permitisse estudos de ganho de função que resultassem no causador da COVID-19. Portanto, o corpo de evidências sobre coronavírus anteriores — incluindo o surgimento no mesmo período da SARS que indica certo grau de sazonalidade nas infecções causadas por coronavírus (GRONVALL, 2021)—, a associação epidemiológica aos mercados de venda de animais, a ampla subamostragem de CoVs em morcegos e animais selvagens, bem como a evolução natural do SARS-CoV-2 durante a pandemia apontam para uma origem natural (HOLMES et al., 2021; LYTRAS et al., 2021). Além disso, uma investigação minuciosa dos primeiros casos em Wuhan demonstrou que a maioria deles se concentrou em torno do mercado de Huanan — mesmo na ausência de contato direto por trabalho ou visita —, o que sugere que havia transmissão comunitária do SARS-CoV-2 em Wuhan em dezembro de 2019 antes mesmo de sua detecção e notificação às autoridades chinesas (WOROBEY, 2021).

[45] Termo qualitativo relacionado à capacidade potencial de desencadear doenças.

1.8.2. NOMENCLATURA DINÂMICA E ACOMPANHAMENTO DO ESPALHAMENTO VIRAL

Desde o sequenciamento do primeiro genoma do SARS-CoV-2 (ZHOU et al., 2020b), esforços internacionais sem precedentes de sequenciamento viral permitiram a disponibilização de cerca de sete milhões de genomas do SARS-CoV-2 no banco de dados *Global initiative on sharing all influenza data* (GISAID) até o momento (SHU; MCCAULEY, 2017), que podem ser utilizados para estudos de epidemiologia genômica que permitem acompanhar a história e a dinâmica evolutiva do SARS-CoV-2 ao longo do espaço e do tempo (RAMBAUT et al., 2020a).

Para capturar a história evolutiva e o espalhamento geográfico do SARS-CoV-2 no mundo foi criada uma nomenclatura dinâmica para classificação dos isolados. Nesta abordagem, uma linhagem é um conjunto (*cluster*) de sequências que estão associadas a um evento epidemiológico, isto é, a introdução (isto é, importação) do vírus em uma área geográfica distinta com evidência de propagação posterior.

De um modo geral, os rótulos das principais linhagens começam com uma letra. Na raiz[46] da filogenia do SARS-CoV-2 estão duas linhagens que foram denominadas A (sequência de 05 de janeiro de 2020: Wuhan/WH04/2020) e B (representante de 26 de dezembro de 2019: Wuhan-Hu-1). Para as demais linhagens descendentes da linhagem A ou B se atribui um valor numérico (por exemplo, A.1 ou B.2), seguindo as seguintes condições: (1) cada linhagem descendente deve apresentar evidência filogenética de aparecimento em outra

[46] Linhagem ancestral que deu origem a todos os organismos da árvore filogenética, sendo estimada com base nas sequências observadas.

população geograficamente distinta, o que implica uma transmissão substancial. Para mostrar evidência filogenética, a nova linhagem deve: (a) apresentar uma ou mais substituições nucleotídicas compartilhadas da linhagem ancestral; (b) possuir, pelo menos, cinco genomas com >95% do genoma sequenciado; (c) apresentar, pelo menos, uma mudança nucleotídica compartilhada entre seus genomas; e (d) um valor de suporte (*bootstrap*) >70% para o ramo da árvore que leva à linhagem. (2) as linhagens identificadas no passo 1 podem agir como ancestrais para linhagens de vírus que subsequentemente emergem em outras áreas geográficas ou em momentos posteriores, desde que satisfaçam os critérios (a-d). Isto resulta em uma nova designação de linhagem (por exemplo, A.1.1); (3) o procedimento iterativo na etapa 2 pode proceder para um máximo de três subníveis (por exemplo, A.1.1.1), após o qual as novas linhagens descendentes recebem uma letra (em ordem alfabética inglesa de C), de modo que A.1.1.1.1 se tornaria C.1 e A.1.1.1.2 se tornaria C.2 (RAMBAUT et al., 2020a).

Portanto, essa nomenclatura é capaz de: (i) capturar padrões locais e globais de diversidade genética do vírus de maneira coerente; (2) rastrear linhagens emergentes à medida que se movem entre países e populações dentro de cada país; (3) ser suficientemente robusta e flexível para acomodar a nova diversidade de vírus à medida que ela é gerada; e (4) ser dinâmica, de modo que possa incorporar tanto o nascimento quanto a morte de linhagens virais ao longo do tempo (RAMBAUT et al., 2020a). A atribuição das linhagens globais (PANGO *lineages*) seguindo tal nomenclatura foi implementada no *software* pangolin (https://github.com/cov-lineages/pangolin), que é amplamente usado pela comunidade científica para a classificação dos genomas recém sequenciados.

Adicionalmente, o grupo responsável disponibiliza uma página onde podem ser feitos filtros por linhagem para obter os países com maior frequência, a quantidade de sequências atribuídas àquela linhagem, a data da primeira detecção, entre outros (https://cov-lineages.org/lineages.html).

Iniciativas como o Nextstrain (https://nextstrain.org/; (HADFIELD et al., 2018) fornecem um fluxo padronizado e interativo para a interpretação e visualização dos resultados filogenéticos e filogeográficos, sendo diariamente atualizado e gerenciado por um grupo ativo de pesquisadores e desenvolvedores. Diariamente, são amostrados genomas representativos e construída uma árvore filogenética global (https://nextstrain.org/ncov/global) e análises focadas nos continentes (por exemplo, América do Sul: https://nextstrain.org/ncov/south-america). Além do SARS-CoV-2, essa ferramenta também pode ser usada para a vigilância genômica de outros patógenos, já tendo sido aplicado aos vírus Influenza, *West Nile virus*, *Zika virus*, *Ebolavirus*, Dengue virus, entre outros.

A ferramenta CoV-GLUE (http://cov-glue.cvr.gla.ac.uk/#/replacement; SINGER et al., 2020) permite o acompanhamento de substituições não-sinônimas, inserções e deleções nas sequências do SARS-CoV-2 depositadas no GISAID, fornecendo uma interface interativa e a disponibilidade de vários filtros que permitem a visualização das mutações mais frequentes no mundo e em regiões específicas.

Uma ótima ferramenta para acompanhar relatórios de mutações e linhagens específicas é o outbreak.info (MULLEN et al., 2021a), que agrega a frequência destas em diferentes países ao longo do tempo e permite customizações para

investigações mais direcionadas. Também permite comparar as mutações comuns e diferentes que definem as linhagens.

No Brasil, pesquisadores da Fundação Oswaldo Cruz (Fiocruz) disponibilizam alguns gráficos que mostram a quantidade de genomas sequenciados no Brasil por mês e por estado, a frequência das principais linhagens no país, nas regiões geográficas e nos estados, a quantidade de linhagens por estado, entre outras informações (http://www.genomahcov.fiocruz.br/grafico/).

1.8.4. ESPALHAMENTO VIRAL NO TERRITÓRIO BRASILEIRO

No Brasil (Figura 8), o SARS-CoV-2 chegou oficialmente em 25 de fevereiro de 2020, em um viajante que retornava da Itália, e os primeiros esforços foram feitos tanto em nível nacional quanto regional para caracterizar as introduções internacionais e a transmissão viral comunitária nesta primeira onda epidêmica.

A dinâmica inicial de transmissão do SARS-CoV-2 no território brasileiro foi investigada por meio do sequenciamento de 490 genomas amostrados de modo proporcional ao número de casos por estado brasileiro até o final de abril de 2020. Foi determinado que: (i) B.1.1 e linhagens derivadas foram predominantes no início da pandemia; (ii) houve >100 introduções internacionais independentes no país; (iii) um movimento significativo do vírus entre as regiões brasileiras foi observado após restrições de viagens internacionais; e (iv) medidas não-farmacológicas foram capazes de reduzir o número de reprodução (R_0) de >3 para ≅1. Adicionalmente, a maioria (76%) desses genomas brasileiros pertencia a três clados que foram introduzidos da Europa entre 22 de fevereiro e 11 de março de 2020. O clado 1 circulava predominantemente no estado de São Paulo (n =159, 85,4%), enquanto

Figura 8. Mapa do Brasil, evidenciando as cinco regiões brasileiras e os 26 estados brasileiros somados ao Distrito Federal.
Fonte: MUNDO EDUCAÇÃO, 2021.

o clado 2 foi considerada a linhagem mais difundida espacialmente, representado por sequências de 16 estados do Brasil. O clado 3, por outro lado, concentrou-se no estado do Ceará (n=16, 89%), fazendo parte de um *cluster*[47] global composto principalmente por sequências da Europa (CANDIDO et al., 2020b).

Em Pernambuco, no Nordeste brasileiro, 88% das 101 sequências foram classificadas como linhagem B.1.1 e foram observados seis clados locais, com a estimativa de, pelo menos, cinco eventos de importação internacional (PAIVA et al., 2020). Em Minas Gerais (Sudeste), 92,5% dos 40 genomas de março de 2020

[47] Conceito similar a clado, no qual é formado um grupo de sequências de diferentes pacientes que compartilham um ancestral comum.

pertenciam à linhagem B (principalmente B.1.1) e uma análise epidemiológica revelou que a distribuição de casos e mortes foi mais uniforme espacialmente, enquanto em outros estados do Sudeste foi mais centralizada em torno das capitais (XAVIER et al., 2020). No Amazonas, 250 genomas de diferentes municípios do estado foram amostrados entre março de 2020 e janeiro de 2021 e observou-se que a primeira fase de crescimento exponencial foi impulsionada, principalmente, pela disseminação da linhagem B.1.195, que foi gradualmente substituída pela linhagem B.1.1.28 (NAVECA et al., 2021a).

Dentre os três clados brasileiros relatados por CANDIDO et al. (2020), dois deles se espalharam amplamente pelo Brasil, sendo denominados como linhagens B.1.1.33 e B.1.1.28. Após um amplo predomínio destas linhagens por grande parte do ano de 2020, as linhagens P.1 (FARIA et al., 2021; NAVECA et al., 2021) e P.2 (VOLOCH et al., 2021) — ambas derivadas de B.1.1.28 — substituíram-nas rapidamente. A variante P.1 rapidamente foi classificada como uma VOC (denominada Gama) devido à constelação de mutações na proteína *Spike*, bem como sua associação a maiores cargas virais (NAVECA et al., 2021a), transmissibilidade, evasão imune e letalidade (FARIA et al., 2021), suplantando todas as demais linhagens no Brasil até meados de 2021. Durante este período, também foram identificadas mais algumas novas linhagens no Brasil, como: P.4 (BITTAR et al., 2021) no interior de São Paulo, N.9 (RESENDE et al., 2021c) e N.10 (RESENDE et al., 2021b) majoritariamente no Norte e Nordeste e P.1.2 no Sudeste (FRANCISCO JUNIOR et al., 2021b) e Sul (FRANCESCHI et al., 2021b) do país, entre outras; bem como foram detectadas VOCs importadas ao país, como: B.1.1.7 (Alfa) (CLARO et al., 2021) a B.1.351 (Beta) (SLAVOV et al., 2021), e B.1.617.2

(Delta) (LAMARCA et al., 2021a), a qual passou a predominar no Brasil juntamente com suas sublinhagens em Agosto de 2021.

1.8.5. EMERGÊNCIA DE VARIANTES DE INTERESSE E PREOCUPAÇÃO

A mutação é um aspecto comum no ciclo de vida de um vírus de RNA. Como estes vírus empregam uma RNA polimerase intrinsecamente propensa a erros na replicação, seus genomas acumularão mutações durante cada ciclo de cópia. Além disso, estes ciclos ocorrem na ordem de horas, assegurando que uma população diversificada de vírus será gerada mesmo dentro de um único hospedeiro infectado. Contudo, espera-se que a maioria das mutações tenha impacto negativo no funcionamento do vírus fazendo com que estas sejam removidas por seleção natural (GRUBAUGH; PETRONE; HOLMES, 2020). Sendo assim, embora uma mutação (ou uma combinação delas) que muda a forma como um vírus é transmitido, sua patogenicidade, sua gama de hospedeiros[48] ou sua antigenicidade[49] possa aparecer prontamente em uma população viral, não se espalhará rapidamente, a menos que seja evolutivamente vantajosa (GRUBAUGH; PETRONE; HOLMES, 2020; HARVEY et al., 2021; PEACOCK et al., 2021b). O espalhamento de mutações e variantes virais é governado por uma série de fatores, incluindo processos demográficos, como crescimento populacional, expansão geográfica, efeitos fundadores[50], deriva genética, bem como efeitos da seleção positiva em casos de aumento de transmissibilidade (VOLZ et al., 2021a).

[48] Amplitude de organismos que um parasita é capaz de infectar.

[49] A capacidade de induzir uma resposta imune.

[50] Perda de variação genética que ocorre quando uma nova população é estabelecida por um número muito pequeno de indivíduos de uma população maior.

A diversidade limitada do SARS-CoV-2 relatada durante o ano de 2020 pode ser parcialmente atribuída ao mecanismo de revisão durante a replicação do RNA (*proofreading)* realizado por uma exoribonuclease presente no complexo não-estrutural (nsp10-nsp14), a qual aumenta significativamente a fidelidade na incorporação de nucleotídeos, mantendo a integridade do seu genoma especialmente grande para vírus de RNA (LIU et al., 2021b; MA et al., 2015; OGANDO et al., 2020). Por este motivo, inicialmente foi postulado que as vacinas baseadas em uma única sequência da proteína *Spike*, provavelmente, gerariam proteção imunológica para todas as variantes circulantes (DEARLOVE et al., 2020).

Porém, desde o final de 2020, algumas linhagens do SARS-CoV-2 emergiram, carregando principalmente mutações na glicoproteína *Spike* (S), que é responsável por mediar a interação com o receptor hACE2. Trata-se, portanto, do principal alvo dos anticorpos neutralizantes (anti-SARS-CoV-2) e do desenvolvimento de vacinas (WALLS et al., 2020). Atualmente, existem cinco linhagens de maior preocupação mundial: B.1.1.7, B.1.351, P.1, B.1.617.2 e B.1.1.529/BA.1. Para evitar a estigmatização dos países de origem e promover a facilitação da comunicação entre especialistas e público leigo, foi criada pela OMS uma nomenclatura baseada em letras gregas. Desta forma, as VOCs são chamadas Alfa, Beta, Gama, Delta e Ômicron, respectivamente (WORLD HEALTH ORGANIZATION, 2021d).

A primeira (Alfa) surgiu no Reino Unido em meados de setembro de 2020 e se caracteriza por 14 substituições de aminoácidos de linhagem específica e se espalhou rapidamente pelo Reino Unido e pela Europa desde sua primeira aparição (RAMBAUT et al., 2020b). Duas substituições presentes nesta linhagem merecem

atenção especial: N501Y e P681H. O sítio 501, no qual ocorre a mutação de uma asparagina para uma tirosina (N501Y), está localizado na porção S1 do RBD e é um dos quatro principais sítios de contato que interagem com hACE2, já o sítio 681, onde ocorre a substituição de uma prolina por uma histidina (P681H), é um dos quatro sítios que compõem a inserção que cria um sítio de clivagem de furinas entre S1 e S2, que não é encontrado nos coronavírus proximamente relacionados (XIA et al., 2020).

A segunda variante (Beta), provavelmente, emergiu na África do Sul em agosto de 2020 e abriga três mutações principais no RBD: K417N, E484K e N501Y (TEGALLY et al., 2021). Ainda em 2020, emergiu a terceira VOC (Gama), denominada P.1. Esta é derivada da B.1.1.28 e foi identificada inicialmente em japoneses que estavam em Manaus (estado do Amazonas, Brasil) e retornaram infectados ao seu país. Possui as mesmas três mutações presentes no RBD da linhagem primeiramente detectada na África do Sul, exceto pela substituição na posição 417, onde a lisina original (K) é substituída por uma treonina (T), em vez de uma asparagina (N). Porém, seu surgimento foi independente das demais linhagens (FARIA et al., 2021). Na primeira metade de 2021, a linhagem B.1.617.2, primeiramente detectada na Índia, também foi caracterizada como VOC, principalmente, por possuir uma constelação de mutações na proteína *Spike* (especialmente L452R e P681R) (PEACOCK et al., 2021b), por seu amplo espalhamento pelo mundo suplantando inclusive outras VOCs (MULLEN et al., 2021b) e pela reduzida sensibilidade a anticorpos de indivíduos vacinados (PLANAS et al., 2021b).

Em novembro de 2021, a variante B.1.1.529/BA.1 (Ômicron) foi identificada, inicialmente, em países com baixas taxas de vacinação (África do Sul e em Botswana), sendo rapidamente associada ao aumento de casos nessas localidades. Além de mutações na proteína S comuns a outras VOCs e VOIs, esta possui muitas substituições adicionais (*e. g.*, A67V, N211I, Δ212, G339D, S371L, S373P, S375F, Q493R, G496S, Q498R, Y505H, T457K, N764K, entre outras). A rápida identificação e caracterização permitiram o acompanhamento em tempo real de seu espalhamento globalmente. Até a metade de dezembro de 2021, foram depositadas mais de 7 mil sequências de 67 países no banco de dados GISAID. Sendo assim, muito embora estudos ainda estejam em andamento, extrapola-se que se trata de uma variante mais transmissível e capaz de evadir a resposta imune devido às suas mutações e sua rápida difusão, já substituindo a variante Delta e se tornando predominante mundialmente (KARIM; KARIM, 2021).

Além das cinco VOCs, linhagens como B.1.427/429 (Califórnia, EUA) (DENG et al., 2021), B.1.526 (Nova Iorque, EUA) (LASEK-NESSELQUIST et al., 2021a), C.37 / Lambda (Peru), B.1.617 (Índia), B.1.621 / Mu (Colômbia) e P.2 (Brasil) (VOLOCH et al., 2021) foram categorizadas como VOIs. O surgimento independente destas linhagens (VOCs e VOIs) com substituições compartilhadas — como K417N/T, L452R, E484K, D614G, N501Y, P681H/R (Figura 9) — sugere processos evolutivos convergentes[51] e uma grande mudança global nos padrões seletivos do SARS-CoV-2 ocorrendo desde o final de 2020 (FARIA et al., 2021; MARTIN et al., 2021; RAMBAUT et al., 2020b; TEGALLY et al., 2021). Tais padrões

[51] Processo pelo qual organismos não intimamente relacionados (não monofiléticos), adquirem independentemente características semelhantes na necessidade de adaptação a ambientes similares.

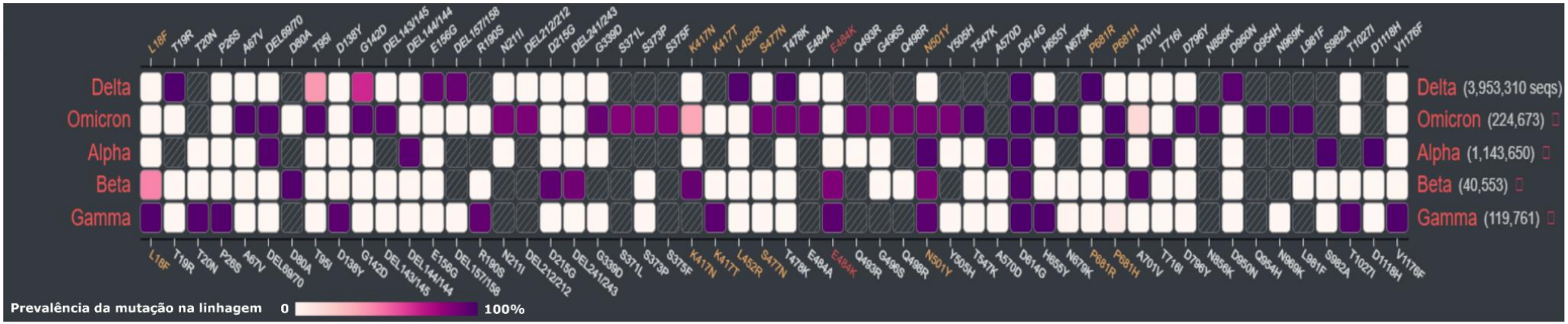

Figura 9. Mutações na proteína *Spike* do SARS-CoV-2 compartilhadas entre as cinco principais VOCs. Linhas representam as diferentes VOCs e colunas mostram as substituições. O gradiente de cores iniciando em rosa claro (0%) até roxo escuro (100%) indica a frequência da substituição dentro da linhagem. Dados atualizados em 11/01/2022. Ao longo de 2021, várias linhagens foram categorizadas como VOIs, mas, ao serem superadas pelas VOCs, foram removidas desta denominação, restando atualmente Lambda e Mu como VOIs juntamente com as VOCs Alfa (B.1.1.7), Beta (B.1.351), Gama (P.1), Delta (B.1.617.2) e Ômicron (B.1.1.529/BA.1).
Fonte: MULLEN et al. (2021c).

evolutivos virais podem comprometer a eficácia das campanhas de vacinação mundiais iniciadas no final de 2020, mas que seguem ritmos diferentes conforme o planejamento, os investimentos e a situação da pandemia nos diferentes países.

Após estudos aprofundados com as mutações presentes especialmente em VOCs (Tabela 3), há evidências importantes da alteração da antigenicidade da proteína *Spike* do SARS-CoV-2. Essas substituições e deleções de aminoácidos na proteína *Spike* que impactam a ação dos anticorpos neutralizantes estão presentes em frequências significativas na população global do vírus. Além disso, há evidências de variantes que exibem resistência à imunidade mediada por anticorpos produzidos pelas vacinas. Como em geral tais variantes também estão associadas à maior transmissibilidade, observa-se um aumento preocupante em sua frequência pelo mundo. Contudo, historicamente, a evolução da resistência às vacinas normalmente ocorre mais lentamente e menos frequentemente do que a evolução da resistência a medicamentos como os antimicrobianos, por exemplo (KENNEDY; READ, 2017). Ainda assim, diante da reduzida eficácia das vacinas frente às VOCs, os fabricantes devem preparar plataformas para uma possível atualização das sequências utilizadas nas vacinas, acompanhando de perto as mudanças genéticas e antigênicas na população viral e os experimentos que elucidam os impactos fenotípicos dessas mutações (HARVEY et al., 2021).

Tabela 3. Principais características moleculares, epidemiológicas e clínicas das cinco VOCs melhor caracterizadas mundialmente desde o início da pandemia

Linhagem PANGO	B.1.1.7	B.1.351	P.1	B.1.617.2	B.1.1.529/BA.1
Clado GISAID	GRY (inicialmente GR/501Y.V1)	GH/501Y.V2	GR/501Y.V3	G/452R.V3	GRA
Clado Nextstrain	20I/S:501Y.V1	20H/S:501Y.V2	20J/S:501Y.V3	21A/S:478K	21K/L/M
Nome OMS	Alpha	Beta	Gamma	Delta	Ômicron
Primeiras amostras documentadas	Setembro 2020	Maio 2020	Novembro 2020	Outubro 2020	Novembro 2021
País de origem provável	Reino Unido	África do Sul	Brasil	Índia	Botswana / Hong Kong / África do Sul
Países com sequências depositadas	175	115	72	160	115
Mutações-chave	ORF1ab: del3675/3677 S: del69/70, del144/145, N501Y, A570D, P681H	ORF1ab: del3675/3677 S: K417N, E484K, N501Y	ORF1ab: del3675/3677 S: K417T, E484K, N501Y	S: del157/158, L452R, T478K, P681R	>25 mutações definidoras em S
Transmissibilidade	+++	+	++	+++	++++
Evasão imune	—	++++	+++	++	+++++
Redução na efetividade de vacinas*	+	+++	++	+++	++++

Para transmissibilidade, evasão imune e efetividade das vacinas é feito um comparativo entre as cinco VOCs, de modo que aquela com mais sinais possui efeitos mais significativos. Informações mais detalhadas sobre estes tópicos serão abordadas mais detalhadamente na discussão.

Fonte: MULLEN et al. (2021a); O'TOOLE et al. (2021).

2. OBJETIVOS

2.1. OBJETIVO GERAL

Realizar análises genômicas de amostras isoladas de SARS-CoV-2 a fim de compreender a distribuição de mutações e linhagens virais em nível municipal (Esteio, Rio Grande do Sul [RS], Brasil), estadual (RS) e nacional (Brasil).

2.2. OBJETIVOS ESPECÍFICOS

- Realizar análises mutacionais de *variant calling* a fim de identificar mutações sofridas pelo vírus durante a pandemia da COVID-19;
- Realizar análises filogenéticas de máxima verossimilhança em genomas de SARS-CoV-2 para identificar clados bem-suportados;
- Realizar estudos filogeográficos e filodinâmicos Bayesianos dos genomas de SARS-CoV-2, a fim de compreender os padrões geográficos e temporais de dispersão local e sua relação com a dispersão regional, nacional e global do vírus;
- Discutir o impacto de mutações positivamente selecionadas para o *fitness* viral, aumento de transmissibilidade e escape imunológico frente à infecção natural e vacinação.

3. CAPÍTULO I

O manuscrito que constitui este capítulo, intitulado “Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil” objetivou acompanhar a evolução molecular e a propagação do SARS-CoV-2 no município de Esteio (RS) usando inferências filogenéticas e filodinâmicas utilizando 21 novos genomas amostrados entre maio e outubro de 2020 no contexto regional e global. Encontra-se publicado na revista BMC Genomics (https://bmcgenomics.biomedcentral.com/), com fator de impacto JCR 2021 = 3,969 e Qualis/CAPES = A2. O manuscrito e os materiais suplementares estão disponíveis na íntegra (*Open access*) no seguinte *link*: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-021-07708-w. Todas as análises descritas no manuscrito, assim como a sua redação, foram realizadas pelo aluno Vinícius Bonetti Franceschi, sendo os demais autores responsáveis por colaborações na escrita ou análises, bem como na sua orientação e obtenção de fomento. Abaixo, todas as páginas do manuscrito publicado foram anexadas para compor o Capítulo I da presente dissertação.

Franceschi *et al. BMC Genomics* (2021) 22:371
https://doi.org/10.1186/s12864-021-07708-w

BMC Genomics

RESEARCH Open Access

Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil

Vinícius Bonetti Franceschi[1], Gabriel Dickin Caldana[2], Amanda de Menezes Mayer[1], Gabriela Bettella Cybis[3], Carla Andretta Moreira Neves[2], Patrícia Aline Gröhs Ferrareze[2], Meriane Demoliner[4], Paula Rodrigues de Almeida[4], Juliana Schons Gularte[4], Alana Witt Hansen[4], Matheus Nunes Weber[4], Juliane Deise Fleck[4], Ricardo Ariel Zimerman[5], Lívia Kmetzsch[1], Fernando Rosado Spilki[4] and Claudia Elizabeth Thompson[1,2,6*]

Abstract

Background: Brazil is the third country most affected by Coronavirus disease-2019 (COVID-19), but viral evolution in municipality resolution is still poorly understood in Brazil and it is crucial to understand the epidemiology of viral spread. We aimed to track molecular evolution and spread of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Esteio (Southern Brazil) using phylogenetics and phylodynamics inferences from 21 new genomes in global and regional context. Importantly, the case fatality rate (CFR) in Esteio (3.26%) is slightly higher compared to the Rio Grande do Sul (RS) state (2.56%) and the entire Brazil (2.74%).

Results: We provided a comprehensive view of mutations from a representative sampling from May to October 2020, highlighting two frequent mutations in spike glycoprotein (D614G and V1176F), an emergent mutation (E484K) in spike Receptor Binding Domain (RBD) characteristic of the B.1.351 and P.1 lineages, and the adjacent replacement of 2 amino acids in Nucleocapsid phosphoprotein (R203K and G204R). E484K was found in two genomes from mid-October, which is the earliest description of this mutation in Southern Brazil. Lineages containing this substitution must be subject of intense surveillance due to its association with immune evasion. We also found two epidemiologically-related clusters, including one from patients of the same neighborhood. Phylogenetics and phylodynamics analysis demonstrates multiple introductions of the Brazilian most prevalent lineages (B.1.1.33 and B.1.1.248) and the establishment of Brazilian lineages ignited from the Southeast to other Brazilian regions.

Conclusions: Our data show the value of correlating clinical, epidemiological and genomic information for the understanding of viral evolution and its spatial distribution over time. This is of paramount importance to better inform policy making strategies to fight COVID-19.

Keywords: COVID-19, Severe acute respiratory syndrome coronavirus 2, Infectious diseases, Sequencing, Molecular epidemiology

* Correspondence: cthompson@ufcspa.edu.br; thompson.ufcspa@gmail.com
[1]Center of Biotechnology, Graduate Program in Cell and Molecular Biology (PPGBCM), Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil
[2]Graduate Program in Health Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, RS, Brazil
Full list of author information is available at the end of the article

Background

In December 2019, the causative agent of Coronavirus disease-2019 (COVID-19) pandemic named Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), emerged in Wuhan, Hubei, China [1]. As of 28 April, 2021, there are 148.963.836 confirmed cases and 3.140.213 million deaths in 192 countries [2]. Unprecedented international efforts of viral sequencing have allowed the submission of ~ 1.3 million genomes in the Global Initiative on Sharing All Influenza Data (GISAID) up to date [3], which are now available for studies of genomic epidemiology to follow the evolutionary history and dynamics of SARS-CoV-2 through space and time. In this sense, some important studies were already conducted in highly-affected countries, including USA [4–7], Italy [8], Netherlands [9], Australia [10, 11], and Brazil [12–14].

By using a nomenclature developed to capture local and global patterns of genetic diversity of the virus, two main lineages (A and B) were identified, both originated in Wuhan and with simultaneous spreading around the world [15]. The dynamics of viral transmission in the Brazilian territory was investigated through the sequencing of ~ 500 genomes until the end of April, 2020. It was determined that: (i) B.1 and derived lineages were predominant at the beginning of the pandemic; (ii) > 100 independent international introductions occurred in the country; (iii) a significant movement of the virus among the Brazilian regions was observed after international travel restrictions; and (iv) non-pharmacological measures were able to reduce the reproduction number (R_0) from > 4 to ≈1 [12].

The genetic diversity of SARS-CoV-2 has been extensively studied, evidencing the presence of recurrent mutations such as S: D614G, S:E484K, S:N501Y across the world [16–18], related to increased pathogenicity and transmissibility (higher viral loads, increased replication on lung epithelial cells, and enhanced binding affinity) [19–24]. Furthermore, the E484K mutation was associated with immune evasion from neutralizing antibodies produced in response to currently available vaccines [25–27]. In addition to SARS-CoV-2 mutations, co-infection with other pathogens (e. g. *Staphylococcus aureus*, *Haemophilus influenzae*, rhinovirus/enterovirus, respiratory syncytial virus, and seasonal coronaviruses) may be associated with poor clinical outcomes, although these rates appear to be limited [28–31].

The viral mutations are not the only factor affecting the COVID-19 pathology and SARS-CoV-2 infectious capacity. Human host factors such as: (i) rare genetic variants governing interferon immunity [32], (ii) DNA polymorphisms in key host factors (e. g. Angiotensin-converting enzyme 2 [ACE2] and transmembrane protease serine 2 [TMPRSS2]) [33, 34], (iii) heritage and ethnicity [35], (iv) the presence of comorbidities (hypertension, diabetes, obesity, and immunological diseases) [36, 37] were already associated to increased disease severity, although more integrative studies are still needed to identify the relative contribution of each of these factors. By analyzing 27 candidate genes and Human leukocyte antigen (HLA) alleles in 954 admixed Brazilian exomes, 395 nonsynonymous variants were found. Of these, six were previously associated with the rate of infection or clinical prognosis of COVID-19. Seventy were identified exclusively in the Brazilian sample, and seven (10%) of these were predicted to affect protein function using in silico analysis [38].

As of March 10, 2020, a 60-year-old man who had been in Italy, became the first confirmed case in the southernmost state of Brazil (Rio Grande do Sul - RS) [39], which is the most populous state in the South Region of Brazil and the fifth in the whole country (~ 11.5 million inhabitants) [40]. As of April 28, 2021, Brazil has ~ 9.7% of worldwide cases (~ 14.4 million) and is the third worst-hit country [2]. The RS State reported ~ 956,030 cases and 24,458 deaths, with ~ 8% of cases requiring hospitalization [39]. The municipality of Esteio, located in the metropolitan region of RS capital, reported 9272 cases (total population: 83,202) and 302 deaths [41]. As Esteio is a commuter town, many workers move to and return from the state capital every day. Importantly, the case fatality rate (CFR) in Esteio (3.26%) was slightly higher compared to the RS state (2.56%), and both are greater than previous CFR estimates (~ 1%) [42, 43].

Thus, we aimed to characterize the main circulating lineages in Esteio (RS, Brazil) and their relationship with global, national and regional lineages using phylogenetics and phylodynamics inference from 21 SARS-CoV-2 genome sequences, including the investigation of putative viral mutations related to poor outcomes. Additionally, due to our typical subtropical climate and therefore high occurrence of respiratory infections, we investigated the occurrence of co-infections with other viral pathogens in these samples. The choice of a small municipality as the target of this study was important since we could more easily and precisely follow infected individuals, allowing a more detailed surveillance on the spread of the virus and detection of variability.

Results

SARS-CoV-2 genomes were obtained with an average coverage depth of 1380.51× (median: 213.28×, standard deviation: 2296.16×) (Additional File 1). All consensus genomes passed the quality control steps. Considering the 21 samples (Table 1), 52.4% of the patients were female and the mean age was 41.3 years (range: 19–72 years). The mean Cycle threshold (Ct) values was 16.12 (range: 12.53–19.94). None of the patients reported

Table 1 Epidemiological data of the 21 sequenced samples from Esteio, RS, Brazil

GISAID Accession	Ct value	Collection month	Age range	Sex	Clinical status	Pangolin Lineage	Nextstrain Clade
EPI_ISL_831678	16.15	May 2020	20–30	M	Mild	B.1.1.33	20B
EPI_ISL_831474	16.48	June 2020	20–30	M	Mild	B.1.1.33	20B
EPI_ISL_831645	16.72	June 2020	60+	M	Moderate	B.1.1.248	20B
EPI_ISL_831646	16.45	June 2020	50–60	F	Mild	B.1.1.33	20B
EPI_ISL_831660	15.53	June 2020	20–30	F	Mild	B.1.1.248	20B
EPI_ISL_831681	16.72	July 2020	40–50	F	Mild	B.1.1	19A
EPI_ISL_831683	15.50	July 2020	30–40	F	Moderate	B.1.1.33	20B
EPI_ISL_831685	16.65	July 2020	10–20	F	Mild	B.1.1.33	20B
EPI_ISL_831688	15.52	July 2020	40–50	M	Mild	B.1.1.248	20B
EPI_ISL_831689	14.37	August 2020	30–40	M	Mild	B.1.1.248	20B
EPI_ISL_831892	15.12	August 2020	60+	M	Mild	B.1.1.33	20B
EPI_ISL_831898	14.14	August 2020	40–50	M	Mild	B.1.1.33	20B
EPI_ISL_831913	12.53	August 2020	30–40	F	Mild	B.1.1.49	20B
EPI_ISL_831938	15.32	August 2020	50–60	M	Mild	B.1.1.248	20B
EPI_ISL_831939	15.58	September 2020	40–50	M	Mild	B.1.1	20B
EPI_ISL_831940	17.10	September 2020	20–30	F	Mild	B.1.1.33	20B
EPI_ISL_832009	14.99	September 2020	30–40	F	Mild	B.1.1.248	20B
EPI_ISL_832010	18.31	October 2020	20–30	F	Mild	B.1.1.248	20B
EPI_ISL_832011	17.33	October 2020	40	M	Mild	B.1.1.248	20B
EPI_ISL_832012	19.94	October 2020	60+	F	Mild	B.1.1.33	20B
EPI_ISL_832013	18.15	October 2020	40–50	F	Mild	B.1.1	20B

All samples were nasopharyngeal swabs collected from patients of the municipality of Esteio. *Sample ID* Sample identifier; *M* Male; *F* Female

interstate or international travels. Regarding clinical status, 90.5% of patients were classified as mild infection and 9.5% as moderate.

Virome analysis

To investigate whether the severity of the infection presented by the patients could be linked to co-infection with another respiratory viral pathogen, we analyzed the viral composition of these samples. We found through taxonomic classification at the level of nucleotides and amino acids that none of the investigated patients had a viral infection other than COVID-19. All samples had high assignment (> 99%) to the *Betacoronavirus* genus.

SARS-CoV-2 mutations found in the patient samples and lineages

The number of SNPs per genome ranged from 1 to 19 (mean: 12.8, median: 14.0) (Fig. 1a). All genomes were different from each other. We identified 80 different SNPs in the 21 genomes analyzed. Thirty two (40.0%) of them were observed in more than one sample (Supplementary Table 1). Of these, 18 (56.2%) were missense (non-synonymous). High frequency (> 5 genomes) missense mutations were observed in the following positions (absolute nucleotide position: amino acid inside the gene): ORF1ab (C12053T: L3930F), Surface (S) glycoprotein (A23403G: D614G; G25088T: V1176F), ORF6 (T27299C: I33T), and Nucleocapsid (N) protein (GGG28881-28883 ACC: RG203-204KR; T29148C: I292T) (Fig. 1b). A new mutation in the Receptor Binding Domain (RBD) of the spike protein (G23012A: E484) was found in two genomes (9.5%) (GISAID IDs: EPI_ISL_832010 and EPI_ISL_832013) from mid-October 2020. Since the municipality of Esteio has a higher CFR (3.26) than the national CFR (as the Brazilian states of São Paulo, Amazonas, Pernambuco, and Rio de Janeiro that were highly affected by the pandemic) (Supplementary Table 2), it is possible that the emergence of new viral mutations and lineages combined with genetic factors in these populations [38, 44] are partially associated with differential COVID-19 severity.

We were able to identify four different viral lineages, all descendants of lineage B (Table 1). Two lineages associated with community-transmission in Brazil, B.1.1.33 ($n = 9$; 42.9%) and B.1.1.248 (reassigned later to B.1.1.28) ($n = 8$; 38.1%) were the most prevalent. All B.1.1.33 sequences shared T27299C (ORF6:I33T), GGG28881-28883AAC (N:RG203-204KR), and T29148C (N:I292T) mutations. All B.1.1.248 sequences shared C241T (5′ UTR), C3037T (ORF1ab nsp3:F924), C12053T (ORF1ab nsp7:L3930F), C14408T (ORF1ab RdRp:L4715), A23403G

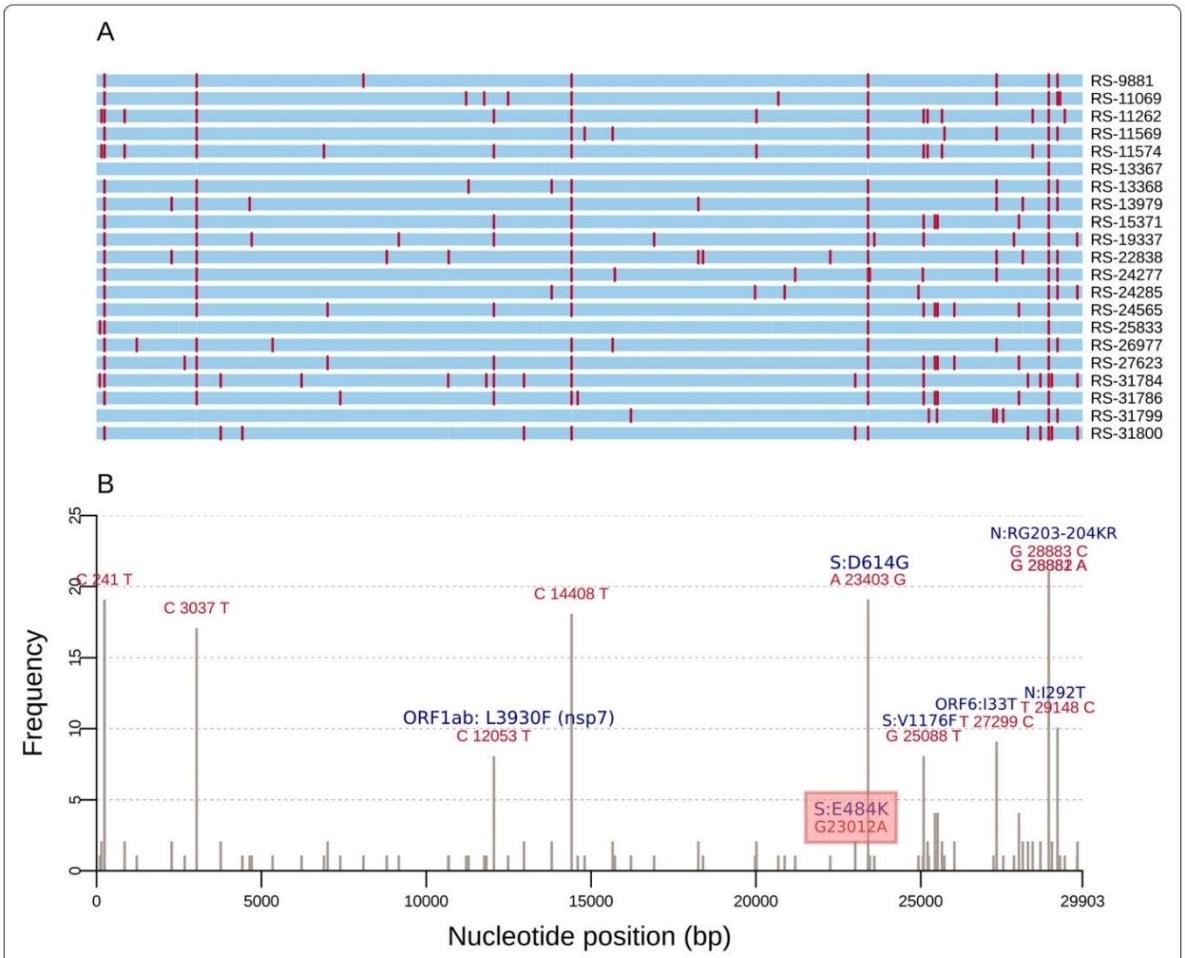

Fig. 1 Mutations of SARS-CoV-2 genomes from patients of Esteio, RS, Brazil. **a** Genome map for the 21 genomes sequenced (indicated in the right). SNPs are colored in red. **b** Frequency of SNPs per SARS-CoV-2 genome position among the 21 genome sequences. These mutations correspond to the red lines in (**a**) and only those represented by > 5 genomes are indicated above the bars, regardless of being synonymous or missense. The abbreviations of genes affected and respective amino acid changes (where applicable) are indicated above the nucleotide changes

(S:D614G), G25088T (S:V1176F), and GGG28881-28883 AAC (N:RG203-204KR) replacements.

Both lineages are represented by > 70% of Brazilian sequences in global context (https://cov-lineages.org/). They are the most representative lineages in the South Region of Brazil and in the whole country (Fig. 2, Supplementary Figures 1, 2 and 3). Three genomes were classified as B.1.1 lineage, which are the most globally widespread lineages characterized by RG203-204KR mutations in the nucleocapsid phosphoprotein (https://cov-lineages.org [15];). Finally, a relatively rare lineage (B.1.1.49) mostly found in Wales and Denmark was also assigned.

After inspecting these sequences assigned to global lineages (B.1.1 and B.1.1.49), we verified that in all cases there were characteristic mutations of B.1.1.248 and B.1.1.33 lineages flagged as undetermined bases (N character; depth of coverage (DP) < 10) in the consensus genome. After reclassifying these sequences using low coverage variants, two were attributed to B.1.1 lineage, one to B.1.1.248 and one to B.1.1.33 (Additional File 2). Therefore, due to the absence of confirmed relationships with other countries and the presence of other defining-lineage variants with low coverage, it seems probable that these sequences are also the result of community transmission in Brazil and were not introduced independently in the municipality of Esteio from other countries.

We also detected two novel epidemiologically-related clusters until then unknown. Two patients had three unique mutations in genome positions 25,207 (in the S2 subunit of spike), 25,642 (ORF3a), and 28,393

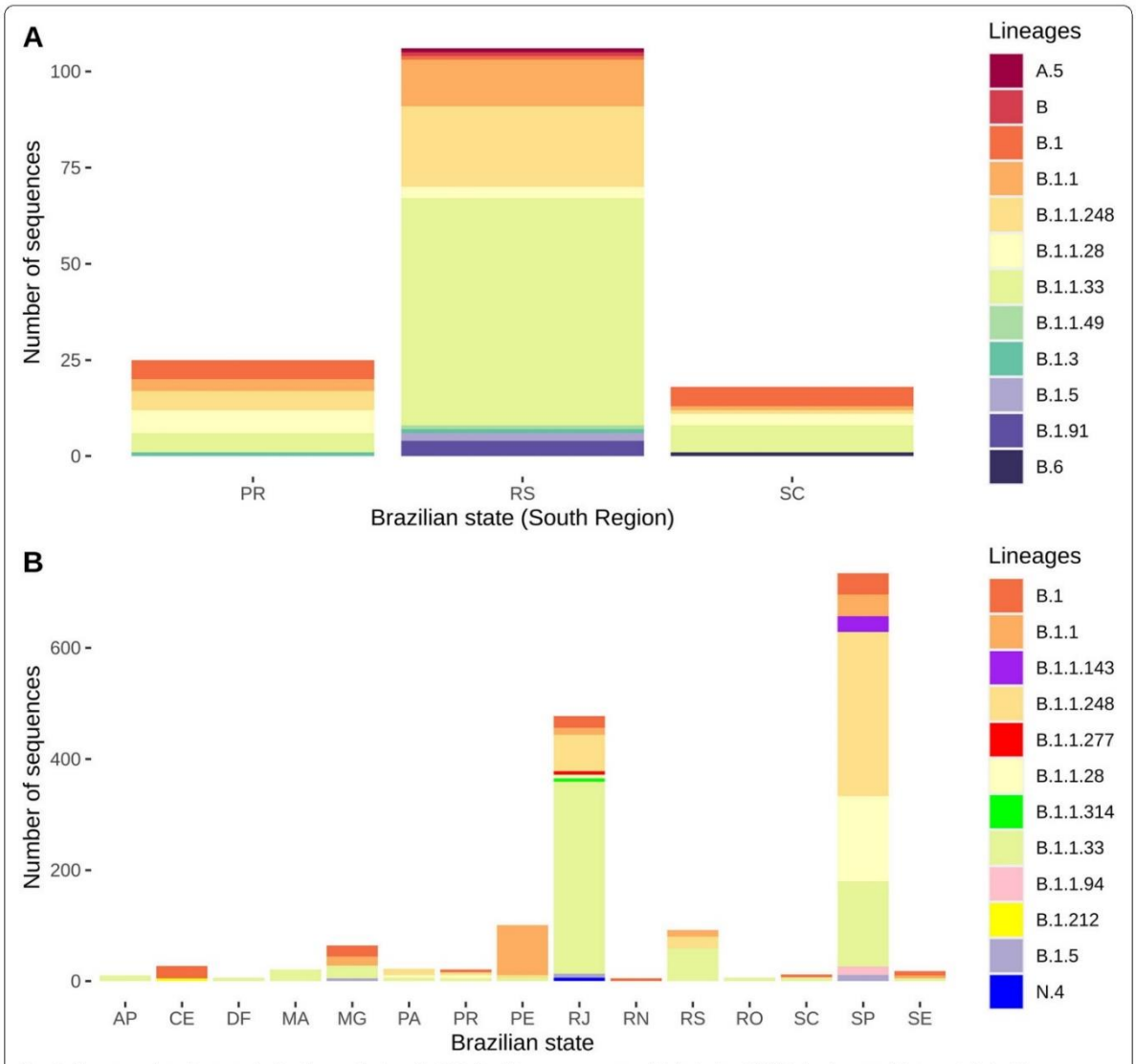

Fig. 2 Most prevalent lineages in Brazil considering all 1778 Brazilian sequences available in the GISAID database. **a** All sampled viral lineages from Southern Brazil, including Paraná (PR), Rio Grande do Sul (RS) and Santa Catarina (SC) states. The RS sequences include the 21 viral genomes from this work and 85 from GISAID. **b** Most prevalent lineages from Brazilian states. Only states that have more ≥5 sequences labelled as the lineage under consideration were included. AP = Amapá; CE = Ceará; DF=Distrito Federal; MA = Maranhão; MG = Minas Gerais; PA = Pará; PE = Pernambuco; RJ = Rio de Janeiro; RN = Rio Grande do Norte; RO = Rondônia; SP=São Paulo; SE = Sergipe

(Nucleocapsid), all synonymous substitutions. The patients both live in the same neighborhood, about 100 m from each other and did the test in a 2-day interval. We also identified another cluster of four patients characterized by three unique mutations: 25429 (ORF3a: V13L), 25,509 (ORF3a), and 27,976 (ORF8: H28R). The tests of these patients were performed in a 3-month interval (July 15–October 13), suggesting a fixation of these mutations through time, possibly forming a new sublineage. The two clusters are linked to viruses belonging to B.1.1.248 lineage, suggesting the existence of specific mutation signatures even within lineages.

Phylogenetics analysis

After running the Nextstrain pipeline for quality control and subsampling, we obtained 3758 time-, geographical- and genetic-representative genomes to proceed phylogenetic inferences. Of these, 393 were from Africa, 800 were from Asia, 1203 from Europe, 235 from North America,

127 from Oceania, and 1000 from South America. Considering the latter, 609 were from Brazil and 98 from Rio Grande do Sul (21 from this study plus 77 from GISAID that passed the quality control criteria) (Additional File 3).

The time-resolved ML phylogeny confirmed that the majority of the time-representative sequences from Esteio are the result of community transmission within Brazil. The sequences grouped mostly in two clades (Fig. 3a, Supplementary Figure 4) corresponding to lineages B.1.1.248 and B.1.1.33. Clade 1 comprised 147 sequences: 20 from RS, 58 from SP, 46 from RJ, 2 from PR, and 2 from SC, therefore mostly widespread in

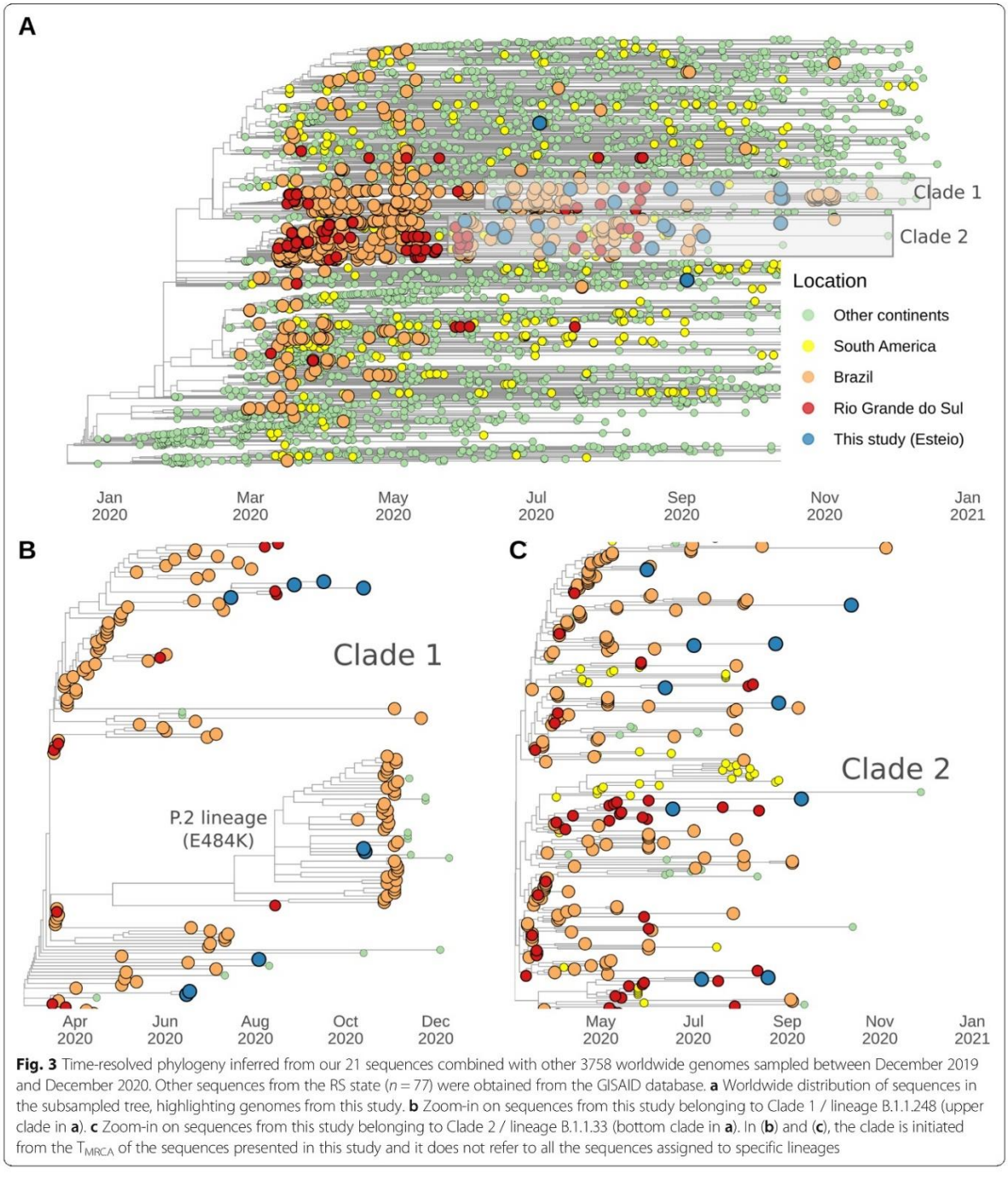

Fig. 3 Time-resolved phylogeny inferred from our 21 sequences combined with other 3758 worldwide genomes sampled between December 2019 and December 2020. Other sequences from the RS state ($n = 77$) were obtained from the GISAID database. **a** Worldwide distribution of sequences in the subsampled tree, highlighting genomes from this study. **b** Zoom-in on sequences from this study belonging to Clade 1 / lineage B.1.1.248 (upper clade in **a**). **c** Zoom-in on sequences from this study belonging to Clade 2 / lineage B.1.1.33 (bottom clade in **a**). In (**b**) and (**c**), the clade is initiated from the T_{MRCA} of the sequences presented in this study and it does not refer to all the sequences assigned to specific lineages

Brazilian states (Fig. 3b). Clade 2 included 277 genomes: 50 from RS, 72 from RJ, 20 from SP, 5 from SC, 30 from Chile, 10 from Argentina, and 6 from Uruguay (Fig. 3c). These results suggest a clade distributed through South American countries. Esteio sequences are relatively evenly distributed through these clades mostly represented by Brazilian and RS genomes. Exceptions to this observation were the two previously described epidemiologically-linked clusters, whose sequences grouped together in the B.1.1.248 lineage as expected (Fig. 3b). Given the low mutation rate of SARS-CoV-2 (6.59×10^{-4} substitutions/site/year, ~19 mutations per year) (Supplementary Figure 5), we believe that this would indicate at least three introductions of lineage B.1.1.248 and six introductions of lineage B.1.1.33 in the municipality of Esteio, probably from other locations in Brazil, and a national movement of the virus even to more distant places like the southernmost state of Brazil. Likewise, despite the large representativeness of Brazilian samples within these two major clades, we also found other sequences from Asia, Europe, Oceania, and South America. Therefore, sequences from these clades seem to have been directly transmitted from Brazil to other countries.

Phylodynamics and phylogeographic analysis

The estimate for population exponential growth rate for B.1.1.248 was 1.141 (95% Highest Posterior Density (HPD) interval: 0.4436–1.8268), while for B.1.1.33 it was 2.5871 (95% HPD: 2.062–3.0872). This can be taken as preliminary evidence that the B.1.1.33 lineage initially spread faster than B.1.1.248, since most of the coalescent events inform earlier periods of the pandemic (February–May, 2020). However, for better population dynamic inference, further analysis with more appropriate prior models for population dynamics would be required.

The Bayesian model estimates for the substitution rate are 7.28×10^{-4} subst/site/year (95% HPD: 6.32×10^{-4} - 8.25×10^{-4}) for B.1.1.248 and 6.16×10^{-4} subst/site/year (95% HPD: 5.61×10^{-4} - 6.76×10^{-4}) for B.1.1.33. While both intervals overlap with the overall estimate of the time-resolved ML tree built from 3758 representative genomes (6.59×10^{-4} subst/site/year), the B.1.1.248 lineage seems to have higher mutation rates. However, the phylogeographic model estimates similar overall migration rates for both lineages, 0.1710 (95% HPD: 0.0651–0.3523) for B.1.1.33 and 0.1980 (95% HPD: 0.0812–0.5002) for B.1.1.248.

Time-measured phylogeographic analysis highlighted the major contribution of Southeast in Brazilian and worldwide diffusion of both lineages (Figs. 4 and 5). Southeast is a common source of B.1.1.248 migrations, since we identified transition events between this region and Northern, Northeast, and Southern Brazil, as well as Asia, Europe, North America and Oceania (Bayes Factor (BF) > 30; Posterior Probability (PP) > 0.8) (Fig. 4a and b). The four subclades from Southern Brazil in the B.1.1.248 Maximum-Clade Credibility (MCC) tree were probably introduced from Southeast (Fig. 4a and b), and we were able to confirm that at least three independent introductions occurred in the municipality of Esteio as suggested previously by the ML analysis (Fig. 4a). Most importantly, the introduction of the P.2 lineage that harbors the E484K mutation was dated on September 09, 2020 (95% HPD: September 09–October 05, 2020) probably introduced from the Rio de Janeiro state. Interestingly, sequences from the USA and England formed a monophyletic clade with our sequence, demonstrating the spread from Brazil to other countries (Fig. 4a).

Southeast also seems to be determinant for the viral diffusion of B.1.1.33 lineage. Of note, the tree reconstruction showed important migrations from Southeast to Northern, Northeast, and Southern Brazil, as well as Europe, North and South America (BF > 30; PP > 0.8) (Fig. 5a and b). Well-supported rates were also identified between Northeast and Africa (BF = 31.10; PP = 0.79) and between South America and Oceania (BF = 56.99; PP = 0.87). Viruses belonging to this lineage appear to have a major contribution in Southern Brazil epidemics, since its sequences formed a monophyletic clade with > 50 sequences in the B.1.1.33 MCC tree (Fig. 5a). Furthermore, this analysis confirmed that Southern Brazil is a probable source of importations of available B.1.1.33 sequences to South-American countries (BF = 1467.97; PP = 0.99) and Northern Brazil (BF = 9.45; PP = 0.53) (Fig. 5a and b). Of note, we also validated that at least 6 introductions should have happened in the municipality of Esteio, two from the Southeast and four from other municipalities of Southern Brazil (especially from the RS state).

Discussion

In the present work, we accessed SARS-CoV-2 mutations, circulating lineages and phylogenetic patterns of SARS-CoV-2 from a time- and age-representative set of patients admitted in a municipal healthcare system from the Southern region of Brazil, the third country most affected by the Covid-19 pandemic. As the study was conducted in a small municipality, we were able to track two clusters of viral mutations in epidemiologically-linked patients, highlighting the importance of viral dissemination in small areas of the community.

The SARS-CoV-2 spike (S) glycoprotein mediates the interaction with the ACE2 receptor in the host cells and it is the primary target of neutralizing antibodies [45]. There are structural unique spike features that contribute to its pandemic capacity: (i) a flat sialic acid-binding domain enables faster viral surfing over the epithelial

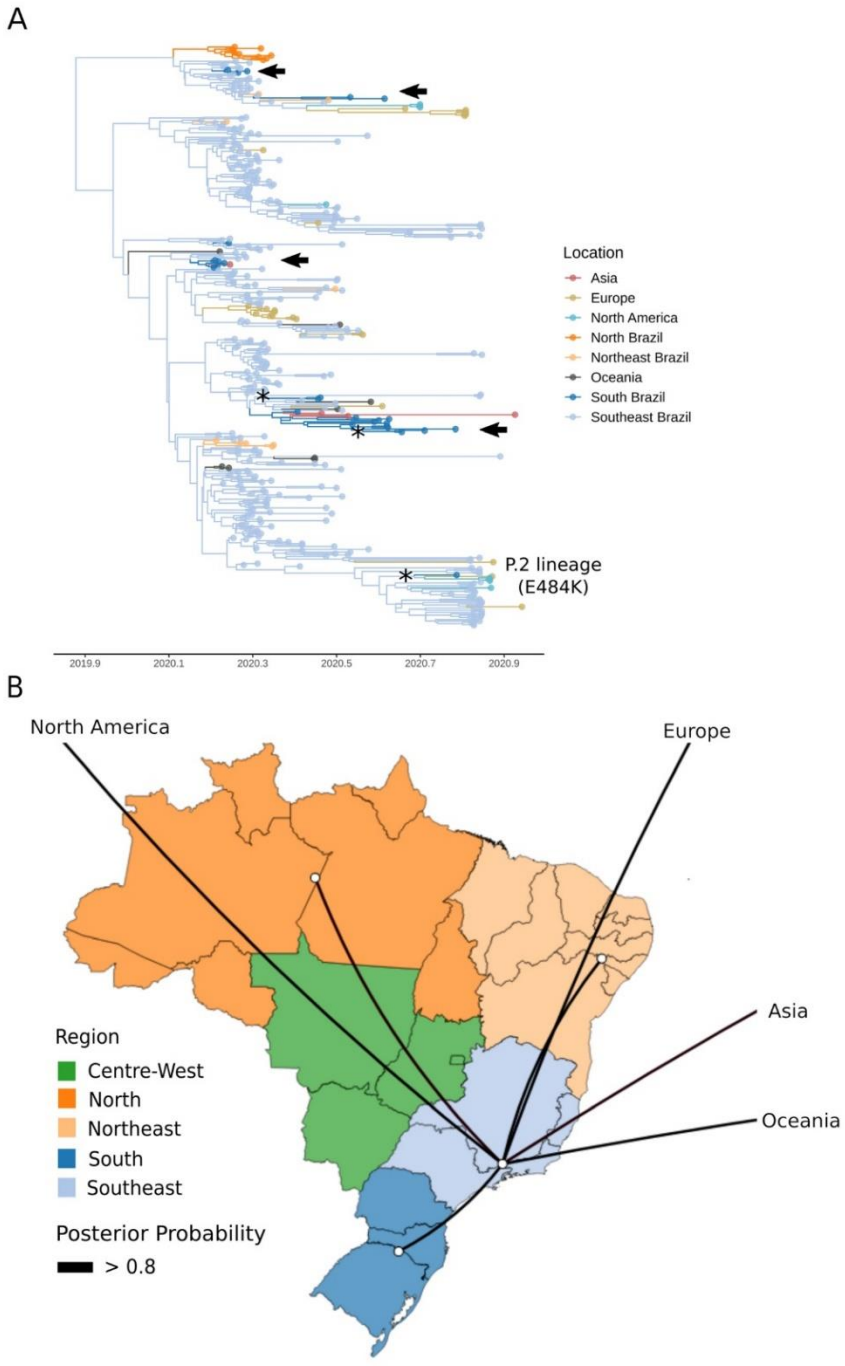

Fig. 4 Spatiotemporal diffusion of B.1.1.248 lineage across Brazilian regions and other continents. **a** Time-resolved MCC tree of 405 genomes belonging to B.1.1.248 lineage. **b** Well-supported transition rates between Brazilian regions and other continents in discrete phylogeographic reconstructions using BSSVS procedure and Bayes Factor tests. Transition rates with posterior probabilities < 0.5 were cut-off. The following states belong to each Brazilian region: Centre-West: DF, GO, MS, MT. North: AC, AM, AP, PA, RO, RR, TO. Northeast: AL, BA, CE, MA, PB, PE, PI, RN, SE. South: PR, RS, SC. Southeast: ES, MG, RJ, SP. The RS state is located in the Southern region. Subclades represented by > 1 Southern Brazilian sequence are indicated by arrows and introductions in Esteio are represented by asterisks. The introduction of P.2 lineage (E484K mutation) is indicated

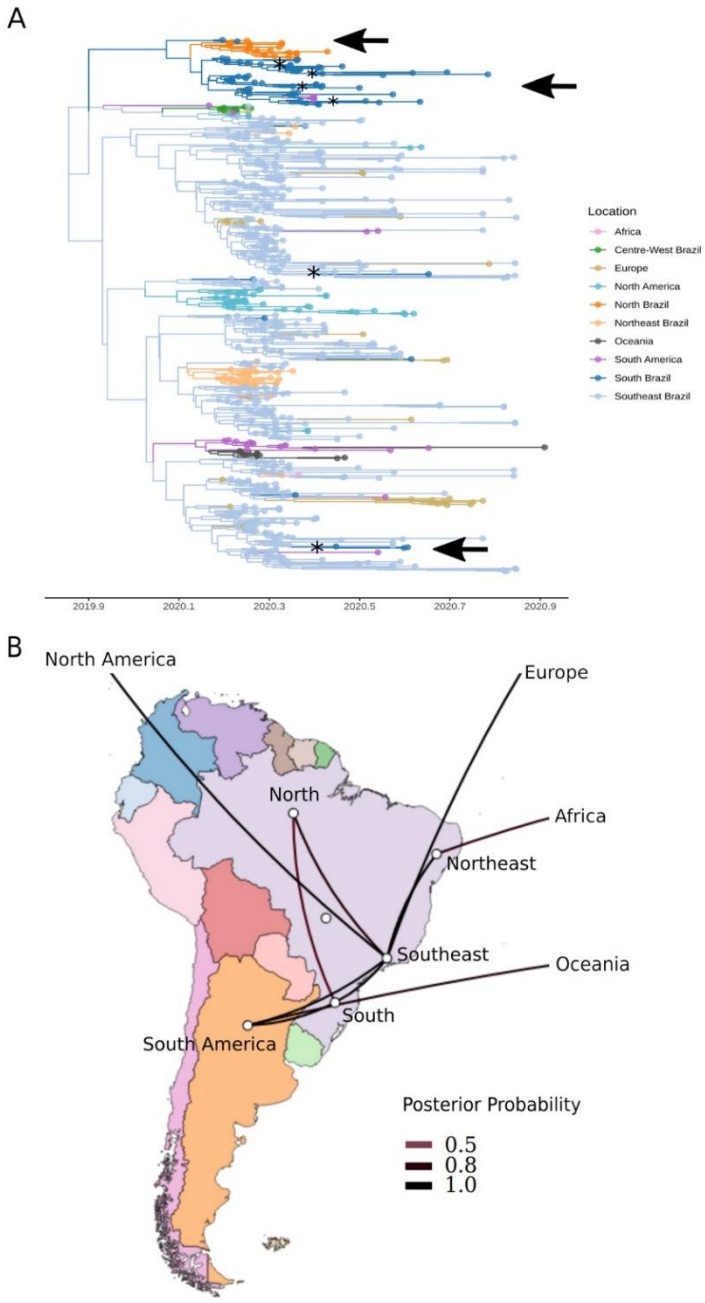

Fig. 5 Spatiotemporal diffusion of B.1.1.33 lineage across Brazilian regions, South America and other continents. **a** Time-resolved MCC tree of 725 genomes belonging to B.1.1.33 lineage. **b** Well-supported transition rates between Brazilian regions and other continents in discrete phylogeographic reconstructions using BSSVS procedure and Bayes Factor tests. Transition rates with posterior probabilities < 0.5 were cut-off. The subclades represented by > 1 Southern Brazilian sequence are indicated by arrows and introductions in Esteio are represented by asterisks

surface before receptor interaction; (ii) tight and nearly perfect binding to the ACE2 entry receptor; (iii) the capacity to use furin and other proteases for cell entry [46]. A mutation in S (D614G) was recently associated with higher viral loads [19], increased replication on human lung epithelial cells [23], and younger age of patients [22], but not with the disease severity [19, 22]. This mutation is associated with an abolition in the hydrogen

bond between the aspartate originally located at 614 position and a threonine residue in the 859 neighbouring protomer of the spike trimer, thus increasing the probability of RBD being found in the open state [47]. This promotes the binding with the ACE2 receptor leading to enhanced infectivity [23]. Importantly, this replacement was detected in 90.5% of our samples, compatible with the dominance of this variant in global context until December 14, 2020 (221,700 sequences [48];). Also concerning the S protein, we identified the V1176F variant in 38.1% of our samples and in all of them there was a co-occurrence with D614G. Importantly, 349/535 (65.2%) of the sequences isolated in the world including this replacement were from Brazil [48]. More recently (April 28, 2021), 1167 (53.7%) of worldwide sequences were from Brazil, representing mostly the spread of B.1.1.28 and P.1 lineages that harbor this mutation. This substitution is located on the C-terminal portion (S2 subunit), more specifically on the heptad repeat 2 (HP2), which is a major target of adaptive evolution in MERS-CoV-related viruses and carry sites associated with expanded host range in other coronaviruses [49].

Another replacement in the RDB of S protein (E484K) was also assigned. Until mid-December 2020, 157 genomes have this mutation globally, 114 (72.6%) isolated from South Africa, where a new lineage (B.1.351 or 501Y.V2) characterized by three RBD mutations (K417N, E484K and N501Y) recently emerged [50]. E484K also emerged independently in multiple lineages, including P.1, P.2 and B.1.1.33 firstly identified in Brazil between late 2020 and early 2021 [51]. As of 28 April, 2021, 38,436 genome sequences harbor this mutation, which are found in ~10% of the sequences generated to date on average (https://outbreak.info/situation-reports?muts=S%3AE484K) In Brazil, > 90% of the sequenced samples since February 2021 carry this mutation, which is present in P.1 and P.2 emergent lineages (https://outbreak.info/situation-reports?muts=S%3AE484K&selected=BRA&loc=BRA). Recent evidence showed that E484K replacement enables viral escape from neutralizing monoclonal antibodies or polyclonal sera [25–27] facilitating reinfection by emerging lineages harboring this mutation as reported in Brazil [52]. Importantly, two sequences (9.5%) from this study had four of five shared mutations with a new lineage (subsequently called P.2) reported in the Rio de Janeiro state (synonymous C28253T (F120F; ORF8), missense G28628T (A119S) and G28975T (M234I) in N protein and C29754U (3′ UTR)), in addition to E484K RBD replacement [53]. Our sequences from Esteio were sampled in mid-October, as the first cases of the RJ lineage (Additional File 2). Phylodynamics inferences pointed that this lineage emerged in early July, approximately four months before the detection of its first genomes [53]. Moreover, its identification in a small municipality from the RS state (located 1.5 thousand kilometers from RJ) demonstrates that it emerged months before October and is already widely distributed in the Brazilian territory, but went unnoticed so far by the lack of appropriate genomic surveillance in Brazil.

The substitution of a negatively charged amino acid (Glutamic Acid) for the positively charged Lysine has a profound impact upon a highly flexible loop at the RBD. More specifically, it creates a strong ion interaction between lysine and amino acid 75 of hACE-2 (the main SARS-CoV-2 receptor). This link is not present in the wild type E484. Shifting key elements of RBD responsible for interactions may have a major impact in the risk of immune evasion. Also E484K could be related to enhanced infectivity, which may be associated with the rapid dissemination of these escape mutants [54]. The publication of a recent reinfection case of SARS-CoV-2 harboring E484K and the presence of this mutation in COVID-19 patients during the current second wave in Northern Brazil [55] are highly suggestive that this mutation is critical for viral evolution and thus must be investigated thoroughly [52].

The ORF6 accessory protein plays a critical role in antagonizing host antiviral responses and viral replication. Therefore, it potentially inhibits both type I interferon (IFN) production and downstream signaling [56, 57]. We have found the ORF6:I33T mutation in 42.9% of our samples, raising its potential association with immune suppression. The N protein packages the genomic RNA, playing a fundamental role during viral self-assembly [58]. It is also associated with replication-transcription complexes [59], and used as a target for diagnostic and immunogenic applications. Interestingly, a tri-nucleotide mutation in the Nucleocapsid gene (RG203KR) resulting in a double amino acid change was observed in high frequency in this study (~ 100), and are characteristic of B.1.1 and derived lineages that spread rapidly around the world [48]. The missense mutation in ORF1ab (L3930F) was reported in other 429 sequences (125 from Brazil) and also in two sequences of lineage B.1.1.248 from the Philippines [60]. Interestingly, the I33F and I292T mutations that were found in ORF6 and N, respectively, have been considered dominant mutations in the SARS-CoV-2 sequences from Brazil [61, 62]. The co-occurrence of these mutations represents the signature of Clade 2 (subsequently named B.1.1.33 lineage), which were one of the three most prevalent Brazilian viral lineage groups in the beginning of the pandemic, highly widespread in 16 states from Brazil [12]. It is important to emphasize that functional studies are necessary to characterize the effect of each viral mutation on transmissibility and pathogenicity. It is expected that most mutations will not have a great impact on viral evolution and its relationship with the human host. However, some of them may increase the viral fitness or represent some viral advantage in the pathogen-host interaction and, consequently, may become

fixed in the population. Further studies to elucidate the interaction between human gene variants (especially in the ACE2 receptor) and SARS-CoV-2 mutations are necessary to establish the possible impacts of spike amino acid replacements with regard to ACE2 binding and function [63].

The knowledge of ACE2 physiological functions and specific features could explain how comorbidities like hypertension, diabetes, obesity, and immunological diseases can enhance the severity of symptoms. Thus, modulation of ACE2's function might promote pulmonary inflammation, thrombosis, obesity-induced hypertension, and cardiac failure, which are especially unfavorable to COVID-19 patients [36]. The expression of other cell receptors potentially involved with COVID-19 infection — depending on age, gender, and characteristics such as obesity, smoking, and polymorphisms — can contribute to patterns of severe symptoms [64]. Moreover, the hypertension-associated elevated immunological activity is a noteworthy factor that promotes the increased risk of hypertensive patients for critical COVID-19 outcomes. The increased immune activation can be observed in these patients and might elucidate the hyperinflammatory response (cytokine storm) [37].

The Brazilian population was mainly formed through an admixture process that comprises mostly European, sub-Saharan African, and Native-American ancestry [65–67]. The broad spectrum of symptoms, days to symptoms onset, and the unpredictability of outcome in COVID-19 patients can also be linked with the vastly admixed Brazilian inhabitants. Moreover, recent reports identified HLA alleles previously associated with SARS-CoV-2 counteraction [38], and showed positive selection in genes associated with obesity, type II diabetes, lipid levels, and waist circumference [44]. Regarding COVID-19, epigenetics, specific variants, ACE2 and TMPRSS2 polymorphisms, ethnicity, as well as inborn immunity errors, have been reported worldwide. This suggests that different host genetic backgrounds might contribute to discrepancies in SARS-CoV-2 aggressiveness [32–35].

We observed a higher CFR in the municipality of Esteio when compared to the RS state and the majority of Brazilian states, which might be linked to the emergence of new viral variants. However, COVID-19-related mortality is determined by both intrinsic factors of the infected individuals (age, comorbidities, and genetic characteristics) [68] and extrinsic aspects such as the access to healthcare assistance (hospital beds, mechanical ventilators, medicines). Additionally, the Southern Brazilian states (Rio Grande do Sul, Santa Catarina, and Paraná) have important determinants of mortality: older population than other regions and highest historical incidence of SARS in the country [69]. Their proximity to the states of São Paulo and Rio de Janeiro (that account for ~ 30% of the national population) also facilitates travel between the two regions and the rapid dissemination of emerging lineages.

Importantly, CFR is highly influenced by the underreporting of confirmed cases and deaths. States with low testing capacity tend to generate higher CFRs, and recently many deaths with an undetermined cause have been reported in Brazil, which also affects the quality of the records [70]. Therefore, the analysis of lethality should take this combination of factors into account [71]. In the case of Esteio, it is the municipality that tested more proportionally to the number of cases (38,416 tests per 100,000 inhabitants), thus these underreporting biases should be less significant, in contrast to the observed in other Brazilian states.

Compared to three other studies conducted in Brazil [12], in the states of Minas Gerais [14] and Pernambuco [13] in the early phase of pandemic — which showed the introduction of viral lineages from other continents (mainly Europe) by international returning travelers —, this study suggests a minor role of international lineages in the ongoing viral transmission in Esteio. We speculate a trend towards the perpetuation and diversification of the lineages found in this study (B.1.1.248 and B.1.1.33) inside Brazil. The dissemination of these lineages were also reported in the Uruguayan-Brazilian border, driving viral introductions mainly from Southeast and Southern Brazil (especially RS state) to Uruguay [72]. In this study, we found consistent results, mainly regarding B.1.1.33 diffusion from Southern and Southeast to South-American countries (*e. g. Argentina*, Chile and Uruguay). These lineages have already formed new sublineages (https://cov-lineages.org/lineages.html). B.1.1.33 has evolved in 10 new sublineages (N.1 to N.10). Furthermore, B.1.1.248 has evolved in P.1 (lineage first identified in Manaus [55] associated with a constellation of spike mutations like B.1.1.7 [73] and B.1.351 [50]), P.2 (lineage firstly identified in RJ state and also found in this study), and P.3. Importantly, all these three sublineages harbor the E484K mutation, which arose independently in both of them and appear to be evolving under diversifying positive selection [50, 51].

We built a time-resolved phylogeny prioritizing sequences that are genetically and spatially closer, but maintaining a global representativity of viral spread. This allowed us to confidently identify that our sequences fell into two main clades, with a broad presence of Brazilian and local sequences. We also inferred the spatiotemporal diffusion of these main lineages in regional and global context, finding the key role of Southeast in disseminating these lineages across Brazilian states and other continents. We also found a broader clade represented by Southern Brazil sequences and its important contribution in disseminating B.1.1.33 to South American

countries. Moreover, we found four broader clades in the B.1.1.248 MCC phylogeny, suggesting multiple introduction events from Southeast followed by community transmission.

Our evolutionary rate estimates for both the broader ML tree and lineage-specific MCC trees (6 to 7×10^{-4} subst/site/year) were slightly smaller than previous findings (8 to 9×10^{-4} subst/site/year) [74, 75]. These differences have contributed to the estimates of the T_{MRCA} for both B.1.1.248 and B.1.1.33, which were dated to late 2019, contrasting with the first description of these lineages [12]. Furthermore, other probable sources of these inconsistencies are: closely related samples having the same age (phylo-temporal clustering), among-lineage rate variation and non-random sampling [76]. Although it is possible that different branches of the phylogeny have different rates, when we used a model that allows different rates across the tree (uncorrelated lognormal relaxed clock), the T_{MRCA} estimates have remained unchanged.

An important caveat for the phylodynamic analysis is that samples are not equally distributed geographically or temporally. This is a consequence of episodic sampling efforts prompted by research resource availability, and does not necessarily resemble a representative uniform sample. Unequal temporal distribution implies that some of the conclusions are disproportionately influenced by events in heavily sampled periods (February–May). Additionally, a large proportion of the samples come from the Southeastern region of Brazil. While this is in fact a heavily hit region and an economic and travel hub for the country, other regions such as the North are underrepresented. Thus, the strong support for the Southeast as prime center for viral dispersion and location of the root of both clades might be somewhat inflated, and epidemiological links between other regions could be downplayed due to undersampling. However, a study from the beginning of pandemic (February–March 2020) estimated that the main destinations of the international passengers arriving to Brazil were São Paulo (46.1%), Rio de Janeiro (21%) and Belo Horizonte (4.1%), three capitals from the Southeast and therefore routes for COVID-19 importation [77]. Moreover, during mid-February and mid-March, SARS-CoV-2 spread mostly locally and within-state borders. In contrast, during mid-March and mid-April there was an ignition of the epidemic from the Southeast region to other states [12], which is consistent with our findings.

Some limitations should be considered. Firstly, it was not possible to analyze a larger sample size. Moreover, the low quantity and spatial representativity of sequences from the RS state to contextualize our sequences limited the inference of events of introduction and movement of the virus with municipal and state resolution. Still in this respect, we have observed a dramatic drop in the sequencing efforts from Brazil after April 2020 [12], which made it difficult to measure the main circulating lineages in the country during our investigation period (May–October, 2020) and may introduce confounding factors.

Since the E484K mutation identified in this study has been associated with loss of neutralizing activity from convalescent plasma (immune evasion) and enhanced interaction with hACE-2, lineages containing this substitution must be the subject of intense surveillance. More specifically, it is critical that immune strategies such as convalescent plasma and vaccines be tested against these new variants. Attempts to demonstrate activity against S mutants should be a priority effort for all vaccine and monoclonal antibody makers. Second generation immune therapies might have to be directed at more conservative neutralizing binding sites (such as S2 fusion domain) or elicit strong cellular response in order to keep on long term protection. Finally, human genetic factors, patient heritage and health conditions should also be studied in an integrated way for a broader understanding of vaccine effectiveness in different populations.

Conclusions

Our results provide a comprehensive view of SARS-CoV-2 mutations from a time- and age-representative sample from May to October 2020, highlighting two frequent mutations in spike glycoprotein (D614G and V1176F), an emergent mutation in spike RBD (E484K) characteristic of B.1.351 and P.1 lineages, and the adjacent replacement of 2 amino acids in Nucleocapsid phosphoprotein (R203K and G204R). In particular, to our best knowledge, we described the earliest SARS-CoV-2 sequences harboring E484K in Southern Brazil. A significant viral diversity was evidenced by the absence of identical isolates in our samples. Furthermore, we identified patterns of SARS-CoV-2 viral diversity inside Southern Brazil, demonstrating the major role of community transmission in viral spreading and the establishment of Brazilian lineages ignited from the Southeast to other Brazilian regions. Our data show the value of correlating clinical, epidemiological and genomic information for the understanding of viral evolution and its spatial distribution over time. This is of paramount importance to better inform policy making strategies to fight COVID-19.

Methods

Sample collection and clinical testing

Nasopharyngeal samples were obtained from patients of the Hospital São Camilo, Secretaria Municipal de Esteio and Vigilância em Saúde from Esteio, RS, Brazil. Nasopharyngeal swabs were collected and placed in Viral

Transport Medium (VTM, Copan Universal Transport Medium). Samples were transported to the Molecular Microbiology Laboratory from Feevale University and tested on the same day for SARS-CoV-2 by reverse-transcriptase quantitative polymerase chain reaction (RTq-PCR). Remnant samples were stored at − 80 °C. SARS-CoV-2 diagnosis was performed using Real Time Reverse-transcriptase Polymerase Chain Reaction (Charité RT-qPCR assays). The RTq-PCR assay used primers and probes recommended by the World Health Organization (WHO) targeting the Nucleocapsid (N1 and N2) genes [78].

We selected 21 samples with RT-qPCR positive results, collected from May 31 to October 13, 2020 from patients residing in the municipality of Esteio, RS, Brazil. We included patients who presented symptoms such as fever, cough, sore throat, dyspnea, anosmia, fatigue, diarrhea and/or vomiting. The clinical status classification was based on the COVID-19 Clinical management guide recommended by the WHO [79]. Additionally, samples were selected based on cycle threshold (Ct) values ≤ 20. Electronic medical records were reviewed to compile epidemiological metadata (e. g., date of collection, sex, age, symptoms, exposure history, and clinical status).

RNA extraction, library preparation and sequencing

We submitted the RT-qPCR positive for SARS-CoV-2 swabs to genomic RNA extraction. This process was performed in the automated nucleic acid purification system KingFisher™ Duo Prime Purification System (ThermoFisher Scientific, Waltham, USA) along with the MagMax™ CORE Nucleic Acid Purification Kit (ThermoFisher Scientific, Waltham, USA).

The extracted and purified genomic RNA was transcribed to cDNA using Maxima H Minus Double-Stranded cDNA Synthesis Kit, catalog number K2561 (ThermoFisher Scientific, Waltham, USA) following the manufacturer's instructions. Library preparation was achieved using Nextera™ Flex for Enrichment with RNA Probes (Illumina, San Diego, USA). Briefly, we performed tagmentation in a pre-programmed thermocycler incubation temperature, until holding at 10 °C. This step uses the Enrichment Bead-Linked Transposomes (Enrichment BLT, eBLT) to tagment DNA followed by post tagmentation clean up. The PCR procedure adds pre-paired 10 base pair adapters and sequences required for sequencing cluster generation. The viral cDNA was used as input for multiple overlapping PCR reactions that spanned the viral genome (Enhanced PCR Mix reagent and nuclease-free water). The amplified tagmented DNA was cleaned with AMPure XP magnetic beads (Beckman Coulter Inc., Indianapolis, USA) to remove shorter DNA fragments and other impurities. We then quantified the cleaned libraries employing Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific, Waltham, USA).

Sequencing was performed on an Illumina Miseq® (Illumina, San Diego, USA) using Reagent Kit v3 with 150 cycles in a paired-end run, following the manufacturer's instructions. All experiments were performed in a biosafety level 2 laboratory.

Consensus calling

Reference mapping and consensus calling was performed using an in-house developed pipeline managed with Snakemake [80]. Briefly, quality control was performed FastQC v0.11.9 and low-quality reads and adapters were removed using Trimmomatic v0.39 [81]. PCR duplicates were discarded using Picard MarkDuplicates v2.23.8 (https://broadinstitute.github.io/picard/). Reads were mapped to the reference SARS-CoV-2 genome (GenBank accession number NC_045512.2) using burrows–wheeler aligner (BWA-MEM) v0.7.17 [82] and unmapped reads were discarded. Consensus sequences were generated using bcftools mpileup combined with bcftools consensus v1.9 [83]. Positions covered by fewer than 10 reads (DP < 10) were considered a gap in coverage and converted to ns. Coverage values for each genome were calculated using bedtools v2.26.0 [84] and plotted using the karyoploteR v1.12.4 package [85]. Finally, we assessed genome consensus sequence quality using Nextclade v0.8.1 (https://clades.nextstrain.org/) and CoV-GLUE (http://cov-glue.cvr.gla.ac.uk/ [48];)

Virome analysis

As the respiratory panel kit used allows the detection of ~ 40 respiratory viral pathogens, the viral composition of each sample (all mapped and unmapped reads against reference) was verified using Kaiju v1.7.3 [86] and Kraken v2.0.7-beta [87] against a reference database of viral sequences. The viral database for each tool was built with the following commands, respectively: kaiju-makedb -s visuses and kraken2-build --download-library viral. Taxonomic classification interactive charts were visualized using Krona [88].

Mutation analysis

Sequence positions in this work refer to GenBank RefSeq sequence NC_045512.2, a genome isolated and sequenced from Wuhan (China), early in the pandemic. Single Nucleotide Polymorphisms (SNPs) and insertions/deletions (INDELs) were assessed in each sample by using snippy variant calling and core genome alignment pipeline v4.6.0 (https://github.com/tseemann/snippy), which uses FreeBayes v1.3.2 [89] variant caller and snpEff v5.0 [90] to annotate and predict the effects of variants on genes and proteins. Genome map and histogram of SNPs were generated after running MAFFT v7.471 alignment using a modified code from Lu et al. 2020 (https://github.com/laduplessis/SARS-CoV-2_

Guangdong_genomic_epidemiology/). Moreover, we identified global virus lineages using Nextclade v0.8.1 (https://clades.nextstrain.org/) and Pangolin v2.1.3 (https://github.com/cov-lineages/pangolin [15];).

Phylogenetics analysis

All available SARS-CoV-2 genomes (285,411 sequences) were obtained from GISAID on December 24, 2020. Available sequences were then subjected to analysis inside the NextStrain ncov pipeline (https://github.com/nextstrain/ncov [91];). Briefly, this pipeline uses the augur toolkit to (i) exclude short and low quality sequences or those with incomplete sampling date; (ii) align filtered sequences using MAFFT v7.471 [92]; (iii) mask uninformative sites and ends from the alignment; (iv) perform context subsampling using genetically closely-related genomes to our focal subset prioritizing sequences geographically closer to RS state, Brazil; (v) build maximum likelihood (ML) phylogenetics tree using IQ-TREE v2.0.3 [93], employing the General time reversible model (GTR) with unequal rates and base frequencies [94], (vi) generate a time-scaled tree resolving polytomies and internal nodes with TreeTime v0.7.6, and under a strict clock under a skyline coalescent prior with a rate of 8×10^{-4} substitutions per site per year [95]; (vii) label clades, assign mutations and infer geographic movements; and (viii) export results to JSON format to enable interactive visualization through Auspice. The ML tree was inspected in TempEst v1.5.3 [96] to investigate the temporal signal through regression of root-to-tip genetic divergence against sampling dates.

Phylodynamic and phylogeographic analysis

All global sequences (until December 24, 2020) belonging to lineages B.1.1.248 ($n = 405$) and B.1.1.33 ($n = 725$), found in high frequency in this study, were recovered from the filtered MAFFT alignment performed inside Nextstrain ncov pipeline in the previous step. The T_{MRCA} and the spatial diffusion of these important circulating lineages through Brazil were separately estimated for each lineage using a Bayesian Markov Chain Monte Carlo (MCMC) approach as implemented in BEAST v1.10.4 [97], using the BEAGLE library v3 [98] to save computational time. Time-scaled Bayesian trees were estimated in BEAST using: $HKY + \Gamma$ nucleotide substitution model, a strict molecular clock model with a Continuous Time Markov Chain (CTMC) prior (mean rate $= 8 \times 10^{-4}$) for the clock rate [99], and a parametric exponential growth model.

Two MCMC chains were run for at least 120 million generations and convergence of the MCMC chains was inspected using Tracer v1.7.1 [100]. After removal of 10% burn-in, log and tree files were combined using LogCombiner v1.10.4 [97] to ensure stationarity and good mixing. Maximum clade credibility (MCC) summary trees were generated using TreeAnnotator v1.10.4 [97]. MCC trees were visualized using FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/) and additional annotations were performed in ggtree R package v2.0.4 [101].

Viral migrations across time were reconstructed using a reversible discrete asymmetric phylogeographic model [102] in order to estimate locations of each internal node of the phylogeny. SpreaD3 [103] was used to map spatiotemporal information embedded in MCC trees. A discretization scheme of 10 possible states defined as Brazilian regions (Centre-West, North, Northeast, South, and Southeast) or other continents (Africa, Europe, North America, Oceania, and South America) was applied. For map plotting, latitudes and longitudes were attributed to a randomly selected point next to the center of each region or continent. Location exchange rates that dominate the diffusion process were identified using the Bayesian stochastic search variable selection (BSSVS) procedure [102] using Bayes Factor tests to identify well-supported rates.

Abbreviations

ACE2: Angiotensin-converting enzyme 2; BF: Bayes Factor; BSSVS: Bayesian Stochastic Search Variable Selection; CFR: Case fatality rate; CTMC: Continuous Time Markov Chain; CONEP: Brazilian's National Ethics Committee (Comissão Nacional de Ética em Pesquisa); DP: Depth of coverage; GISAID: Global Initiative on Sharing Avian Influenza Data; GTR: General time reversible model; HPD: Highest Posterior Density; INDEL: Insertion/deletion; MCC: Maximum clade credibility; MCMC: Markov chain Monte Carlo; ML: Maximum Likelihood; ORF: Open Reading Frame; PP: Posterior Probability; R_0: Reproduction number; RBD: Receptor Binding Domain; RS: Rio Grande do Sul state, Southern Brazil; RTq-PCR: Real Time Reverse-transcriptase Polymerase Chain Reaction; SARS-CoV-2: *Severe acute respiratory syndrome coronavirus 2*; SNP: Single Nucleotide Polymorphism; TMPRSS2: transmembrane protease serine 2; VTM: Virus Transport Medium; WHO: World Health Organization

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07708-w.

Additional file 1.
Additional file 2.
Additional file 3.
Additional file 4.

Acknowledgements

We thank the administrators and curators of the GISAID database and research groups across the world for supporting rapid and transparent sharing of genomic data during the COVID-19 pandemic. We also thank the Mayor's Office, Health Department and São Camilo Hospital (Esteio, RS, Brazil), Leonardo Duarte Pascoal and Ana Regina Boll for their work in combating COVID-19, for providing the samples and epidemiological information of patients, and for the financial support for the sequencing; the Molecular Microbiology Laboratory of Feevale University for conducting the sequencing; Arnaldo Zaha, Augusto Schrank, Macley Silva Cardoso, Maiara Monteiro Oliveira, and Marilene Henning Vainstein from Universidade Federal do Rio Grande do Sul for writing corrections and insightful discussions on the best way to present the data of this manuscript.

Authors' contributions

CET and LK conceived the study. AMM, CAMN, GDC collected clinical samples and metadata. AMM, AWH, FRS, GDC, JDF, JSG, MD, MNW, PRA generated sequencing data. CET, GBC, LK, VBF contributed bioinformatics tools. VBF generated the visualizations. CET, GBC, LK, PAGF, VBF performed the data analysis. CET, GBC, LK, RAZ, VBF wrote the manuscript. The authors read and approved the final manuscript.

Funding

This work was supported by grants from Prefeitura Municipal de Esteio (Esteio Mayor's Office). Students' scholarships were supplied by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The funders had no role in the study design, data generation and analysis, decision to publish or the preparation of the manuscript.

Availability of data and materials

The FASTQ data generated in this study have been submitted to the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra) under BioProject accession number PRJNA707583. Consensus genomes have been submitted to the GISAID database (https://www.gisaid.org/) under accession numbers listed in Table 1 (EPI_ISL_831678, EPI_ISL_831474, EPI_ISL_831645, EPI_ISL_831646, EPI_ISL_831660, EPI_ISL_831681, EPI_ISL_831683, EPI_ISL_831685, EPI_ISL_831688, EPI_ISL_831689, EPI_ISL_831892, EPI_ISL_831898, EPI_ISL_831913, EPI_ISL_831938, EPI_ISL_831939, EPI_ISL_831940, EPI_ISL_832009, EPI_ISL_832010, EPI_ISL_832011, EPI_ISL_832012, EPI_ISL_832013). The publicly available sequences used are listed in the Additional File 3. An interactive Microreact visualization is available at https://microreact.org/project/9ia4CJbx1mF52Ew1g7e9uT. Additional information used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Ethical approval was obtained from the Brazilian's National Ethics Committee (Comissão Nacional de Ética em Pesquisa — CONEP) under process number 30934020.5.0000.0008. The study was performed in accordance with the Declaration of Helsinki. Patients were informed in detail about the study and written informed consent was obtained from all participants. Their samples were anonymized before received by the study investigators, following Brazilian and international ethical standards.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Center of Biotechnology, Graduate Program in Cell and Molecular Biology (PPGBCM), Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil. [2]Graduate Program in Health Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, RS, Brazil. [3]Department of Statistics, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil. [4]Molecular Microbiology Laboratory, Universidade Feevale, Novo Hamburgo, RS, Brazil. [5]Irmandade Santa Casa de Misericórdia de Porto Alegre, Porto Alegre, RS, Brazil. [6]Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), 245/200C Sarmento Leite St, Porto Alegre, RS 90050-170, Brazil.

Received: 23 February 2021 Accepted: 11 May 2021
Published online: 20 May 2021

References

1. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5.
2. Johns Hopkins Coronavirus Resource Center. COVID-19 Map. Johns Hopkins Coronavirus Resource Center. 2021. https://coronavirus.jhu.edu/map.html. Accessed 10 Nov 2020.
3. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. 2017;22(13). https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.
4. Deng X, Gu W, Federman S, du Plessis L, Pybus OG, Faria N, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into northern California. Science. 2020;369(6503):582–7. https://doi.org/10.1126/science.abb9263.
5. Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, et al. Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. Cell. 2020;181:990–6 e5.
6. Ladner JT, Larsen BB, Bowers JR, Hepp CM, Bolyen E, Folkerts M, et al. An Early Pandemic Analysis of SARS-CoV-2 Population Structure and Dynamics in Arizona. mBio. 2020;11. https://doi.org/10.1128/mBio.02107-20.
7. Maurano MT, Ramaswami S, Zappile P, Dimartino D, Boytard L, Ribeiro-dos-Santos AM, et al. Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City Region. Genome Res. 2020;30:1781–88. https://doi.org/10.1101/gr.266676.120.
8. Bartolini B, Rueca M, Gruber CEM, Messina F, Carletti F, Giombini E, et al. SARS-CoV-2 phylogenetic analysis, Lazio region, Italy, February–march 2020. Emerg Infect Dis. 2020;26(8):1842–5. https://doi.org/10.3201/eid2608.201525.
9. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med. 2020;26(9):1405–10. https://doi.org/10.1038/s41591-020-0997-y.
10. Rockett RJ, Arnott A, Lam C, Sadsad R, Timms V, Gray K-A, et al. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. Nat Med. 2020;26:1398–1404. https://doi.org/10.1038/s41591-020-1000-7.
11. Seemann T, Lane CR, Sherry NL, Duchene S, Gonçalves da Silva A, Caly L, et al. Tracking the COVID-19 pandemic in Australia using genomics. Nat Commun. 2020;11:4376.
12. Candido D, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science. 2020;369(6508):1255–60. https://doi.org/10.1126/science.abd2161.
13. Paiva MHS, Guedes DRD, Docena C, Bezerra MF, Dezordi FZ, Machado LC, et al. Multiple introductions followed by ongoing community spread of SARS-CoV-2 at one of the largest metropolitan areas of Northeast Brazil. Viruses. 2020;12(12):1414. https://doi.org/10.3390/v12121414.
14. Xavier J, Giovanetti M, Adelino T, Fonseca V, da Costa AVB, Ribeiro AA, et al. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. Emerg Microbes Infect. 2020;9(1):1824–34. https://doi.org/10.1080/22221751.2020.1803146.
15. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol. 2020;5(11):1403–7. https://doi.org/10.1038/s41564-020-0770-5.
16. Takahiko Koyama, Daniel Platt, Laxmi Parida. WHO | Variant analysis of SARS-CoV-2 genomes. https://www.who.int/bulletin/volumes/98/7/20-253591/en/. Accessed 24 Nov 2020.
17. Laamarti M, Alouane T, Kartti S, Chemao-Elfihri MW, Hakmi M, Essabbar A, 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. 2020;15(11):e0240345. https://doi.org/10.1371/journal.pone.0240345.
18. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol. 2020;83:104351. https://doi.org/10.1016/j.meegid.2020.104351.
19. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell. 2020;182(4):812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043.
20. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. Cell. 2020;182:1284–94 e9.
21. Toyoshima Y, Nemoto K, Matsumoto S, Nakamura Y, Kiyotani K. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J Hum Genet. 2020;65(12):1075–82. https://doi.org/10.1038/s10038-020-0808-9.

22. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell. 2020;184(1):64–75.e11. https://doi.org/10.1016/j.cell.2020.11.020.
23. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature. 2021;592:116–21. https://doi.org/10.1038/s41586-020-2895-3.
24. Gu H, Chen Q, Yang G, He L, Fan H, Deng Y-Q, et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. Science. 2020;369(6511):1603–7. https://doi.org/10.1126/science.abc4730.
25. Baum A, Fulton BO, Wloga E, Copin R, Pascal KE, Russo V, et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science. 2020;369(6506):1014–8. https://doi.org/10.1126/science.abd0831.
26. Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe. 2020;29(1):44–57.e9. https://doi.org/10.1016/j.chom.2020.11.007.
27. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife. 2020;9:e61312. https://doi.org/10.7554/eLife.61312.
28. Calcagno A, Ghisetti V, Burdino E, Trunfio M, Allice T, Boglione L, et al. Coinfection with other respiratory pathogens in COVID-19 patients. Clin Microbiol Infect. 2020;0. https://doi.org/10.1016/j.cmi.2020.08.012.
29. Kim D, Quinn J, Pinsky B, Shah NH, Brown I. Rates of co-infection between SARS-CoV-2 and other respiratory pathogens. JAMA. 2020;323(20):2085–6. https://doi.org/10.1001/jama.2020.6266.
30. Nowak MD, Sordillo EM, Gitman MR, Mondolfi AEP. Coinfection in SARS-CoV-2 infected patients: where are influenza virus and rhinovirus/enterovirus? J Med Virol. 2020;92(10):1699–700. https://doi.org/10.1002/jmv.25953.
31. Peddu V, Shean RC, Xie H, Shrestha L, Perchetti GA, Minot SS, et al. Metagenomic analysis reveals clinical SARS-CoV-2 infection and bacterial or viral superinfection and colonization. Clin Chem. 2020;66(7):966–72. https://doi.org/10.1093/clinchem/hvaa106.
32. Zhang Q, Bastard P, Liu Z, Le Pen J, Moncada-Velez M, Chen J, et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science. 2020;370(6515):eabd4570. https://doi.org/10.1126/science.abd4570.
33. Hou Y, Zhao J, Martin W, Kallianpur A, Chung MK, Jehi L, et al. New insights into genetic susceptibility of COVID-19: an ACE2 and TMPRSS2 polymorphism analysis. BMC Med. 2020;18(1):216. https://doi.org/10.1186/s12916-020-01673-z.
34. Choudhary S, Sreenivasulu K, Mitra P, Misra S, Sharma P. Role of genetic variants and gene expression in the susceptibility and severity of COVID-19. Ann Lab Med. 2021;41(2):129–38. https://doi.org/10.3343/alm.2021.41.2.129.
35. De La Cruz M, Nunes DP, Bhardwaj V, Subramanyan D, Zaworski C, Roy P, et al. Colonic epithelial angiotensin-converting enzyme 2 (ACE2) expression in blacks and whites: potential implications for pathogenesis Covid-19 racial disparities. J Racial Ethn Health Disparities. 2021. https://doi.org/10.1007/s40615-021-01004-9.
36. Guilger-Casagrande M, de Barros CT, Antunes VAN, de Araujo DR, Lima R. Perspectives and challenges in the fight against COVID-19: the role of genetic variability. Front Cell Infect Microbiol. 2021;11. https://doi.org/10.3389/fcimb.2021.598875.
37. Trump S, Lukassen S, Anker MS, Chua RL, Liebig J, Thürmann L, et al. Hypertension delays viral clearance and exacerbates airway hyperinflammation in patients with COVID-19. Nat Biotechnol. 2020:1–12. https://doi.org/10.1038/s41587-020-00796-1.
38. Secolin R, de Araujo TK, Gonsales MC, Rocha CS, Naslavsky M, Marco LD, et al. Genetic variability in COVID-19-related genes in the Brazilian population. Hum Genome Var. 2021;8:1–9.
39. Rio Grande do Sul Department of Health. SES-RS - Coronavírus. https://ti.saude.rs.gov.br/covid19/. Accessed 24 Nov 2020.
40. Brazilian Institute of Geography and Statistics - IBGE. Cidades e Estados: Rio Grande do Sul. https://www.ibge.gov.br/cidades-e-estados/rs.html. Accessed 24 Nov 2020.
41. Esteio Department of Health. Monitoramento COVID-19 Esteio. http://covid.esteio.rs.gov.br/. Accessed 24 Nov 2020.
42. Petersen E, Koopmans M, Go U, Hamer DH, Petrosillo N, Castelli F, et al. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. Lancet Infect Dis. 2020;20(9):e238–44. https://doi.org/10.1016/S1473-3099(20)30484-9.
43. Yang W, Kandula S, Huynh M, Greene SK, Wye GV, Li W, et al. Estimating the infection-fatality risk of SARS-CoV-2 in New York City during the spring 2020 pandemic wave: a model-based analysis. Lancet Infect Dis. 2021;21(2):203–12. https://doi.org/10.1016/S1473-3099(20)30769-6.
44. Secolin R, Gonsales MC, Rocha CS, Naslavsky M, De Marco L, Bicalho MAC, et al. Exploring a Region on Chromosome 8p23.1 Displaying Positive Selection Signals in Brazilian Admixed Populations: Additional Insights Into Predisposition to Obesity and Related Disorders. Front Genet. 2021;12. https://doi.org/10.3389/fgene.2021.636542.
45. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181:281–92 e6.
46. Seyran M, Takayama K, Uversky VN, Lundstrom K, Palù G, Sherchan SP, et al. The structural basis of accelerated host cell entry by SARS-CoV-2. FEBS J. 2020. https://doi.org/10.1111/febs.15651.
47. Mansbach RA, Chakraborty S, Nguyen K, Montefiori DC, Korber B, Gnanakaran S. The SARS-CoV-2 Spike variant D614G favors an open conformational state. Sci Adv. 2021;7:eabf3671. https://doi.org/10.1126/sciadv.abf3671.
48. Singer J, Gifford R, Cotten M, Robertson D. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. 2020. https://doi.org/10.20944/preprints202006.0225.v1.
49. Forni D, Filippi G, Cagliani R, De Gioia L, Pozzoli U, Al-Daghri N, et al. The heptad repeat region is a major selection target in MERS-CoV and related coronaviruses. Sci Rep. 2015;5(1):14480. https://doi.org/10.1038/srep14480.
50. Tegally H, Wilkinson E, Giovanetti M, Iranzadeh A, Fonseca V, Giandhari J, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature. 2021;592:438–43. https://doi.org/10.1038/s41586-021-03402-9.
51. Ferrareze PAG, Franceschi VB, de Menezes Mayer A, Caldana GD, Zimerman RA, Thompson CE. E484K as an innovative phylogenetic event for viral evolution: Genomic analysis of the E484K spike mutation in SARS-CoV-2 lineages from Brazil. bioRxiv. 2021; 2021.01.27.426895. https://doi.org/10.1101/2021.01.27.426895.
52. Nonaka CKV, Franco MM, Gräf T, Barcia CA de L, Mendonça RN de Á, Sousa KAF de, et al. Genomic Evidence of SARS-CoV-2 Reinfection Involving E484K Spike Mutation, Brazil. Emerg Infect Dis. 2021;27:1522. https://doi.org/10.3201/eid2705.210191.
53. Voloch CM, Francisco R da S, Almeida LGP de, Cardoso CC, Brustolini OJ, Gerber AL, et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. J Virol. 2021;95. https://doi.org/10.1128/JVI.00119-21.
54. Nelson G, Buzko O, Spilman P, Niazi K, Rabizadeh S, Soon-Shiong P. Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. bioRxiv. 2021; 2021.01.13.426558. https://doi.org/10.1101/2021.01.13.426558.
55. Faria N, Claro IM, Candido D, Franco LAM, Andrade PS, Coletti TM, et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. Virological. 2021; https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586. Accessed 14 Jan 2021.
56. Lei X, Dong X, Ma R, Wang W, Xiao X, Tian Z, et al. Activation and evasion of type I interferon responses by SARS-CoV-2. Nat Commun. 2020;11(1):3810. https://doi.org/10.1038/s41467-020-17665-9.
57. Schultze JL, Aschenbrenner AC. COVID-19 and the human innate immune system. Cell. 2021;184(7):1671–92. https://doi.org/10.1016/j.cell.2021.02.029.
58. Chang C, Hou M-H, Chang C-F, Hsiao C-D, Huang T. The SARS coronavirus nucleocapsid protein--forms and functions. Antivir Res. 2014;103:39–50. https://doi.org/10.1016/j.antiviral.2013.12.009.
59. Verheije MH, Hagemeijer MC, Ulasli M, Reggiori F, Rottier PJM, Masters PS, et al. The coronavirus Nucleocapsid protein is dynamically associated with the replication-transcription complexes. J Virol. 2010;84(21):11575–9. https://doi.org/10.1128/JVI.00569-10.
60. Velasco JM, Chinnawirotpisan P, Joonlasak K, Manasatienkij W, Huang A, Valderama MT, et al. Coding-complete genome sequences of 23 SARS-CoV-2 samples from the Philippines. Microbiol Resour Announc. 2020;9(43). https://doi.org/10.1128/MRA.01031-20.
61. Franco-Muñoz C, Álvarez-Díaz DA, Laiton-Donato K, Wiesner M, Escandón P, Usme-Ciro JA, et al. Substitutions in spike and Nucleocapsid proteins of SARS-CoV-2 circulating in South America. Infect Genet Evol. 2020;85:104557. https://doi.org/10.1016/j.meegid.2020.104557.

62. Singh J, Singh H, Hasnain SE, Rahman SA. Mutational signatures in countries affected by SARS-CoV-2: Implications in host-pathogen interactome. bioRxiv. 2020; 2020.09.17.301614. https://doi.org/10.1101/2020.09.17.301614.
63. Villoutreix BO, Calvez V, Marcelin A-G, Khatib A-M. In Silico investigation of the new UK (B.1.1.7) and south African (501Y.V2) SARS-CoV-2 variants with a focus at the ACE2-spike RBD Interface. Int J Mol Sci. 2021;22(4). https://doi.org/10.3390/ijms22041695.
64. Radzikowska U, Ding M, Tan G, Zhakparov D, Peng Y, Wawrzyniak P, et al. Distribution of ACE2, CD147, CD26, and other SARS-CoV-2 associated molecules in tissues and immune cells in health and in asthma, COPD, obesity, hypertension, and COVID-19 risk factors. Allergy. 2020;75(11):2829–45. https://doi.org/10.1111/all.14429.
65. Kehdy FSG, Gouveia MH, Machado M, Magalhães WCS, Horimoto AR, Horta BL, et al. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proc Natl Acad Sci. 2015;112(28):8696–701. https://doi.org/10.1073/pnas.1504447112.
66. Lima-Costa MF, Rodrigues LC, Barreto ML, Gouveia M, Horta BL, Mambrini J, et al. Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (the Epigen initiative). Sci Rep. 2015;5(1):9812. https://doi.org/10.1038/srep09812.
67. de Moura RR, Coelho AVC, de Queiroz Balbino V, Crovella S, Brandão LAC. Meta-analysis of Brazilian genetic admixture and comparison with other Latin America countries. Am J Hum Biol. 2015;27(5):674–80. https://doi.org/10.1002/ajhb.22714.
68. Feng Y, Ling Y, Bai T, Xie Y, Huang J, Li J, et al. COVID-19 with different severities: a multicenter study of clinical features. Am J Respir Crit Care Med. 2020;201(11):1380–8. https://doi.org/10.1164/rccm.202002-0445OC.
69. Bastos LS, Niquini RP, Lana RM, Villela DAM, Cruz OG, Coelho FC, et al. COVID-19 e hospitalizações por SRAG no Brasil: uma comparação até a 12a semana epidemiológica de 2020. Cad Saúde Pública. 2020;36(4):e00070120. https://doi.org/10.1590/0102-311x00070120.
70. Alves THE, Souza TA de, Samyla de Almeida Silva, Ramos NA, SV de Oliveira. Underreporting of death by COVID-19 in Brazil's second Most populous State. Front Public Health 2020;8. doi:https://doi.org/10.3389/fpubh.2020.578645.
71. Souza CDF de, Paiva JPS de, Leal TC, Silva LF da, Santos LG, Souza CDF de, et al. Spatiotemporal evolution of case fatality rates of COVID-19 in Brazil, 2020. J Bras Pneumol. 2020;46. doi:https://doi.org/10.36416/1806-3756/e20200208.
72. Mir D, Rego N, Resende PC, López-Tort F, Fernandez-Calero T, Noya V, et al. Recurrent dissemination of SARS-CoV-2 through the Uruguayan-Brazilian border. medRxiv. 2021; 2021.01.06.20249026. https://doi.org/10.1101/2021.01.06.20249026.
73. Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, Carabelli A, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. Virological. 2020; https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563. Accessed 4 Jan 2021.
74. Rambaut A. Phylodynamic analysis | 176 genomes | 6 mar 2020 - SARS-CoV-2 coronavirus / nCoV-2019 genomic epidemiology. Virological. 2020; https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356. Accessed 11 Feb 2021.
75. Su YCF, Anderson DE, Young BE, Linster M, Zhu F, Jayakumar J, et al. Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. mBio. 2020;11. https://doi.org/10.1128/mBio.01610-20.
76. Tong KJ, Duchêne DA, Duchêne S, Geoghegan JL, Ho SYW. A comparison of methods for estimating substitution rates from ancient DNA sequence data. BMC Evol Biol. 2018;18(1):70. https://doi.org/10.1186/s12862-018-1192-3.
77. Candido D, Watts A, Abade L, Kraemer MUG, Pybus OG, Croda J, et al. Routes for COVID-19 importation in Brazil. J Travel Med. 2020;27(3). https://doi.org/10.1093/jtm/taaa042.
78. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance. 2020;25:2000045.
79. World Health Organization. COVID-19 Clinical management: living guidance. 2021. https://www.who.int/publications-detail-redirect/WHO-2019-nCoV-clinical-2021-1. Accessed 1 May 2021.
80. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28(19):2520–2. https://doi.org/10.1093/bioinformatics/bts480.
81. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
82. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinforma Oxf Engl. 2009;25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324.
83. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93. https://doi.org/10.1093/bioinformatics/btr509.
84. Quinlan AR. BEDTools: the Swiss-Army tool for genome feature analysis. Curr Protoc Bioinforma. 2014;47:11.12.1–34.
85. Gel B, Serra E. karyoploteR: an R/bioconductor package to plot customizable genomes displaying arbitrary data. Bioinforma Oxf Engl. 2017;33(19):3088–90. https://doi.org/10.1093/bioinformatics/btx346.
86. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7(1):11257. https://doi.org/10.1038/ncomms11257.
87. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. Genome Biol. 2019;20(1):257. https://doi.org/10.1186/s13059-019-1891-0.
88. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a web browser. BMC Bioinformatics. 2011;12(1):385. https://doi.org/10.1186/1471-2105-12-385.
89. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio. 2012; http://arxiv.org/abs/1207.3907. Accessed 14 Nov 2020.
90. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6(2):80–92. https://doi.org/10.4161/fly.19695.
91. 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. https://doi.org/10.1093/bioinformatics/bty407.
92. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80. https://doi.org/10.1093/molbev/mst010.
93. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74. https://doi.org/10.1093/molbev/msu300.
94. Tavare S. Some probabilistic and statistical problems in the analysis of DNA sequences. Some Math Quest Biol DNA Seq Anal Ed Robert M Miura 1986. https://agris.fao.org/agris-search/search.do?recordID=US201301755037. Accessed 1 May 2021.
95. Sagulenko P, Puller V, Neher RA. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4(1). https://doi.org/10.1093/ve/vex042.
96. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly path-O-gen). Virus Evol. 2016;2(1). https://doi.org/10.1093/ve/vew007.
97. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4:vey016.
98. Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, et al. BEAGLE: an application programming Interface and high-performance computing library for statistical Phylogenetics. Syst Biol. 2012;61(1):170–3. https://doi.org/10.1093/sysbio/syr100.
99. Ferreira MAR, Suchard MA. Bayesian analysis of elapsed times in continuous-time Markov chains. Can J Stat. 2008;36(3):355–68. https://doi.org/10.1002/cjs.5550360302.
100. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian Phylogenetics using tracer 1.7. Syst Biol. 2018;67(5):901–4. https://doi.org/10.1093/sysbio/syy032.
101. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017;8:28–36.
102. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography finds its roots. PLoS Comput Biol. 2009;5(9):e1000520. https://doi.org/10.1371/journal.pcbi.1000520.
103. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. Mol Biol Evol. 2016;33(8):2167–9. https://doi.org/10.1093/molbev/msw082.

Publisher's Note

4. CAPÍTULO II

O manuscrito que constitui este capítulo, intitulado “Mutation hotspots and spatiotemporal distribution of SARS-CoV-2 lineages in Brazil, February 2020-2021” visou acompanhar a evolução molecular e a distribuição espaço-temporal do SARS-CoV-2 no território brasileiro realizando inferências filogenéticas e filogeográficas a partir de genomas virais depositados até Fevereiro de 2021. Encontra-se publicado na revista Virus Research (https://www.journals.elsevier.com/virus-research), com fator de impacto JCR 2021 = 3,303 e Qualis/CAPES = A3. O manuscrito e os materiais suplementares estão disponíveis mediante assinatura no seguinte *link*: https://www.sciencedirect.com/science/article/abs/pii/S0168170221002392. Todas as análises descritas no manuscrito, assim como a sua redação, foram realizadas pelo aluno Vinícius Bonetti Franceschi, sendo os demais autores responsáveis por colaborações na escrita ou análises, bem como na sua orientação e obtenção de fomento. Abaixo, todas as páginas do manuscrito publicado foram anexadas para compor o Capítulo II da presente dissertação.

Virus Research 304 (2021) 198532

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Mutation hotspots and spatiotemporal distribution of SARS-CoV-2 lineages in Brazil, February 2020-2021

Vinícius Bonetti Franceschi [a], Patrícia Aline Gröhs Ferrareze [b], Ricardo Ariel Zimerman [c], Gabriela Bettella Cybis [d], Claudia Elizabeth Thompson [a,b,e,*]

[a] Graduate Program in Cell and Molecular Biology (PPGBCM), Center of Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil
[b] Graduate Program in Health Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, RS, Brazil
[c] Department of Infection Control and Prevention, Hospital da Brigada Militar, Porto Alegre, RS, Brazil
[d] Department of Statistics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil
[e] Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), 245/200C Sarmento Leite St, Porto Alegre, RS 90050-170, Brazil

ARTICLE INFO

Keywords:
COVID-19
Severe acute respiratory syndrome coronavirus 2
Infectious diseases
High-Throughput Nucleotide Sequencing
Molecular Epidemiology
Phylogeography

ABSTRACT

The COVID-19 pandemic has already reached more than 110 million people and is associated with 2.5 million deaths worldwide. Brazil is the third worst-hit country, with approximately 10.2 million cases and 250 thousand deaths. International efforts have been established to share information about *Severe acute respiratory syndrome coronavirus 2* (SARS-CoV-2) epidemiology and evolution to support the development of effective strategies for public health and disease management. We aimed to analyze the high-quality genome sequences from Brazil from February 2020-2021 to identify mutation hotspots, geographical and temporal distribution of SARS-CoV-2 lineages by using phylogenetics and phylodynamics analyses. We describe heterogeneous sequencing efforts, the progression of the different lineages along time, evaluating mutational spectra and frequency oscillations derived from the prevalence of specific lineages across different Brazilian regions. We found at least seven major (1–7) and two minor clades related to the six most prevalent lineages in the country and described its spatial distribution and dynamics. The emergence and recent frequency shift of lineages (P.1 and P.2) carrying mutations of concern in the spike protein (e. g., E484K, N501Y) draws attention due to their association with immune evasion and enhanced receptor binding affinity. Improvements in genomic surveillance are of paramount importance and should be extended in Brazil to better inform policy makers about better decisions to fight the COVID-19 pandemic.

1. Introduction

After its initial emergence in China in late 2019 (Huang et al., 2020), *Severe Acute Respiratory Syndrome 2 virus* (SARS-CoV-2) has spread rapidly around the world causing the COVID-19 pandemic (World Health Organization, 2020). New epicenters of the disease have been established throughout 2020, mainly in Europe, USA, and South America (Pei et al., 2020; Ruiu, 2020). As of February 22, 2021, more than 110 million cases and approximately 2.5 million deaths worldwide have been confirmed (Johns Hopkins Coronavirus Resource Center, 2021). Several countries are currently experiencing second waves of infections, decreasing optimism regarding a brief solution to the pandemic.

Since the sequencing of the first SARS-CoV-2 genome (Zhou et al., 2020), international efforts have been established through data sharing in the Global initiative on sharing all influenza data (GISAID) database (https://www.gisaid.org/, Shu and McCauley 2017). These sequencing and metadata information were made public and have enabled the study of the viral spread pattern through space and time. However, sequencing facilities and research investments are very heterogeneous across the world. Asian, European, North American and Oceanian countries have contributed with more data proportionally to the number of cases (Furuse, 2021), while African and South American genomic surveillance have been more limited. Disparities are even deeper on an individual country basis. For instance, as of February 2021, while the United

* Corresponding author at: Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), 245/200C Sarmento Leite St, Porto Alegre, RS 90050-170, Brazil.
E-mail address: cthompson@ufcspa.edu.br (C.E. Thompson).

https://doi.org/10.1016/j.virusres.2021.198532
Received 19 May 2021; Received in revised form 21 July 2021; Accepted 29 July 2021
Available online 5 August 2021

Abbreviations

CTMC:	Continuous Time Markov Chain
GISAID:	Global initiative on sharing all influenza data
hACE2:	human Angiotensin-converting enzyme 2 receptor
HKY:	Hasegawa-Kishino-Yano nucleotide substitution model
HPD:	Highest Posterior Density
INDEL:	Insertion/deletion
JSON:	JavaScript Object Notation
MCC:	Maximum clade credibility
MCMC:	Markov Chain Monte Carlo
RBD:	Receptor Binding Domain
SARS-CoV-2:	Severe acute respiratory syndrome coronavirus 2
SNP:	Single nucleotide polymorphism
TMRCA	: Time of the most recent common ancestor
VOC:	Variants of Concern

Kingdom (UK) leads sequencing efforts (5.85% of cases sequenced), Brazil has sequenced only 0.03% of all cases, despite being the third worst-hit country, with approximately 10.2 million cases and 250 thousand deaths. Currently (mid-July 2021), the UK has expanded sequencing capacity (11.1%), as modestly occurred in Brazil (0.12%). In South America, Chile (0.30%), Ecuador (0.22%), and Uruguay (0.15%) lead sequencing rates, while the remaining countries (except Brazil) that reported > 100,000 cases have < 0.1% sequenced cases (https://www.gisaid.org/, Johns Hopkins Coronavirus Resource Center, 2021).

Many studies were performed to characterize early viral introductions and the transmission dynamics in several countries (*e. g.* China Lu et al. 2020, USA Deng et al. 2020, Maurano et al. 2020 and Worobey et al. 2020, Australia Seemann et al. 2020, Italy Bartolini et al. 2020, United Kingdom da Silva Filipe et al. 2021 and du Plessis et al. 2021). In Brazil, SARS-CoV-2 arrived officially on February 25, 2020, in a returning traveller from Italy, and early efforts were made both at the national (Candido et al., 2020a) and regional levels (Paiva et al., 2020; Xavier et al., 2020) to further characterize the viral introduction and spread. B.1 and derived lineages were prevalent in the country at the beginning of the pandemic and significant movements between state borders after international travel restrictions have been demonstrated (Candido et al., 2020a). Unfortunately, little is known about the viral evolution in the entire Brazilian territory after these earliest studies.

More recently, new SARS-CoV-2 lineages have emerged and are considered as "Variants of Concern" (VOC), mainly those carrying mutations in the spike (S) glycoprotein due to its role in binding to the human Angiotensin-converting enzyme 2 receptor (hACE2). Up to July 2021, there are four VOCs described worldwide, namely: B.1.1.7, B.1.351, P.1, and B.1.617.2. More recently, the World Health Organization (WHO) assigned easy to remember labels based on Greek letters, mainly to avoid discriminatory associations related to places of first detection. Therefore, the four VOCs are named, respectively, Alpha, Beta, Gamma, and Delta following the WHO-nomenclature. The former (B.1.1.7) emerged in England in mid-September 2020 and it is characterized by 14 lineage-specific amino acid substitutions, especially N501Y (a key contact residue interacting with hACE2) and P681H (one of four amino acids comprising the insertion that creates a novel furin cleavage site between S1 and S2) (Rambaut et al., 2020b). The second (B.1.351) emerged in South Africa in October 2020 and harbor a constellation of mutations in the Receptor Binding Domain (RBD) (especially K417N, E484K, and N501Y) (Tegally et al., 2021a). The third is P.1, derived from B.1.1.28, a widespread lineage from Brazil. It was reported in returning travelers from Manaus (Amazonas, Brazil), after arriving in Japan. It has the same three mutations (except for K417T instead of K417N) in RBD as B.1.351, but it arose independently (Faria et al., 2021). More recently, the B.1.617.2 lineage, first detected in India, has also been characterized as a VOC, primarily for carrying a constellation of mutations in the spike protein (especially L452R and P681R) (Peacock et al., 2021), its wide spread worldwide even outperforming other VOCs (Mullen et al., 2021), and reduced antibody sensitivity in vaccinated individuals (Planas et al., 2021). Importantly, B.1.351 and P.1 carry the E484K mutation associated with escape from neutralizing antibodies (Baum et al., 2020; Greaney et al., 2020 and Weisblum et al., 2020). Recently, a E484K harboring virus was identified in a reinfected patient (Nonaka et al., 2021) from Brazil, confirming the ability to evade naturally developed antibodies from previous infection as well. Moreover, all three VOC lineages harbor N501Y mutation, already associated with enhanced receptor binding affinity (Starr et al., 2020), which could lead to increased infectiousness.

It is believed that after more than a year of its emergence, some mutations of SARS-CoV-2 (*e. g.* L452R, E484K, N501Y, P681H/R) have been positively selected, since they may confer adaptive advantages leading to convergent evolution in different lineages spreading across multiple countries (Martin et al., 2021). Despite the low sequencing rate, a deeper analysis of mutations and lineages throughout the Brazilian states would allow a better understanding of viral diversity and spread patterns inside the country. Thus, we aimed to identify mutation hotspots, geographical and temporal distribution of SARS-CoV-2 lineages in the Brazilian territory by using phylogenetics and phylodynamics analyses from high-quality SARS-CoV-2 genome sequences.

2. Materials and methods

2.1. SARS-CoV-2 genomes and epidemiological data retrieval

Complete SARS-CoV-2 genomes (> 29,000 bp) and the associated metadata were obtained from the GISAID database. Considering 2,751 available sequences from Brazil submitted until February 16, 2021, 2,732 were retrieved applying filters for human host and complete collection date. Number of cases per state per day and across Brazil were downloaded from https://covid19br.wcota.me/en/ (Cota, 2020) on the same date. This initiative aggregates data from the Brazilian Ministry of Health and epidemiological bulletins of each federative unit. Geographical maps and general purpose plots were generated using R v3.6.1 (R Core Team, 2020), and the ggplot2 v3.3.2 (Wickham, 2009), geobr v.1.4 (Pereira et al., 2019), and sf v0.9.8 (Pebesma, 2018) packages.

2.2. Mutation analysis

The GenBank RefSeq sequence NC_045512.2 from Wuhan (China) was used as the reference for our analysis. Single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) were assessed by using snippy variant calling pipeline v4.6.0 (https://github.com/tseemann/snippy), which uses FreeBayes v1.3.2 (Garrison and Marth, 2012) as variant caller and snpEff v5.0 (Cingolani et al., 2012) to annotate and predict the effects of variants on genes and proteins. Mutations and lineages were concatenated with associated metadata and counted by Brazilian states using custom Python and R scripts. Histogram of SNPs were generated after running MAFFT v7.471 (Katoh and Standley, 2013) alignment using msastats.py script, plotAlignment and plotSNPHist functions (Du Plessis, 2020).

2.3. Phylogenetics analysis

All available SARS-CoV-2 genomes (537,360 sequences) were retrieved from GISAID on February 16, 2021. These sequences were then subjected to analysis inside NextStrain ncov pipeline (https://github.com/nextstrain/ncov) (Hadfield et al., 2018) through a Brazilian-focused subsampling scheme using time- and worldwide-representative contextual samples.

In this workflow, sequences were filtered out based on high

divergence, incompleteness and sampling date availability. Next, filtered genomes were aligned using nextalign v0.1.6 (https://github.com/neherlab/nextalign) and their ends were masked (100 positions in the beginning, 50 in the end). Maximum likelihood (ML) phylogenetics tree was built using IQ-TREE v2.0.3 (Nguyen et al., 2015), employing the best-fit model of nucleotide substitution as selected by ModelFinder (Kalyaanamoorthy et al., 2017). The root of the tree was placed between lineage A and B (Wuhan/WH01/2019 and Wuhan/Hu-1/2019). The clock-like behavior of the inferred tree was inspected using TempEst v1.5.3 (Rambaut et al., 2016) to generate the root-to-tip regression against sampling dates (correlation coefficient = 0.83, $R^2 = 0.68$). Sequences that deviate more than four interquartile ranges from the root-to-tip regression were removed from the analysis. The subsampled time-scaled ML phylogenetics tree was generated using TreeTime v0.8.1 under a strict clock and a skyline coalescent prior with a rate of 8×10^{-4} substitutions per site per year (Sagulenko et al., 2018). Results were then exported to JSON format to enable interactive genetic and geographical visualization using Auspice. Additionally, ML and time-stamped trees were visualized using FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/) and ggtree R package v2.0.4 (Yu et al., 2017). We identified global lineages using the dynamic nomenclature (Rambaut et al., 2020a) implemented in Pangolin v2.2.2 (https://github.com/cov-lineages/pangolin).

2.4. Bayesian phylogeographic and phylodynamic analysis

All major clades associated with massive community transmission in Brazil (clades 3–6) as defined by the previous maximum likelihood analysis were included in this analysis. Clade 7 was discarded due to the low correlation of genetic distances and sampling dates ($R^2 = 2.11 \times 10^{-2}$, correlation coefficient = 0.1488). Additionally, most of its sequences ranged from 2020.8 and 2021.1 and had high divergence, suggesting important undersampling for this clade (represented by sequences from the P.2 lineage). The spatiotemporal diffusion of these important circulating lineages through Brazil were separately estimated for each clade using a Bayesian Markov Chain Monte Carlo (MCMC) approach as implemented in BEAST v1.10.4 (Suchard et al., 2018), using the BEAGLE library v3 (Ayres et al., 2012) to enhance computational time. Time-scaled Bayesian trees were estimated in BEAST using: HKY+Γ nucleotide substitution model (Hasegawa et al., 1985), a strict molecular clock model with a Continuous Time Markov Chain (CTMC) (Ferreira and Suchard, 2008) prior (mean rate = 8×10^{-4}), and a parametric Exponential Growth coalescent tree prior (Griffiths and Tavare, 1994). Although it is not the most appropriate model to reconstruct the demographic and evolutionary processes, it was used given the difficulty of convergence in more complex models due to the lack of sufficient evolutionary information contained in the data. For clade 4, we used a non-parametric Bayesian Skygrid prior with 10 parameters (Gill et al., 2013).

Viral migrations across time were reconstructed using a Brownian random walk continuous phylogeographic model (Lemey et al., 2010; Pybus et al., 2012) to generate a posterior distribution of 1,000 trees whose internal nodes are associated with geographic coordinates. We assigned sampling coordinates using the centroid point of each municipality (when available) and the centroid of the Brazilian state's capital (when the municipality was unavailable) using a random jitter window size of 0.01 to add a small amount of noise to duplicated sampling coordinates assigned to the tips of the tree.

Two MCMC chains were run for > 100 million generations and convergence of the MCMC chains was inspected using Tracer v1.7.1 (Rambaut et al., 2018). After removal of 10% burn-in, log and tree files were combined using LogCombiner v1.10.4 (Suchard et al., 2018). Maximum clade credibility (MCC) trees were generated using TreeAnnotator v1.10.4 (Suchard et al., 2018). The seraphim R package (Dellicour et al., 2016) and SPREAD3 (Bielejec et al., 2016, p. 3) were used to extract and map the spatiotemporal information embedded in the MCC trees. Finally, these trees were visualized using the ggtree R package v2.0.4 (Yu et al., 2017).

3. Results

3.1. Distribution of Brazilian sequences through time and space

Sequencing efforts from Brazil were concentrated mainly in the first epidemic wave (March to April, 2020) (Fig. 1A and 1B). In March, 503 genomes (8.64% of the confirmed cases) and in April, 942 sequences (1.16% of the cases) were sampled. All following months had a sequencing rate below 1% (Table S1). All Brazilian states sequenced less than 0.1% of the confirmed cases through the first year of the pandemic (Fig. 1C). From the Southeast region, the states of Rio de Janeiro (0.09%) and São Paulo (0.06%) have led the country's sequencing initiatives, followed by Rio Grande do Sul (0.045%) from the South region, Amazonas (0.041%) from the North region and Pernambuco (0.040%) from the Northeast region. The Centre-West was the region with the lowest sequencing rate (Fig. 1C, Table S2).

The Southeast region contributed with 1,704 genomes (62.53%), followed by Northeast (n = 359; 13.17%), South (n = 319; 11.71%), North (n = 310; 11.38%), and Centre-West (1.21%) regions (Fig. 1D, Table S3). In total, 59 different lineages were detected in Brazil, and the states sequencing more genomes (São Paulo, Rio de Janeiro and Rio Grande do Sul) detected higher numbers of circulating lineages (33, 17 and 16, respectively) (Fig. 1E, Table S3). Importantly, all states that sequenced at least 100 genomes identified ≥ 10 lineages (Table S3).

3.2. High frequency mutations

A total of 3,919 mutations were detected across the 2,731 Brazilian genomes and only 354 (12.96%) occurred in >5 sequences, 44 (1.61%) in > 50 genomes, and 38 (1.39%) in > 100 sequences (Fig. 2, Table 1). Twenty-five (65.79%) of these 38 mutations were non-synonymous. Of these, 11 (44.0%) were in the spike protein, 5 (20.0%) in the nucleocapsid protein, and 5 (20.0%) in the ORF1ab polyprotein (Table 1).

Three mutations were found in > 95% of the genomes: A23403G (S:D614G), C14408T (ORF1ab:L4715), and C3037T (ORF1ab:F924), which are signatures of the B.1 and derived lineages that spread early in the pandemic. The adjacent replacement GGG28881AAC (N:RG203-204KR) was found in 85.17%, representing a clear signature of B.1.1 lineage. The defining-mutations of the Brazilian most widespread lineages (B.1.1.28 and B.1.1.33) were also found in high abundance. The G25088T (S:V1176F) replacement from B.1.1.28 occurred in 47.56% of all sequences, while T29148C (N:I292T) and T27299C (ORF6:I33T) from B.1.1.33 in ≈32.5%. The G23012A (S:E484K) mutation in the Receptor Binding Domain (RBD) of spike that recently emerged independently in three Brazilian lineages (B.1.1.33, P.1 and P.2) is already among the most frequent detected up to February, 2021 (11.42%). The viruses containing the E484K mutation have been spreading mostly between mid-2020 up to early-2021. Additionally, the multiple lineage-defining mutations found in emergent P.1 and P.2 lineages from Brazil were observed in > 100 and > 200 genomes, respectively (Fig. 2, Table 1).

3.3. Lineage dynamics of Brazilian epidemic

In March 2020, the majority of Brazilian sequences belonged to 3 lineages: B.1 (n = 101; 20.08%), B.1.1.28 (n = 156; 31.01%) and B.1.1.33 (n = 131; 26.04%). The first was probably introduced in Brazil through multiple imports from other continents, and the others probably emerged from community transmission inside the country (Candido et al., 2020a). Between April and August, > 75% of all sequences per month were classified as B.1.1.28 or B.1.1.33. During October and November, the B.1.1.28 derived lineage P.2 was the most prevalent (n = 32; 37.65% and n = 92; 40.71%), while B.1.1.28 and B.1.1.33

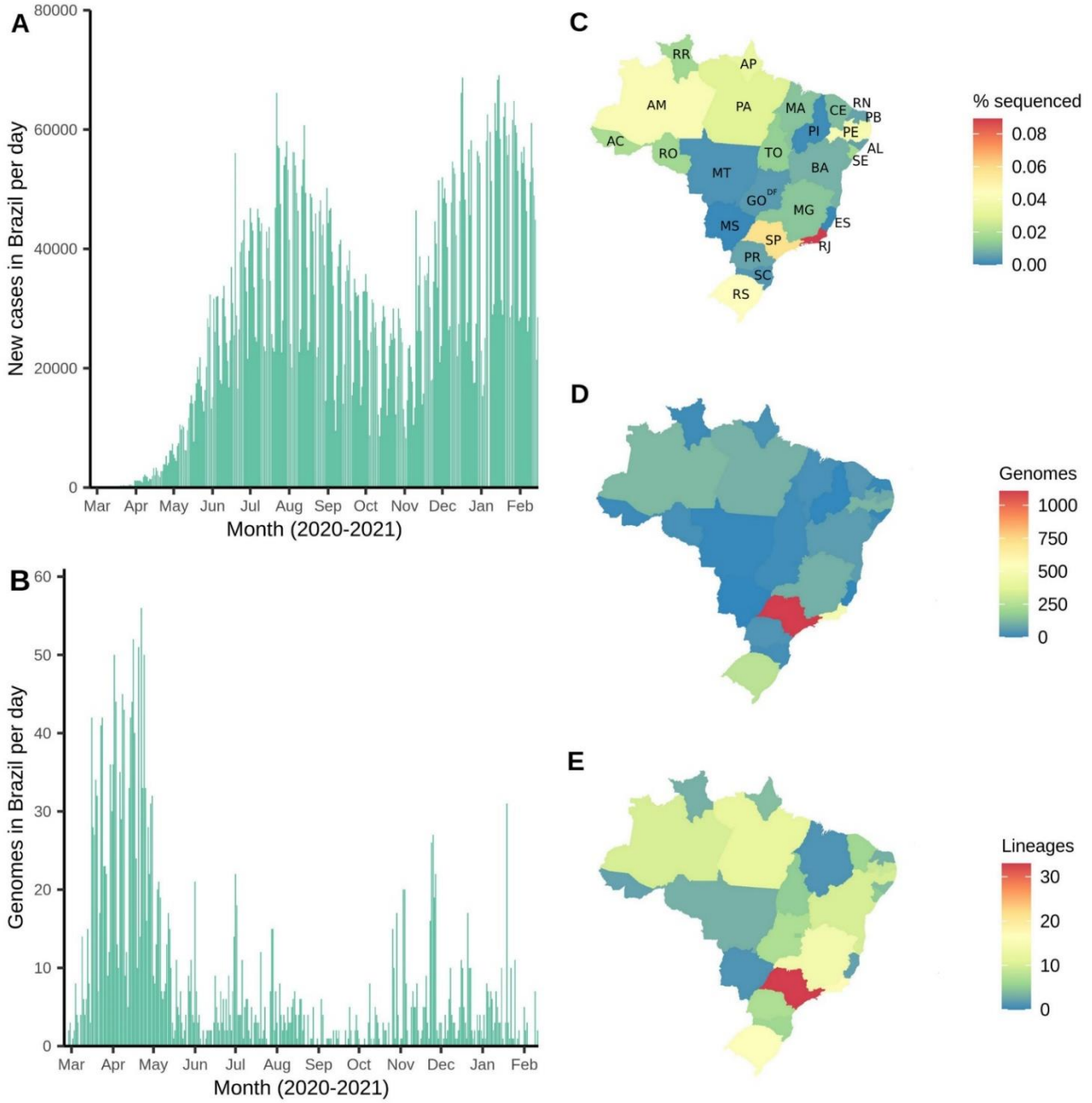

Fig. 1. Distribution of Brazilian genomes through time (end-February 2020 to mid-February 2021) and space (Brazilian states). (A) Number of new cases per day in Brazil over time. (B) Number of genomes (sequenced cases) in Brazil over time (date of sample collection). (C) Fraction of genomes sequenced related to number of cases per Brazilian state. (D) Total number of genomes deposited per Brazilian state. (E) Total number of different SARS-CoV-2 lineages detected per Brazilian state. State abbreviations: AC=Acre; AL=Alagoas; AM=Amazonas; Amapá=AP; BA=Bahia; CE=Ceará; DF=Distrito Federal; ES=Espírito Santo; GO=Goiás; MA=Maranhão; MG=Minas Gerais; MS=Mato Grosso do Sul; MT=Mato Grosso; PA=Pará; PE=Pernambuco; PB=Paraíba; PI=Piauí; PR=Paraná; RJ=Rio de Janeiro; RN=Rio Grande do Norte; RO=Rondônia; RR=Roraima; RS=Rio Grande do Sul; SC=Santa Catarina; SE=Sergipe; SP=São Paulo; TO=Tocantins.

represented together < 50% of the genomes. From December 2020 onward, the VOC P.1 emerged in Manaus and has been established, together with P.2, as the most prevalent lineages represented by sequencing data (Fig. 3A).

Regarding distribution between the five different Brazilian regions, the Southeast and Southern regions sequenced a larger proportion of B.1.1.28 and B.1.1.33 viruses. The Northeast apparently has a slightly different dynamics, since B.1.1.74 (n = 138; 38.44%) is the most prevalent lineage followed by B.1.1.33 (n = 91; 25.35%). In the Northern region, P.2 is already the second most prevalent lineage (n = 81; 26.13%) (Fig. 3B and 3C), but this may be related to the low quantity of sequences from the beginning of the pandemic, and the higher surveillance in the region at present due to enhanced sequencing efforts after the emergence of P.1. In the Centre-West, the extremely low sequencing rate prevents us from making any assumptions about the genetic diversity of the circulating lineages (Fig. 3B).

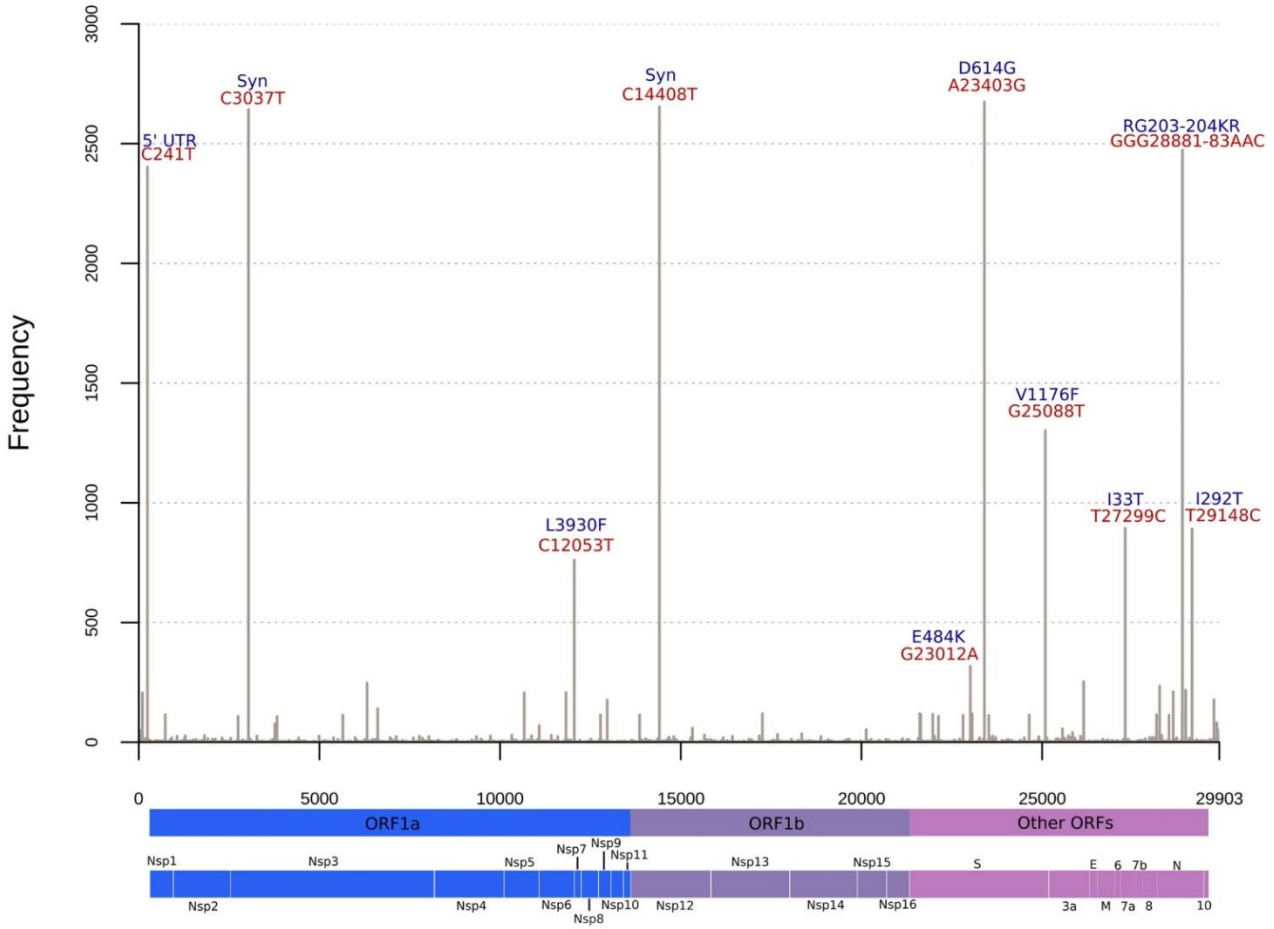

Fig. 2. High frequent mutations across Brazilian sequences. Nucleotide replacements occurring in >250 genomes are indicated in red and associated amino acid substitutions in blue. Other mutations occurring in less than 250 genomes but more than 100 are indicated in Table 1. Syn: Synonymous.

Statewide view shows the wide distribution of lineages B.1.1.28, B.1.1.33 and P.2 across almost all Brazilian states. Although the P.1 lineage has been represented by genomes from only 7 states (AM, PA, RS, RO, RR, SC, and SP) up to February 16, 2021, other states have already reported its detection after this date. States reporting a higher proportion of B.1 lineage (CE, GO, MG, SC, and SE) have concentrated their sequencing efforts mainly in the early phase of the pandemic. B.1.1.74 is overrepresented (> 50%) in two states from the Northeast (PE and PB), suggesting a more geographically restricted distribution. Additionally, the contribution of other lineages that are not among the 10 most frequent in the country are apparently limited (Fig. 3C and Fig. S1).

The median time between the first and last detection of the 18 lineages represented by more than five Brazilian genomes was 179.0 days (~6 months). However, the standard deviation was high (119.58 days). Eleven of the 18 lineages were sampled in different states between the first and last detection, suggesting its spread through the Brazilian territory (Table S4). B and B.1 lineages have presented a relatively low time of spread, while several B.1.1 derived lineages probably have spreaded for a longer period (*e. g*, B.1.1.143, B.1.1.28, B.1.1.314, B.1.1.33, B.1.1.74, B.1.1.94) (Fig. S2). B.1.1.7 (Alpha) was firstly detected in the country in mid-December (Claro et al., 2021), while P.1 was found in November in Manaus (Amazonas, Brazil) and Japan (Faria et al., 2021). P.2 was discovered in mid-April 2020, but has been increasing in frequency since August 2020 (Fig. 3A, Fig. S2).

3.4. Phylogenetic analysis of SARS-CoV-2 genomes from Brazil

We obtained 10,573 time-, geographical- and genetic-representative genomes to proceed with phylogenetic inferences. Of these, 1,135 were from Africa, 1,963 were from Asia, 3,248 from Europe, 612 from North America, 285 from Oceania, and 3,330 from South America. Among the latter, 2,350 were from Brazil.

The maximum likelihood tree showed the evolutionary diversity of SARS-CoV-2 sequences across Brazilian territory (Fig. 4A). Almost all sequences from Brazil harbor the S:D614G and ORF1b:P314L, which were imported from other continents to Brazil (mainly to Southeastern states) in the early epidemic wave of COVID-19. After the importation of B.1.1 lineages characterized by N:R203K and N:G204R mutations, community transmission massively occurred and gave rise to B.1.1.28 (S:V1176F) and B.1.1.33 lineages (ORF6:I33T and N:I292T), widely distributed in Brazilian regions along the first year of the pandemic (Fig. 4A and 4B). B.1.1.28 has diversified into two lineages, P.1 (20J/501Y.V3) and P.2, which are already widely represented by many sequences and distributed across all regions. The larger branch length leading to P.1 lineage draws attention and its emergence is probably driven by an accelerated molecular evolution (Fig. 4B and 5A).

We estimated an evolutionary rate of 7.76×10^{-4} substitutions per site per year (23.22 mutations per year) for the Brazilian-focused subsampling (Fig. 5A). Since the end of 2020, two evolutionary patterns are observed in the root-to-tip regression of sampling dates. Despite the

Table 1
Mutations of Brazilian genomes found in > 100 genomes, associated effects on encoded proteins and major associated lineages.

REF	Genomic position	ALT	Effect	Amino acid	Gene	Product	Number of sequences	% in Brazilian genomes	Major associated lineages
A	23403	G	Missense	D614G	S	Surface glycoprotein	2651	97.07	B.1 and derived
C	14408	T	Synonymous	L4715L	ORF1ab	RdRp	2648	96.96	B.1 and derived
C	3037	T	Synonymous	F924F	ORF1ab	nsp3	2635	96.48	B.1 and derived
C	241	T	Intergenic	-	-	-	2374	86.93	B.1 and derived
GGG	28881	AAC	Missense	RG203-204KR	N	N phosphoprotein	2326	85.17	B.1.1 and derived
G	25088	T	Missense	V1176F	S	Surface glycoprotein	1299	47.56	B.1.1.28, B.1.1.143, B.1.1.94, P.1, and P.2
T	29148	C	Missense	I292T	N	N phosphoprotein	888	32.52	B.1.1.33, B.1.1.314, N.1, N.4
T	27299	C	Missense	I33T	ORF6	ORF6 protein	885	32.41	B.1.1.33, B.1.1.314
C	12053	T	Missense	L3930F	ORF1ab	nsp7	761	27.87	B.1.1.28, B.1.1.74, B.1.1.143, P.2
G	23012	A	Missense	E484K	S	Surface glycoprotein	312	11.42	B.1.1.33, P.1, P.2
T	26149	C	Missense	S253P	ORF3a	ORF3a protein	249	9.12	B.1.1.28, P.1
A	6319	G	Synonymous	P2018P	ORF1ab	nsp3	244	8.93	B.1.1.28, P.1
C	28253	T	Synonymous	F120F	ORF8	ORF8 protein	222	8.13	B.1.1.28, B.1.1.33, P.2
G	28975	T	Missense	M234I	N	N phosphoprotein	217	7.95	B.1.1.28, P.2
G	28628	T	Missense	A119S	N	N phosphoprotein	211	7.73	P.2
C	11824	T	Synonymous	I3853I	ORF1ab	nsp6	207	7.58	P.2
T	10667	G	Missense	L3468V	ORF1ab	3C-like proteinase	206	7.54	P.2
A	12964	G	Synonymous	G4233G	ORF1ab	nsp9	176	6.44	P.2
A	6613	G	Synonymous	V2116V	ORF1ab	nsp3	137	5.02	B.1.1.28, P.1
GTCTGGTTTT	11287	G	Deletion	3675-3677 SGF	ORF1ab	nsp6	130	4.76	B.1.1.7, P.1
C	29754	T	Intergenic	-	-	-	117	4.28	P.2
AGTAGGG	28877	TCTAAAC	Missense	RG203-204KR	N	N phosphoprotein	116	4.25	B.1.1.28, P.1
C	21614	T	Missense	L18F	S	Surface glycoprotein	116	4.25	B.1.1.28, P.1
G	17259	T	Missense	S5665I	ORF1ab	Helicase	116	4.25	P.1
A	23063	T	Missense	N501Y	S	Surface glycoprotein	115	4.21	B.1.1.7, P.1
C	21638	T	Missense	P26S	S	Surface glycoprotein	113	4.14	P.1
T	733	C	Synonymous	D156D	ORF1ab	Leader protein	111	4.06	P.1
C	13860	T	Missense	T4532I	ORF1ab	RdRp	111	4.06	P.1
C	24642	T	Missense	T1027I	S	Surface glycoprotein	111	4.06	P.1
C	12778	T	Synonymous	Y4171Y	ORF1ab	nsp9	111	4.06	P.1
C	21621	A	Missense	T20N	S	Surface glycoprotein	111	4.06	P.1
C	28512	G	Missense	P80R	N	N phosphoprotein	110	4.03	P.1
A	5648	C	Missense	K1795Q	ORF1ab	nsp3	109	3.99	P.1
C	23525	T	Missense	H655Y	S	Surface glycoprotein	109	3.99	P.1
G	28167	A	Missense	E92K	ORF8	ORF8 protein	108	3.95	P.1
G	21974	T	Missense	D138Y	S	Surface glycoprotein	107	3.92	B.1.1.33, P.1
G	22132	T	Missense	R190S	S	Surface glycoprotein	105	3.84	P.1
C	2749	T	Synonymous	D828D	ORF1ab	nsp3	105	3.84	P.1

Ref: Reference nucleotide(s); ALT: Replaced nucleotide(s); UTR= Untranslated region; ORF=Open reading frame; S: Spike; N: Nucleocapsid; nsp: nonstructural protein; RdRp: RNA-dependent RNA polymerase.

majority of sequences following the expected substitution rate for SARS-CoV-2, we observe a rise in P.1 (20J/501Y.V3) and B.1.1.7 (20I/501Y.V1) abundance, both characterized by an abnormal clock rate (Fig. 5A) and a constellation of mutations in the spike protein. Time-resolved maximum likelihood tree highlighted the importation of lineages in the early phase of the pandemic and its rapid diversification through community transmission inside the country, leading to the spread of two main lineages (B.1.1.28 and B.1.1.33) within and between state borders (Fig. 5B) and a more restricted circulation of B.1.1.74 lineage in the Northeast. A time-resolved phylogeny considering only Brazilian sequences reinforces the nationwide distribution of multiple SARS-CoV-2 lineages and clades (Fig. 5C and Fig. S3).

We found at least seven major clades (1–7) (Fig. 6) and two minor (4.2 and 5.3) (Fig. S4) related to the six most prevalent Brazilian lineages. Clade 1 (B.1 lineage) is represented by 33 genomes, > 50% from the Southeast with a few introductions to other regions and a restricted time of spread until around May 2020 (Fig. 6B). Clade 2 includes 98 genomes and is associated with B.1 and B.1.212 lineages. B.1 (Clade 2.1) sequences from this clade are mostly restricted to Southern and Southeast Brazil, while B.1.212 (Clade 2.2) has spread mainly in the Northeast

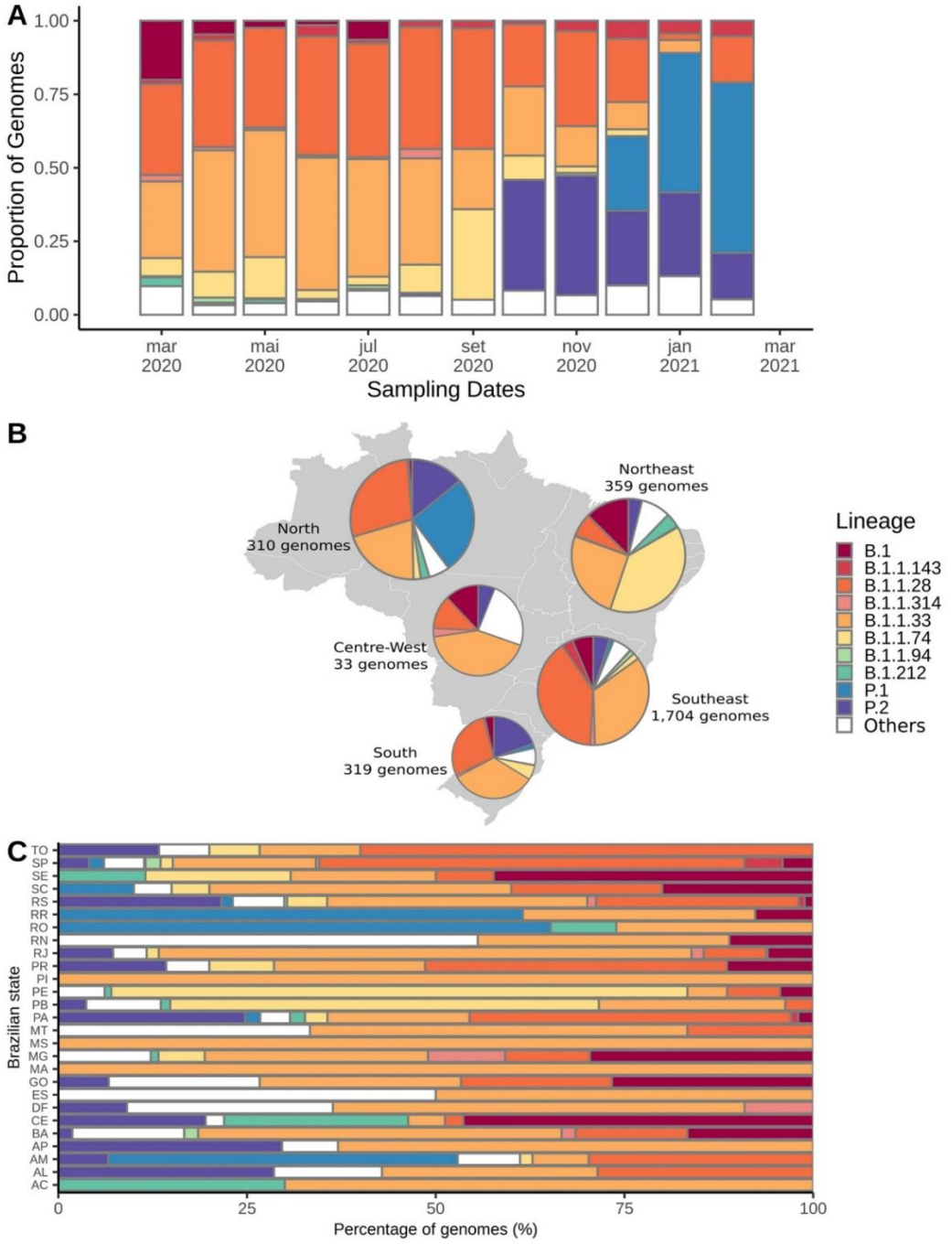

Fig. 3. Distribution of the 10 most prevalent SARS-CoV-2 lineages (n>25) inside Brazil from end-February 2020 to mid-February 2021. (A) Frequency of these lineages through time in the entire Brazil. (B) Map showing the fraction of each of these lineages across all five Brazilian regions. (C) Distribution of these lineages across all Brazilian states, proportional to the number of sequenced genomes.

and Northern regions (Fig. 6C). Clade 3 is represented by 804 genomes, mostly from B.1.1.33 lineage. It reached an unprecedented dissemination through Brazil since March 2020, whose introductions in different regions and states were mostly driven by the Southeast followed by massive community transmission and the establishment of local clusters (Fig. 6D). Clade 4 ($n = 107$) is characterized by B.1.1.74 sequences and it is more geographically restricted to the Northeast, especially Pernambuco ($n = 41$) and Paraiba (n=23). However, it accounts for occasional introductions to Southern and Southeastern regions (Fig. 6E). Clade 5 harbors B.1.1.28 (Clade 5.1) and P.1 (Clade 5.2) sequences. B.1.1.28 sequences are mostly widespread in Northern and Southeastern regions (Fig. 6F), and P.1 genomes ($n = 96$) are mostly found in Northern Brazil, especially Amazonas ($n = 50$) and Rondônia ($n = 15$), and Southeast (São Paulo; $n = 17$) (Fig. 6H). Clade 6 is also represented by genomes that fall into the B.1.1.28 lineage ($n = 529$). However, it is more widespread in the Southeast ($n = 436$), especially in São Paulo ($n = 412$) and

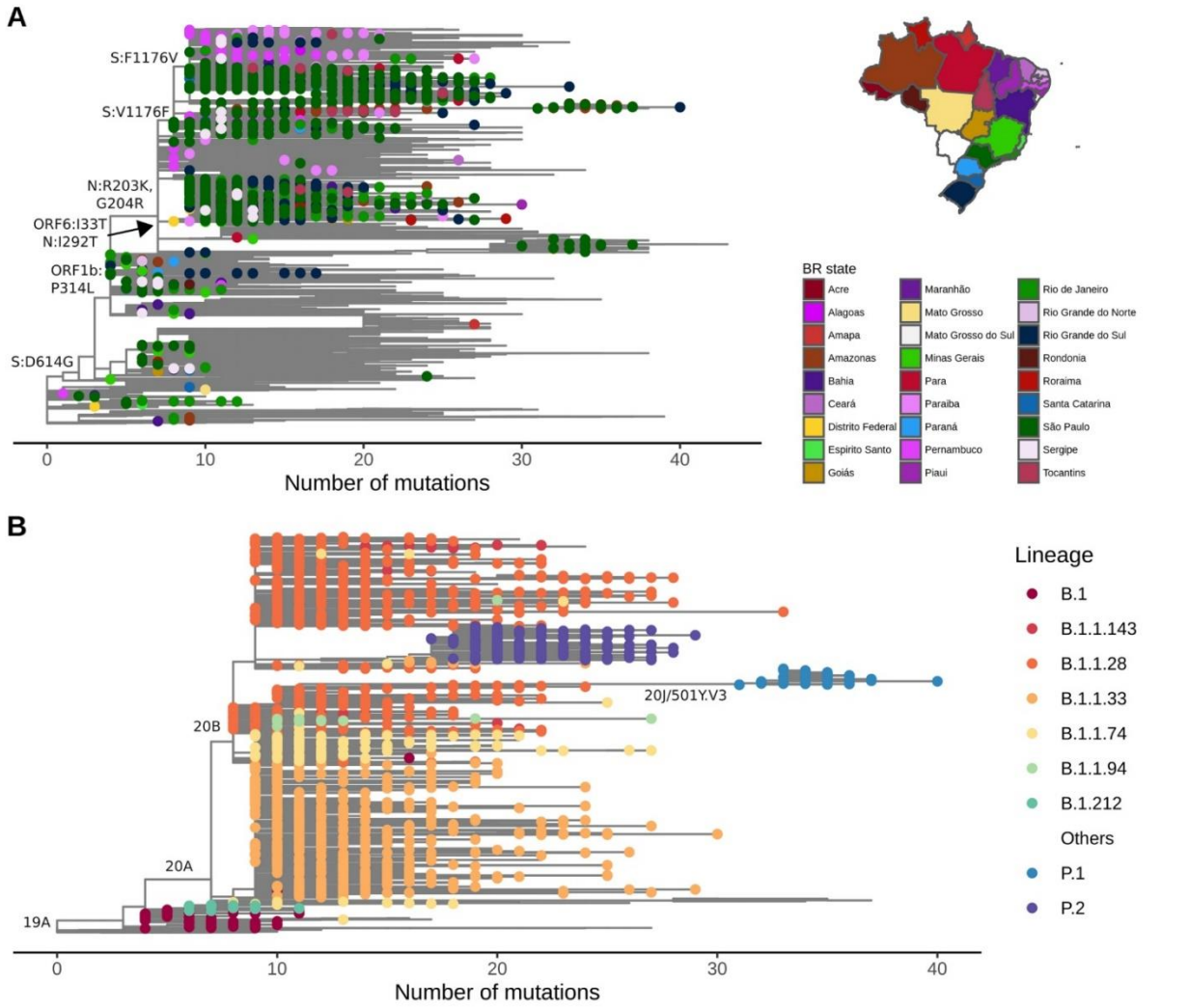

Fig. 4. Evolutionary distribution of SARS-CoV-2 genomes from Brazil. (A) Maximum likelihood phylogenetics tree of 2,346 Brazilian sequences and 8,227 additional global representative genomes. States belonging to each specific Brazilian region are colored using similar colors depicted in the map (Centre-West: yellow; North: red; Northeast: purple and pink; South: blue; Southeast: green). Key mutations are represented in the respective branches. (B) Maximum likelihood phylogenetics tree dropping sequences from other countries and highlighting the most frequent Brazilian lineages presented in Fig. 3. Nextstrain clades are represented in key branches. Evolutionary rate is represented by the number of mutations (divergences) related to SARS-CoV-2 reference sequence (NC_045512.2).

in the Southern region ($n = 72$), especially Rio Grande do Sul ($n = 65$) (Fig. 6G). Clade 7 is highly supported by sequences of P.2 lineage ($n = 204$). This clade is also widely distributed throughout Brazilian regions, especially the Southeast ($n = 83$), South ($n = 63$), and North ($n = 42$), even giving rise to new local clusters in the end of 2020 and early-2021 (Fig. 6H).

3.5. Phylogeographic and phylodynamic patterns of SARS-CoV-2 spread in Brazil

Considering the four major clades mostly represented by Brazilian sequences and with sufficient temporal signal (clades 3 to 6), we reconstructed their spatiotemporal diffusion in the Brazilian territory.

Clade 3 (lineage B.1.1.33) presented a median evolutionary rate of 5.11×10^{-4} (95% Highest Posterior Density (HPD): 4.68×10^{-4} to 5.57×10^{-4}). The phylogeographic reconstruction suggests multiple introductions of this clade in different Brazilian regions followed by massive intrastate spread and few transmissions to surrounding states (Fig. 7A, Fig. S5A, Video S1). It can be illustrated in the MCC tree, which shows five subclades, two mostly represented by sequences from the Southeast, one from the South, one from the Northeast and the other from the North (Fig. S5A).

Clade 4 (lineage B.1.1.74) probably emerged in the Southeast and rapidly spread to Northeast and Southern regions, dispersing progressively from state to state in these regions. A new introduction into the Northeast occurred around September 2020, seeding transmission events to neighboring Northern states (Fig. 7B, Fig. S5B, Video S2). This clade had a median evolutionary rate of 6.66×10^{-4} (95% HPD: 5.49×10^{-4} to 7.79×10^{-4}) and is more geographically restricted to the Brazilian states in the eastern coast, especially states from the Northeast.

Clade 5 (lineages B.1.1.28 and P.1) had a median evolutionary rate of 4.92×10^{-4} (95% HPD: 3.94×10^{-4} to 5.82×10^{-4}). This estimate led to an older dating (2018.63, 95% HPD: 2018.03 to 2019.14) of the most recent common ancestor (TMRCA) and instabilities in the tree rooting. Thus, time- or rate-related inferences for this clade should be considered with extreme caution. A hallmark of this clade is its rapid diffusion to all five Brazilian regions (until around May 2020), followed by massive local transmission within the same state or neighboring states in the

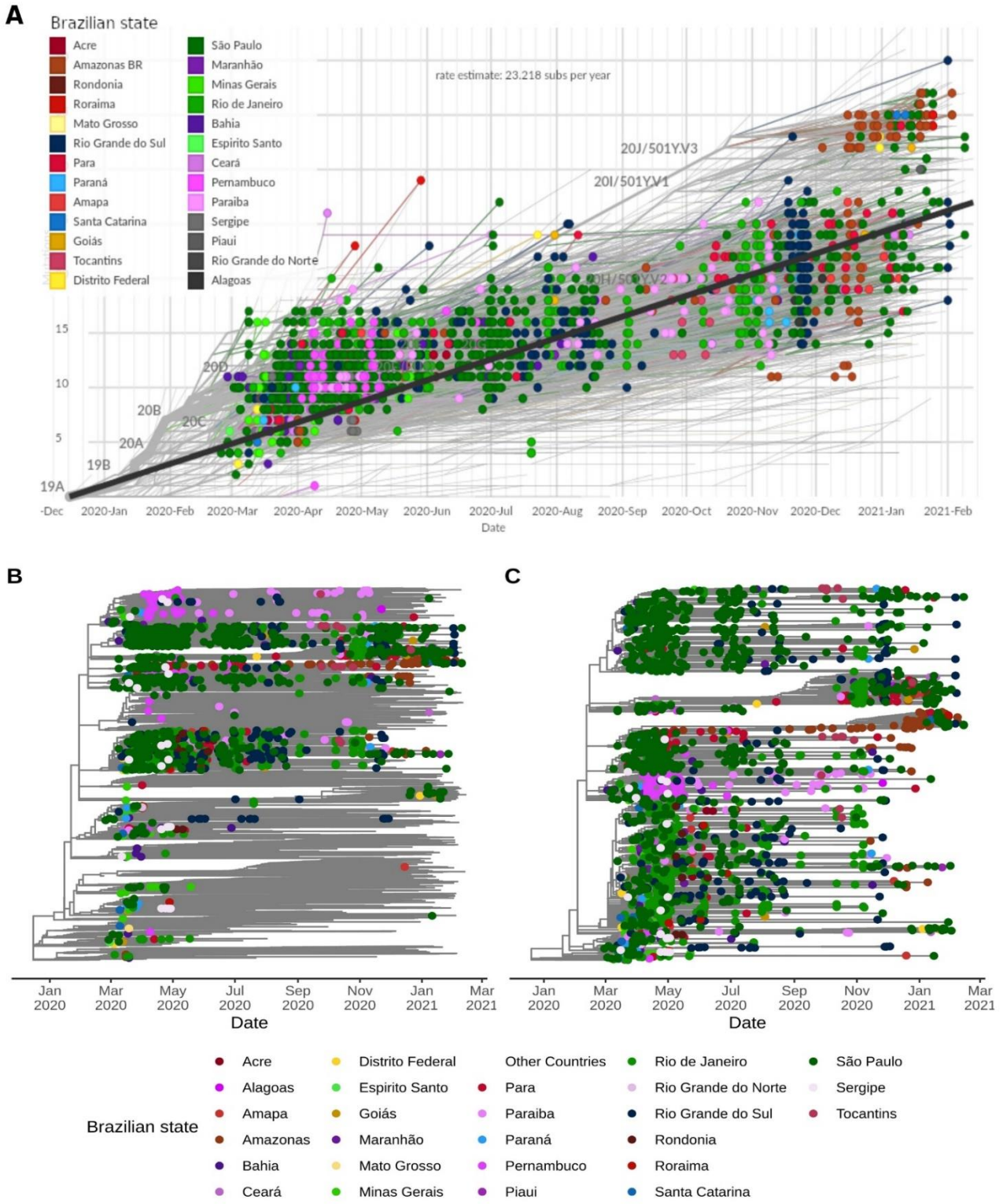

Fig. 5. Molecular clock estimates of SARS-CoV-2 genomes from Brazil. (A) Root-to-tip regression of genetic distances (in number of mutations) against sampling dates filtered by Brazilian sequences. States belonging to each specific Brazilian region are colored using similar colors (Centre-West: yellow; North: red; Northeast: pink and gray; South: blue; Southeast: green). Nextstrain clades are represented in key branches. (B) Time-resolved Maximum Likelihood phylogenetics tree of the 10,573 worldwide genomes included colored by Brazilian states. (C) Maximum likelihood phylogenetics tree considering only the dynamics inside Brazil. In (B) and (C), state colorings follow the same scheme as (A), except for Northeast, where purple and pink colors define sequences from this region. Tree topology remains the same for these two trees, but node ordering is slightly different.

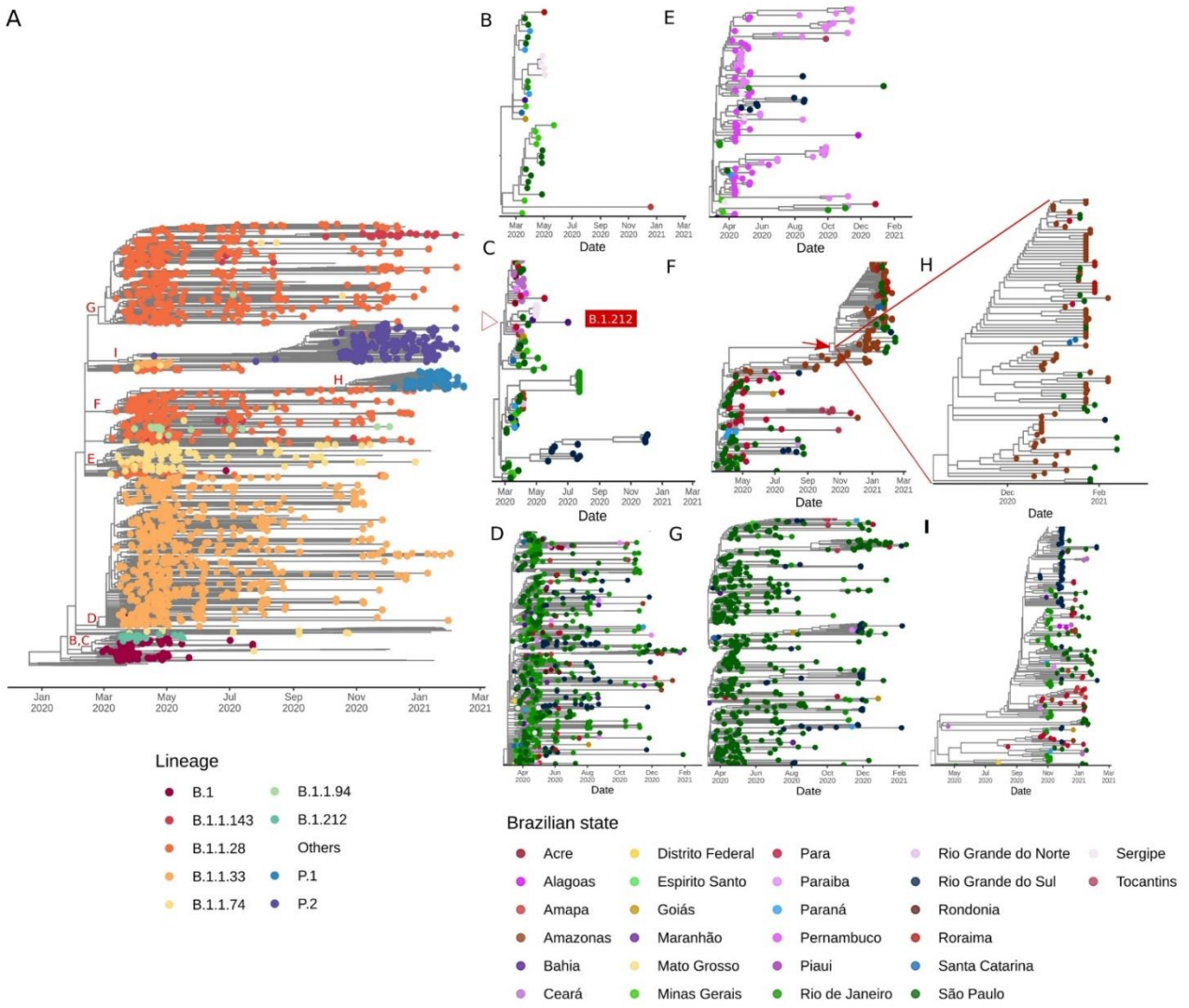

Fig. 6. Zoom-in on clades corresponding to the major six Brazilian lineages. (A) Time-resolved ML tree of Brazilian sequences colored by PANGO lineages. Letters around clades are augmented in the respective figures and colored by Brazilian states. (B) Clade 1 is represented by sequences from the B.1 lineage. (C) Clade 2 has sequences from the B.1 and B.1.212 lineages (D) Clade 3 corresponds to the B.1.1.33 lineage. (E) Clade 4 is represented by B.1.1.74 genomes. (F) Clade 5 harbor sequences from B.1.1.28 and P.1. (G) Clade 6 has sequences from the B.1.1.28 lineage. (H) Zoom-in on P.1 sequences from Clade 5. (I) Clade 7 corresponds to the P.2 lineage. From (B) to (I), states belonging to each specific Brazilian region are colored using similar colors (Centre-West: yellow; North: red; Northeast: purple and pink; South: blue; Southeast: green).

same region, with higher prevalence in Amazonas (North) and São Paulo (Southeast) (Fig. 7C, Fig. S5C, Video S3).

Clade 6 (B.1.1.28) had a median evolutionary rate of 6.06×10^{-4} (95% HPD: 5.46×10^{-4} to 6.37×10^{-4}). Each of the three subclades within this clade are homogeneously distributed in states from the same region (Fig. S5D) and phylogeographic estimates highlighted early introductions in the Southeast, Southern and Northeast regions seeding mostly intra and interstate spread, mainly in the Southeast and Southern Brazil over 2020 (Fig. 7D, Video S4).

4. Discussion

Viral sequencing is essential to track viral evolution and spread patterns. Despite the initial efforts to obtain a representative genomic dataset of the Brazilian first epidemic wave to better characterize viral introductions and early spread (Candido et al., 2020a), the initiatives across different regions have been limited and non-uniform after on. Interestingly, the previously described clades (Candido et al., 2020a) have spread and diversified through the country. Clade 1 (after named B.1.1.28) was mostly restricted to the Southeast (São Paulo) and Clade 2 (after named B.1.1.33) was already present in 16 states in this early phase. In this work, we showed the emergence of at least four clades derived from B.1.1.28. The first was widely distributed in Brazilian regions (Clade 5.1) and evolved to P.1 lineage (Clade 5.2). The second (Minor clade 5.3) and third (Clade 6) are most widespread in the Southeast, accounting for a few introductions in other regions. The fourth gave rise to P.2 lineage (Clade 7), which is distributed in all Brazilian regions. B.1.1.33 continues to be composed of a larger clade (Clade 3) with a wide distribution among all Brazilian regions.

The B.1.1.33 lineage was studied in further detail using 190 genomes from 13 Brazilian states, showing its variable abundance in different states (ranging from 2% in Pernambuco to 80% in Rio de Janeiro), and

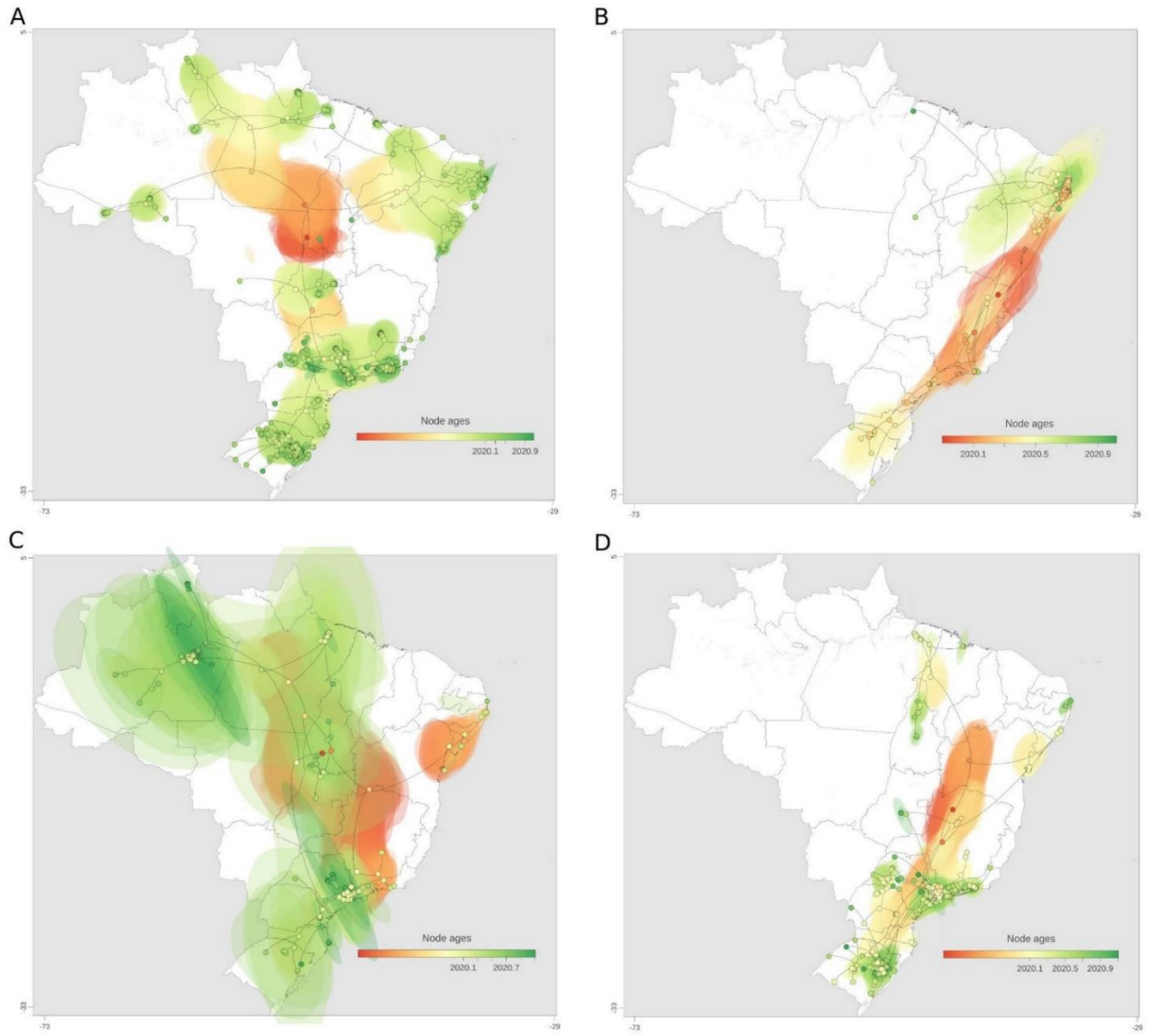

Fig. 7. Spatiotemporal reconstruction of the dispersal history of SARS-CoV-2 in Brazil in the first year of the COVID-19 pandemic. Reconstructions of (A) Clade 3, (B) Clade 4, (C) Clade 5, and (D) Clade 6 are represented. MCC trees and 80% HPD regions are based on 1,000 trees subsampled from the posterior distribution of a continuous phylogeographic analysis. Nodes of the MCC tree are coloured according to their time of occurrence. 80% HPD regions were computed for successive time layers and then superimposed using the same colour scale reflecting time

its moderate prevalence in South American countries (5–18%). Surprisingly, this lineage was firstly detected in early-March in other American countries (*e. g.,* Argentina, Canada, and USA), and additional analysis suggest that an intermediate lineage (B.1.1.33-like) most probably arose in Europe and was later disseminated to Brazil, where its spread gave origin to lineage B.1.1.33 (Resende et al., 2021) and possibly seeded secondary outbreaks in Argentina and Uruguay (Mir et al., 2021; Resende et al., 2021).

The states of Pernambuco (Northeast) and Minas Gerais (Southeast) presented more restricted viral dynamics. In Pernambuco, 88% of 101 early sequences were classified as lineage B.1.1 and six local B.1.1 clades were seeded through both national and international traveling (Paiva et al., 2020). This finding is consistent with the prevalence of B.1.1.74 lineage found in the Northeast, especially in Pernambuco. In Minas Gerais, 92.5% of the 40 genomes from March 2020 belong to the B lineage (mostly B.1.1) and epidemiological analysis revealed that the distribution of cases and deaths was more spatially uniform, while in other Southeastern states it was more centralized around capital cities (Xavier et al., 2020).

Studies from Rio de Janeiro (Southeast) and Rio Grande do Sul (South) identified B.1.1.28 and B.1.1.33 in higher proportion from April to December 2020 (Franceschi et al., 2021; Francisco Jr et al., 2021; Voloch et al., 2021). The emergence of a B.1.1.28-derived lineage carrying the E484K mutation (later named P.2) was dated in July 2020, however it began to appear more frequently and almost simultaneously in October 2020 in the Rio de Janeiro state (Voloch et al., 2021) and in the small municipality of Esteio, Rio Grande do Sul (Franceschi et al., 2021), suggesting its wide distribution in the Southern and Southeastern regions of Brazil and uncertainty regarding its origin. This assumption and the frequency increase of B.1.1.28 and derived lineages were

corroborated by another study from several municipalities of Rio Grande do Sul, which found that 86% of the sequenced genomes were classified as B.1.1.28 and ~50% of these belong to the new lineage P.2 (Francisco Jr et al., 2021). Here, we found that P.2 is already distributed in all Brazilian regions up to mid-February 2021.

Recent findings using 250 genomes (March 2020 to January 2021) from Manaus showed that the first exponential growth phase was driven mostly by the dissemination of B.1.195 which was gradually replaced by B.1.1.28, and the second wave coincides with the emergence of the VOC P.1. This variant probably evolved from a local B.1.1.28 clade in late November and replaced the parental lineage in less than two months. An evolutionary intermediate between B.1.1.28 and P.1 (named P.1-like) was identified, suggesting that the diversity of SARS-CoV-2 variants harboring spike mutations in Manaus could be larger than initially expected and that those variants probably circulated for some time before the emergence and expansion of P.1 (Naveca et al., 2021).

Another study following the circulation of P.1 estimated its emergence to November 2020 preceded by a period of faster molecular evolution. Additionally, virus exchanges between Amazonas and the urban metropolises in Southeast Brazil follow patterns in national air travel mobility, since states reporting P.1 until end-February 2021 received around 100,000 air passengers from Manaus in November. Of the 10 new amino acid mutations in the spike protein (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I) compared to its immediate ancestor (B.1.1.28), molecular selection analyses found evidence that 8 of these 10 mutations are under diversifying positive selection (Faria et al., 2021).

Detecting mutations that are subjected to positive pressure is of paramount importance in order to predict the SARS-CoV-2 pandemic future. By correlating amino acid replacements with expected structural changes, it is possible to anticipate risk of immune evasion with consequent infection recurrence and or vaccine mismatching. Various specific mutations were encountered in different lineages that are potentially associated with selective advantages. RBD and its hACE-2 interacting core, the Receptor Binding Motif (RBM), is of evident importance, since substitutions in this motif were associated with increased receptor binding forces (*e. g.*, N501Y) or immune evasion (*e.g.*, E484K). The E484K mutation seems to be of particular relevance, as its presence shifts the main interaction residue to this site. Molecular dynamic simulation reveals that E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces a higher number of conformational changes than N501Y mutant alone, potentially resulting in an escape mutant (Nelson et al., 2021). Since this site is not involved in interaction with hACE-2 when the original glutamate is in place, its occurrence has been linked to reinfection, convalescent plasma activity abolishment and decreased post-vaccination neutralizing activity (Li et al., 2021; Nelson et al., 2021).

Other mutations likely play important roles by allosteric mechanisms and have been positively selected early during the SARS-CoV-2 pandemic. Almost all Brazilian sequences harbor S:D614G, a hallmark of the ancestral B.1 lineage. Although this mutation is outside the RBD, it is speculated that it abolishes the hydrogen bond between the 614 position in S1 and a threonine residue located at S2 from the neighbour protomer. In consequence, RBD would be locked in its activated "up" position, thus increasing viral infectivity (Ozono et al., 2021). Therefore, the establishment of this mutation in Brazilian sequences seems to be related to both a founder effect based on the importation of primarily G614 variants to Brazil and an evolutionary advantage in comparison with D614.

The precise forces that drive the appearance of complex mutational signatures characteristic of different lineages over short time periods remain largely unknown. Under specific circumstances, the combination of prolonged viral shedding with high selective pressure could lead to major evolutionary leaps. Critically ill and immunosuppressed patients chronically infected with SARS-CoV-2 and treated with convalescent plasma have been linked to viral breakthroughs caused by mutant viruses (Kemp et al., 2021). Whichever phenomena allow for faster viral evolution, they probably have to permit multiple substitutions to occur almost simultaneously, since our molecular clock analysis shows constant and relatively slow rates of mutation accumulation, except for VOC viruses.

Although the proximal origins of the most important VOCs remain to be determined, some conclusions about their nature could be already drawn. First, the mutations shared between P.1 and B.1.1.351 seem to be associated with a rapid increase in cases even in locations where previous attack rates were thought to be very high (Buss et al., 2020). Lineage P.1, which emerged from the Brazilian state of Amazonas between November and December 2020, has accumulated a high number of non-synonymous mutations and is now dispersed across novel Brazilian regions, representing one of the most frequent lineages up to February 2021 (Faria et al., 2021; Naveca et al., 2021). Second, the fact that the set of mutations shared by P.1, B.1.1.7 and B.1.351 seem to have arisen independently, as previously demonstrated with emergence of E484K in others Brazilian lineages (P.2, B.1.1.28 and B.1.1.33), is suggestive of convergent molecular evolution (Faria et al., 2021; Ferrareze et al., 2021).

Our study shows that the Brazilian territory was affected by at least 59 different lineages during the first year COVID-19 pandemic. This is not completely unexpected, considering the size of the country and its touristic and economical relevance. South Africa, the original source of B.1.351 lineage, similarly had multiple and diverse viral introductions. Of note, a recent genomic study detected 42 different circulating lineages in the country, between the first epidemic wave (March) and mid-September, 2020. Moreover, the three main lineages (B.1.1.54, B.1.1.56 and C.1), which represented the majority of cases in the first wave, were responsible for ~42% total of the infections by the end of 2020, since B.1.351 had emerged in an explosive fashion in mid-October (Tegally et al., 2021b). This later lineage is of major concern and South Africa is, up to February 2021, the leader country in COVID-19 related deaths in Africa (Johns Hopkins Coronavirus Resource Center, 2021; Li et al., 2020). In Brazil, the prevalence of lineages B.1.1.28 and its derivatives P.1 and P.2 have been representing progressively more cases of the sequenced genomes at the time of writing. In March 2020, B.1.1.28 was one of the 28 circulating lineages present in the country, with more than 30% of the sequenced genomes. At its side, B.1.1.33 was the identified lineage for approximately 26% of analyzed genomes. Ten months later (January, 2021), the VOC P.1 appeared as the prevalent one among 12 different lineages, reaching more than 45% of the sequenced cases. Next, followed the P.2 lineage (~27%). At that point, B.1.1.28 and B.1.1.33, which achieved 85.5% of the sequenced genomes together in June, matched for, in January, less than 10% of the cases. Therefore, in both countries, despite potentially different initial founding effects that could have led to diverse lineage dissemination patterns, eventually complex VOCs harboring advantageous mutations were selected for viral spread, indicating an increased evolutionary fitness for these viruses.

Our phylogeographic reconstruction demonstrated the widespread dissemination of multiple SARS-CoV-2 lineages inside Brazil and the major role of intrastate diffusion of the most prevalent lineages. The evolutionary rate estimates for clade-specific MCC trees (especially clades 3, 5, and 6) were significantly smaller than previous findings (8 to 9×10^{-4} subst/site/year) (Rambaut, 2020; Su et al., 2020). These differences have contributed to an older dating of the MRCA, since the accumulation of more mutations is expected to occur in a wider timeframe. While these older MRCA datings are probably not realistic in the context of the COVID-19 pandemic, they highlight issues with the data collection process. Potential explanations for this behavior are: related samples having the same age (phylo-temporal clustering), among-lineage rate variation and non-random sampling (Tong et al., 2018).

The dispersion model used is well suited for spread over land, in opposition to long distances traveled by plane, a common practice in the

Brazilian territory due to its continental size. The major consequence is that inferences of the nodes locations near to the root of the tree must be analyzed very cautiously. For example, in Fig. 7C the green ellipses in the northwest-southeast direction are a consequence of sparse sampling and nodes representing travel events between the North and Southeast. Therefore, the strong uncertainty in these estimates (large ellipses) reflects limitations of both the data and the employed model. Despite its limitations, more local dispersal routes represented here closer to terminal nodes were well captured by the phylogeographic model, which was the main objective of this study since there was a greater contribution of intra- and interstate diffusion after the pandemic was established in the country.

Unfortunately, incomplete and erratic sequencing efforts have limited a better SARS-CoV-2 characterization in Brazil, since genomes are not equally distributed in geographical or temporal scales due to episodic sampling efforts prompted by resource availability. This is reflected by a very small fraction of SARS-CoV-2 cases being sequenced. Unequal temporal distribution implies that some of the conclusions are disproportionately affected by events in heavily (March-May 2020) and poorly (June 2020 onward) sampled periods. Therefore, different lineage distributions could be an artefact of distinct sequencing coverage among states and across different time frames. Importantly, as Brazil is currently the new epicenter of the pandemic and identified variants of concern and interest, the country initiatives doubled the quantity of genomes deposited in GISAID (5,468 in April 8th vs 2,751 used in this analysis until February 16). Only 420 of these were collected between mid-February and early-April 2021, showing important delays between sample collection and sequencing or sample storage for long periods before sequencing and submission (*e. g.,* waiting for reagents and/or research investments). Therefore, a significant amount of sequences from 2020 were deposited lately and were not included in this analysis. By mid-July, 23,351 genomes from Brazil are available, of which 17,894 (76.63%) have collection dates in 2021 and almost almost half are from the state of São Paulo (https://www.gisaid.org/). These data demonstrate the absence of countrywide temporally and spatially representative datasets even with increased investment and sequencing capacity.

Notably, the majority of the sequences analyzed here come from the Southeastern region of Brazil. While this is in fact an economic and travel hub for the country, accounting for >70% of the international passengers arriving in Brazil in the beginning of the pandemic (Candido et al., 2020b), inferences regarding this region can be inflated in relation to undersampled regions. However, this is, to the best of our knowledge, the first attempt to characterize sequencing efforts, SARS-CoV-2 mutations, phylogenetics, phylogeography and phylodynamics in the entire Brazil after the study that characterized the first epidemic wave in Brazil using 490 representative genomes (Candido et al., 2020a). In the near future, it will be important to describe and track the spread of the P.1 and P.2 lineages, which have already shown to be replacing the other lineages identified in Brazil (https://outbreak.info/situation-reports).

5. Conclusions

In summary, by systematic analysis of viral genomes distributed across Brazil over time, we were able to confirm the early introductions of multiple lineages, its rapid diversification to constitute new lineages, probable convergent evolution of important mutations (*e. g.,* E484K, N501Y), and the emergence of P.1, arguably one of the most potentially concerning lineages identified worldwide up to February 2021. The occurrence of this lineage and the emergence of novel variants (*e. g.*, B.1.617.2) could jeopardize the efficacy of vaccines and immunotherapies and may lead the health care system to overload. We concluded that enhanced genomic surveillance is, therefore, of paramount importance and should be extended as soon as possible as a means to better inform policy makers and enable precise evidence-based decisions to fight the COVID-19 pandemic.

Supplementary files

Supplementary File 1. GISAID acknowledgement table of worldwide SARS-CoV-2 genomes used in this study.

Table S1. Comparison of genomes sampled from Brazil and confirmed cases per month from March 2020 to mid-February 2021.

Table S2. Comparison of genomes sampled from each Brazilian state and confirmed cases from March 2020 to mid-February 2021. Data are ordered from the higher sequencing rate (genomes per case) to lower.

Table S3. Number of genomes and lineages sampled from each Brazilian state from March 2020 to mid-February 2021. Data are ordered from the higher number of genomes to lower.

Table S4. First and last detection of SARS-CoV-2 lineages represented by > 5 Brazilian genomes, including information on state of detection and minimum time of spread according to available data.

Fig. S1. Distribution of the 10 most prevalent SARS-CoV-2 lineages across all Brazilian states considering the absolute number of sequences per state.

Fig. S2. Time between first and last detection of the 18 lineages observed in more than 5 Brazilian genomes.

Fig. S3. Time-resolved Maximum likelihood phylogenetics tree dropping sequences from other countries and highlighting the most frequent Brazilian lineages.

Fig. S4. Zoom-in on clades corresponding to the two minor clades from B.1.1.74 and B.1.1.28 Brazilian lineages not represented in Fig. 6. (A) Time-resolved ML tree of Brazilian sequences colored by PANGO lineages. Letters around clades are augmented in respective figures. (B) Clade 4.2 is represented by 27 sequences (B.1.1.74 lineage). (C) Clade 5.3 harbors 162 sequences from B.1.1.28 and B.1.1.94. In (B) and (C), states belonging to each specific Brazilian region are colored using similar colors (Centre-West: yellow; North: red; Northeast: purple and pink; South: blue; Southeast: green).

Fig. S5. Maximum clade credibility trees of major Brazilian clades with sufficient temporal signal. (A) Clade 3 (lineage B.1.1.33), (B) Clade 4 (B.1.1.74), (C) Clade 5 (B.1.1.28 and P.1) and (D) Clade 6 (B.1.1.28). In (A), (B), (C), and (D), states belonging to each specific Brazilian region are colored using similar colors (Centre-West: yellow; North: red; Northeast: purple and pink; South: blue; Southeast: green).

Video S1. Animation of the phylogeographic reconstruction of Clade 3. Node polygons and lines represent posterior probabilities (orange to red scale). Colors on the map represent the five Brazilian regions (South, Southeast, Centre-West, Northeast and North).

Video S2. Animation of the phylogeographic reconstruction of Clade 4. Node polygons and lines represent posterior probabilities (orange to red scale). Colors on the map represent the five Brazilian regions (South, Southeast, Centre-West, Northeast and North).

Video S3. Animation of the phylogeographic reconstruction of Clade 5. Node polygons and lines represent posterior probabilities (orange to red scale). Colors on the map represent the five Brazilian regions (South, Southeast, Centre-West, Northeast and North).

Video S4. Animation of the phylogeographic reconstruction of Clade 6. Node polygons and lines represent posterior probabilities (orange to red scale). Colors on the map represent the five Brazilian regions (South, Southeast, Centre-West, Northeast and North).

Funding

This study was supported by the Brazilian National Research Council (CNPq) and Amazon Web Services (AWS) joint grant 032/2019 (Grant No. 440084/2020-2). Scholarships and Fellowships were supplied by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA). The funders had no role in the study design, data generation and analysis, decision to publish or the preparation of the manuscript.

Availability of data and material

A full table acknowledging the authors and corresponding labs submitting sequencing data used in this study can be found in Supplementary File 1. Additional information used and/or analysed during the current study are available from the corresponding author on reasonable request.

CRediT authorship contribution statement

Vinícius Bonetti Franceschi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Patrícia Aline Gröhs Ferrareze:** Methodology, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Ricardo Ariel Zimerman:** Investigation, Writing – original draft, Writing – review & editing. **Gabriela Bettella Cybis:** Methodology, Validation, Formal analysis, Investigation, Writing – review & editing, Supervision. **Claudia Elizabeth Thompson:** Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgment

We thank the administrators of the GISAID database and research groups across the world (especially Brazilians) for supporting the rapid and transparent sharing of genomic data during the COVID-19 pandemic. We also thank the Mayor's Office, Health Department and São Camilo Hospital (Esteio, RS, Brazil), Leonardo Duarte Pascoal and Ana Regina Boll for their work in combating COVID-19 and for supporting the work developed by our research group.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.virusres.2021.198532.

References

Ayres, D.L., Darling, A., Zwickl, D.J., Beerli, P., Holder, M.T., Lewis, P.O., Huelsenbeck, J.P., Ronquist, F., Swofford, D.L., Cummings, M.P., Rambaut, A., Suchard, M.A., 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol. 61, 170–173. https://doi.org/10.1093/sysbio/syr100 https://doi.org/.

Bartolini, B., Rueca, M., Gruber, C.E.M., Messina, F., Carletti, F., Giombini, E., Lalle, E., Bordi, L., Matusali, G., Colavita, F., Castilletti, C., Vairo, F., Ippolito, G., Capobianchi, M.R., Caro, A.D., 2020. SARS-CoV-2 Phylogenetic analysis, Lazio Region, Italy, February–March 2020. Emerg. Infect. Dis. 26 https://doi.org/10.3201/eid2608.201525 https://doi.org/.

Baum, A., Fulton, B.O., Wloga, E., Copin, R., Pascal, K.E., Russo, V., Giordano, S., Lanza, K., Negron, N., Ni, M., Wei, Y., Atwal, G.S., Murphy, A.J., Stahl, N., Yancopoulos, G.D., Kyratsous, C.A., 2020. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. Science 369, 1014–1018. https://doi.org/10.1126/science.abd0831 https://doi.org/.

Bielejec, F., Baele, G., Vrancken, B., Suchard, M.A., Rambaut, A., Lemey, P., 2016. SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. Mol. Biol. Evol. 33, 2167–2169. https://doi.org/10.1093/molbev/msw082 https://doi.org/.

Buss, L.F., Prete, C.A., Abrahim, C.M.M., Mendrone, A., Salomon, T., Almeida-Neto, C.de, França, R.F.O., Belotti, M.C., Carvalho, M.P.S.S., Costa, A.G., Crispim, M.A.E., Ferreira, S.C., Fraiji, N.A., Gurzenda, S., Whittaker, C., Kamaura, L.T., Takecian, P.L., Peixoto, P.da S., Oikawa, M.K., Nishiya, A.S., Rocha, V., Salles, N.A., Santos, A.A.de S., Silva, M.A.da, Custer, B., Parag, K.V., Barral-Netto, M., Kraemer, M.U.G., Pereira, R.H.M., Pybus, O.G., Busch, M.P., Castro, M.C., Dye, C., Nascimento, V.H., Faria, N.R., Sabino, E.C., 2020. Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. Science. https://doi.org/10.1126/science.abe9728 https://doi.org/.

Candido, D., Claro, I.M., Jesus, J.G.de, Souza, W.M., Moreira, F.R.R., Dellicour, S., Mellan, T.A., Plessis, L.du, Pereira, R.H.M., Sales, F.C.S., Manuli, E.R., Thézé, J., Almeida, L., Menezes, M.T., Voloch, C.M., Fumagalli, M.J., Coletti, T.M., Silva, C.A.M.da, Ramundo, M.S., Amorim, M.R., Hoeltgebaum, H.H., Mishra, S., Gill, M.S., Carvalho, L.M., Buss, L.F., Prete, C.A., Ashworth, J., Nakaya, H.I., Peixoto, P.S., Brady, O.J., Nicholls, S.M., Tanuri, A., Rossi, Á.D., Braga, C.K.V., Gerber, A.L., Guimarães, A.P.de C., Gaburo, N., Alencar, C.S., Ferreira, A.C.S., Lima, C.X., Levi, J. E., Granato, C., Ferreira, G.M., Francisco, R.S., Granja, F., Garcia, M.T., Moretti, M. L., Perroud, M.W., Castiñeiras, T.M.P.P., Lazari, C.S., Hill, S.C., Santos, A.A.de S., Simeoni, C.L., Forato, J., Sposito, A.C., Schreiber, A.Z., Santos, M.N.N., Sá, C.Z.de, Souza, R.P., Resende-Moreira, L.C., Teixeira, M.M., Hubner, J., Leme, P.A.F., Moreira, R.G., Nogueira, M.L., , Brazil-UK Centre for Arbovirus Discovery, Diagnosis, Genomics and Epidemiology (CADDE) Genomic Network, Ferguson, N.M., Costa, S. F., Proenca-Modena, J.L., Vasconcelos, A.T.R., Bhatt, S., Lemey, P., Wu, C.H., Rambaut, A., Loman, N.J., Aguiar, R.S., Pybus, O.G., Sabino, E.C., Faria, N.R., 2020a. Evolution and epidemic spread of SARS-CoV-2 in Brazil. Science 369, 1255–1260. https://doi.org/10.1126/science.abd2161 https://doi.org/.

Candido, D., Watts, A., Abade, L., Kraemer, M.U.G., Pybus, O.G., Croda, J., de Oliveira, W., Khan, K., Sabino, E.C., Faria, N.R., 2020b. Routes for COVID-19 importation in Brazil. J. Travel Med. 27 https://doi.org/10.1093/jtm/taaa042 https://doi.org/.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6, 80–92. https://doi.org/10.4161/fly.19695 https://doi.org/.

Claro, I.M., Sales, F.C., da, S., Ramundo, M.S., Candido, D.S., Silva, C.A.M., Jesus, J.G.de, Manuli, E.R., Oliveira, C.M.de, Scarpelli, L., Campana, G., Pybus, O.G., Sabino, E.C., Faria, N.R., Levi, J.E., 2021. Local transmission of SARS-CoV-2 lineage B.1.1.7, Brazil, December 2020. Number 3—March 2021 Emerg. Infect. Dis. J. Volume 27. https://doi.org/10.3201/eid2703.210038. CDChttps://doi.org/.

Cota, W., 2020. Monitoring the number of COVID-19 cases and deaths in Brazil at municipal and federative units level (preprint). https://doi.org/ 10.1590/SciELOPreprints.362.

da Silva Filipe, A., Shepherd, J.G., Williams, T., Hughes, J., Aranday-Cortes, E., Asamaphan, P., Ashraf, S., Balcazar, C., Brunker, K., Campbell, A., Carmichael, S., Davis, C., Dewar, R., Gallagher, M.D., Gunson, R., Hill, V., Ho, A., Jackson, B., James, E., Jesudason, N., Johnson, N., McWilliam Leitch, E.C., Li, K., MacLean, A., Mair, D., McAllister, D.A., McCrone, J.T., McDonald, S.E., McHugh, M.P., Morris, A. K., Nichols, J., Niebel, M., Nomikou, K., Orton, R.J., O'Toole, Á., Palmarini, M., Parcell, B.J., Parr, Y.A., Rambaut, A., Rooke, S., Shaaban, S., Shah, R., Singer, J.B., Smollett, K., Starinskij, I., Tong, L., Sreenu, V.B., Wastnedge, E., Holden, M.T.G., Robertson, D.L., Templeton, K., Thomson, E.C., 2021. Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. Nat. Microbiol. 6, 112–122. https://doi.org/10.1038/s41564-020-00838-z https://doi.org/.

Dellicour, S., Rose, R., Faria, N.R., Lemey, P., Pybus, O.G., 2016. SERAPHIM: studying environmental rasters and phylogenetically informed movements. Bioinformatics 32, 3204–3206. https://doi.org/10.1093/bioinformatics/btw384 https://doi.org/.

Deng, X., Gu, W., Federman, S., Plessis, L.du, Pybus, O.G., Faria, N., Wang, C., Yu, G., Bushnell, B., Pan, C.Y., Guevara, H., Sotomayor-Gonzalez, A., Zorn, K., Gopez, A., Servellita, V., Hsu, E., Miller, S., Bedford, T., Greninger, A.L., Roychoudhury, P., Starita, L.M., Famulare, M., Chu, H.Y., Shendure, J., Jerome, K.R., Anderson, C., Gangavarapu, K., Zeller, M., Spencer, E., Andersen, K.G., MacCannell, D., Paden, C. R., Li, Y., Zhang, J., Tong, S., Armstrong, G., Morrow, S., Willis, M., Matyas, B.T., Mase, S., Kasirye, O., Park, M., Masinde, G., Chan, C., Yu, A.T., Chai, S.J., Villarino, E., Bonin, B., Wadford, D.A., Chiu, C.Y., 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science. https://doi.org/10.1126/science.abb9263 https://doi.org/.

du Plessis, L., 2020. laduplessis/SARS-CoV-2_Guangdong_genomic_epidemiology: Initial release. Zenodo. https://doi.org/10.5281/zenodo.3922606 https://doi.org/.

COVID-19 Genomics UK (COG-UK) Consortium du Plessis, L., McCrone, J.T., Zarebski, A. E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T.R., Faria, N.R., Jackson, B., Loman, N.J., O'Toole, Á., Nicholls, S.M., Parag, K.V., Scher, E., Vasylyeva, T.I., Volz, E.M., Watts, A., Bogoch, I.I., Khan, K., Aanensen, D.M., Kraemer, M.U.G., Rambaut, A., Pybus, O.G., COVID-19 Genomics UK (COG-UK) Consortium, 2021. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. Science. https://doi.org/10.1126/science.abf2946. https://doi.org/.

Faria, N., Mellan, T.A., Whittaker, C., Claro, I.M., Candido, D.da S., Mishra, S., Crispim, M.A.E., Sales, F.C.S., Hawryluk, I., McCrone, J.T., Hulswit, R.J.G., Franco, L.A.M., Ramundo, M.S., Jesus, J.G.de, Andrade, P.S., Coletti, T.M., Ferreira, G.M., Silva, C.A.M., Manuli, E.R., Pereira, R.H.M., Peixoto, P.S., Kraemer, M.U.G., Gaburo, N., Camilo, C.da C., Hoeltgebaum, H., Souza, W.M., Rocha, E.C., Souza, L.M.de, Pinho, M.C.de, Araujo, L.J.T., Malta, F.S.V., Lima, A.B. de, Silva, J.do P., Zauli, D.A.G., Ferreira, A.C.de S., Schnekenberg, R.P., Laydon, D.J., Walker, P.G.T., Schlüter, H.M., Santos, A.L.P.dos, Vidal, M.S., Caro, V.S.D., Filho, R. M.F., Santos, H.M.dos, Aguiar, R.S., Proença-Modena, J.L., Nelson, B., Hay, J.A., Monod, M., Miscouridou, X., Coupland, H., Sonabend, R., Vollmer, M., Gandy, A., Prete, C.A., Nascimento, V.H., Suchard, M.A., Bowden, T.A., Pond, S.L.K., Wu, C.H., Ratmann, O., Ferguson, N.M., Dye, C., Loman, N.J., Lemey, P., Rambaut, A., Fraiji, N.A., Carvalho, M.do P.S.S., Pybus, O.G., Flaxman, S., Bhatt, S., Sabino, E.C., 2021. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. https://doi.org/10.1126/science.abh2644 https://doi.org/.

Ferrareze, P.A.G., Franceschi, V.B., Mayer, A.de M., Caldana, G.D., Zimerman, R.A., Thompson, C.E., 2021. E484K as an innovative phylogenetic event for viral evolution: genomic analysis of the E484K spike mutation in SARS-CoV-2 lineages from Brazil. Infect. Genet. Evol. 93, 104941 https://doi.org/10.1016/j.meegid.2021.104941 https://doi.org/.

Ferreira, M.A.R., Suchard, M.A., 2008. Bayesian analysis of elapsed times in continuous-time Markov chains. Can. J. Stat. 36, 355–368. https://doi.org/10.1002/cjs.5550360302 https://doi.org/.

Franceschi, V.B., Caldana, G.D., de Menezes Mayer, A., Cybis, G.B., Neves, C.A.M., Ferrareze, P.A.G., Demoliner, M., de Almeida, P.R., Gularte, J.S., Hansen, A.W., Weber, M.N., Fleck, J.D., Zimerman, R.A., Kmetzsch, L., Spilki, F.R., Thompson, C.E., 2021. Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil. BMC Genom. 22, 371. https://doi.org/10.1186/s12864-021-07708-w https://doi.org/.

Francisco Jr,, R., da, S., Benites, L.F., Lamarca, A.P., de Almeida, L.G.P., Hansen, A.W., Gularte, J.S., Demoliner, M., Gerber, A.L., de C Guimarães, A.P., Antunes, A.K.E., Heldt, F.H., Mallmann, L., Hermann, B., Ziulkoski, A.L., Goes, V., Schallenberger, K., Fillipi, M., Pereira, F., Weber, M.N., de Almeida, P.R., Fleck, J.D., Vasconcelos, A.T. R., Spilki, F.R., 2021. Pervasive transmission of E484K and emergence of VUI-NP13L with evidence of SARS-CoV-2 co-infection events by two different lineages in Rio Grande do Sul, Brazil. Virus Res 296, 198345. https://doi.org/10.1016/j.virusres.2021.198345 https://doi.org/.

Furuse, Y., 2021. Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. Int. J. Infect. Dis. 103, 305–307. https://doi.org/10.1016/j.ijid.2020.12.034 https://doi.org/.

Garrison, E., Marth, G., 2012. Haplotype-Based Variant Detection From Short-Read Sequencing. ArXiv. http://arxiv.org/abs/1207.3907.

Gill, M.S., Lemey, P., Faria, N.R., Rambaut, A., Shapiro, B., Suchard, M.A., 2013. Improving bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol. Biol. Evol. 30, 713–724. https://doi.org/10.1093/molbev/mss265 https://doi.org/.

Greaney, A.J., Starr, T.N., Gilchuk, P., Zost, S.J., Binshtein, E., Loes, A.N., Hilton, S.K., Huddleston, J., Eguia, R., Crawford, K.H.D., Dingens, A.S., Nargi, R.S., Sutton, R.E., Suryadevara, N., Rothlauf, P.W., Liu, Z., Whelan, S.P.J., Carnahan, R.H., Crowe, J.E., Bloom, J.D., 2020. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe. https://doi.org/10.1016/j.chom.2020.11.007 https://doi.org/.

Griffiths, R.C., Tavare, S., 1994. Sampling theory for neutral alleles in a varying environment. Philos. Trans. Biol. Sci. 344, 403–410.

Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., Neher, R.A., 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34, 4121–4123. https://doi.org/10.1093/bioinformatics/bty407 https://doi.org/.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J., Cao, B., 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395, 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5 https://doi.org/.

Johns Hopkins Coronavirus Resource Center, 2021. COVID-19 Map. Johns Hopkins University [WWW Document]. https://coronavirus.jhu.edu/map.html. accessed 02.22.21.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589. https://doi.org/10.1038/nmeth.4285 https://doi.org/.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010 https://doi.org/.

Kemp, S.A., Collier, D.A., Datir, R.P., Ferreira, I.A.T.M., Gayed, S., Jahun, A., Hosmillo, M., Rees-Spear, C., Mlcochova, P., Lumb, I.U., Roberts, D.J., Chandra, A., Temperton, N., Sharrocks, K., Blane, E., Modis, Y., Leigh, K., Briggs, J., van Gils, M., Smith, K.G.C., Bradley, J.R., Smith, C., Doffinger, R., Ceron-Gutierrez, L., Barcenas-Morales, G., Pollock, D.D., Goldstein, R.A., Smielewska, A., Skittrall, J.P., Gouliouris, T., Goodfellow, I.G., Gkrania-Klotsas, E., Illingworth, C.J.R., McCoy, L.E., Gupta, R.K., 2021. SARS-CoV-2 evolution during treatment of chronic infection. Nature 1–10. https://doi.org/10.1038/s41586-021-03291-y https://doi.org/.

Lemey, P., Rambaut, A., Welch, J.J., Suchard, M.A., 2010. Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. 27, 1877–1885. https://doi.org/10.1093/molbev/msq067 https://doi.org/.

Li, Q., Nie, J., Wu, J., Zhang, Li, Ding, R., Wang, H., Zhang, Y., Li, T., Liu, S., Zhang, M., Zhao, C., Liu, H., Nie, L., Qin, H., Wang, M., Lu, Q., Xiaoyu, L., Liu, J., Liang, H., Shi, Y., Shen, Y., Xie, L., Zhang, Linqi, Qu, X., Xu, W., Huang, W., Wang, Y., 2021. No higher infectivity but immune escape of SARS-CoV-2 501Y.V2 variants. Cell. https://doi.org/10.1016/j.cell.2021.02.042 https://doi.org/.

Li, Q., Wu, J., Nie, J., Zhang, Li, Hao, H., Liu, S., Zhao, C., Zhang, Q., Liu, H., Nie, L., Qin, H., Wang, M., Lu, Q., Li, Xiaoyu, Sun, Q., Liu, J., Zhang, Linqi, Li, Xuguang, Huang, W., Wang, Y., 2020. The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. Cell 182, 1284–1294.e9. https://doi.org/10.1016/j.cell.2020.07.012.

Lu, J., du Plessis, L., Liu, Z., Hill, V., Kang, M., Lin, H., Sun, J., François, S., Kraemer, M. U.G., Faria, N.R., McCrone, J.T., Peng, J., Xiong, Q., Yuan, R., Zeng, L., Zhou, P., Liang, C., Yi, L., Liu, J., Xiao, J., Hu, J., Liu, T., Ma, W., Li, W., Su, J., Zheng, H., Peng, B., Fang, S., Su, W., Li, K., Sun, R., Bai, R., Tang, X., Liang, M., Quick, J., Song, T., Rambaut, A., Loman, N., Raghwani, J., Pybus, O.G., Ke, C., 2020. Genomic epidemiology of SARS-CoV-2 in Guangdong Province. China. Cell 181, 997–1003.e9. https://doi.org/10.1016/j.cell.2020.04.023 https://doi.org/.

Martin, D.P., Weaver, S., Tegally, H., San, E.J., Shank, S.D., Wilkinson, E., Giandhari, J., Naidoo, S., Pillay, Y., Singh, L., Lessells, R.J., , NGS-SA, COVID-19 Genomics UK (COG-UK), Gupta, R.K., Wertheim, J.O., Nekturenko, A., Murrell, B., Harkins, G.W., Lemey, P., MacLean, O.A., Robertson, D.L., Oliveira, T.de, Pond, S.L.K., 2021. The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. medRxiv. https://doi.org/10.1101/2021.02.23.21252268, 2021.02.23.21252268https://doi.org/.

Maurano, M.T., Ramaswami, S., Zappile, P., Dimartino, D., Boytard, L., Ribeiro-dos-Santos, A.M., Vulpescu, N.A., Westby, G., Shen, G., Feng, X., Hogan, M.S., Ragonnet-Cronin, M., Geidelberg, L., Marier, C., Meyn, P., Zhang, Y., Cadley, J.A., Ordoñez, R., Luther, R., Huang, E., Guzman, E., Arguelles-Grande, C., Argyropoulos, K.V., Black, M., Serrano, A., Call, M.E., Kim, M.J., Belovarac, B., Gindin, T., Lytle, A., Pinnell, J., Vougiouklakis, T., Chen, J., Lin, L.H., Rapkiewicz, A., Raabe, V., Samanovic, M.I., Jour, G., Osman, I., Aguero-Rosenfeld, M., Mulligan, M.J., Volz, E. M., Cotzia, P., Snuderl, M., Heguy, A., 2020. Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City Region. Genome Res. https://doi.org/10.1101/gr.266676.120 gr.266676.120. https://doi.org/.

Mir, D., Rego, N., Resende, P.C., López-Tort, F., Fernandez-Calero, T., Noya, V., Brandes, M., Possi, T., Arleo, M., Reyes, N., Victoria, M., Lizasoain, A., Castells, M., Maya, L., Salvo, M., Gregianini, T.S., Rosa, M.T.M.da, Martins, L.G., Alonso, C., Vega, Y., Salazar, C., Ferrés, I., Smircich, P., Sotelo, J., Fort, R.S., Mathó, C., Arantes, I., Appolinario, L., Mendonça, A.C., Benitez-Galeano, M.J., Graña, M., Simoes, C., Motta, F., Siqueira, M.M., Bello, G., Colina, R., Spangenberg, L., 2021. Recurrent dissemination of SARS-CoV-2 through the Uruguayan-Brazilian border. medRxiv. https://doi.org/10.1101/2021.01.06.20249026, 2021.01.06.20249026. https://doi.org/.

Mullen, J.L., Tsueng, G., Latif, A.A., Alkuzweny, M., Cano, M., Haag, E., Zhou, J., Zeller, M., Matteson, N., Andersen, K.G., Wu, C., Su, A.I., Gangavarapu, K., Hughes, L.D., Center for viral systems biology, 2021. B.1.617.2 Lineage Report [WWW Document]. outbreak.info. URL https://outbreak.info/situation-reports?pango=B.1.617.2&loc=IND&loc=GBR&loc=USA&selected=IND accessed 7.13.21.

Naveca, F., Nascimento, V., Souza, V., Corado, A., Nascimento, F., Silva, G., Costa, Á., Duarte, D., Pessoa, K., Mejía, M., Brandão, M., Jesus, M., Gonçalves, L., da Costa, C., Sampaio, V., Barros, D., Silva, M., Tirza, M., Pontes, G., Abdalla, L., Santos, J., Arantes, I., Dezordi, F., Siqueira, M., Wallau, G., Resende, P., Delatorre, E., Graff, T., Bello, G., 2021. COVID-19 epidemic in the Brazilian state of Amazonas was driven by long-term persistence of endemic SARS-CoV-2 lineages and the recent emergence of the new Variant of Concern P.1 [WWW Document]. https://doi.org/ 10.21203/rs.3.rs-275494/v1.

Nelson, G., Buzko, O., Spilman, P., Niazi, K., Rabizadeh, S., Soon-Shiong, P., 2021. Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. bioRxiv. https://doi.org/10.1101/2021.01.13.426558, 2021.01.13.426558https://doi.org/.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300 https://doi.org/.

Nonaka, C.K.V., Franco, M.M., Gräf, T., Mendes, A.V.A., Aguiar, R.S. de, Giovanetti, M., Souza, B.S. de F., 2021. Genomic evidence of a SARS-CoV-2 reinfection case with E484K spike mutation in Brazil. https://doi.org/ 10.20944/preprints202101.0132.v1.

Ozono, S., Zhang, Y., Ode, H., Sano, K., Tan, T.S., Imai, K., Miyoshi, K., Kishigami, S., Ueno, T., Iwatani, Y., Suzuki, T., Tokunaga, K., 2021. SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE_2-binding affinity. Nat. Commun. 12, 848. https://doi.org/10.1038/s41467-021-21118-2 https://doi.org/.

Paiva, M.H.S., Guedes, D.R.D., Docena, C., Bezerra, M.F., Dezordi, F.Z., Machado, L.C., Krokovsky, L., Helvecio, E., da Silva, A.F., Vasconcelos, L.R.S., Rezende, A.M., da Silva, S.J.R., Sales, K.G.da S., de Sá, B.S.L.F., da Cruz, D.L., Cavalcanti, C.E., Neto, A. de M., da Silva, C.T.A., Mendes, R.P.G., da Silva, M.A.L., Gräf, T., Resende, P.C., Bello, G., Barros, M.da S., do Nascimento, W.R.C., Arcoverde, R.M.L., Bezerra, L.C. A., Brandão-Filho, S.P., Ayres, C.F.J., Wallau, G.L., 2020. Multiple introductions followed by ongoing community spread of SARS-CoV-2 at one of the largest metropolitan areas of Northeast Brazil. Viruses 12, 1414. https://doi.org/10.3390/v12121414 https://doi.org/.

Peacock, T.P., Sheppard, C.M., Brown, J.C., Goonawardane, N., Zhou, J., Whiteley, M., Consortium, P.V., Silva, T.I.de, Barclay, W.S., 2021. The SARS-CoV-2 variants associated with infections in India, B.1.617, show enhanced spike cleavage by furin. bioRxiv. https://doi.org/10.1101/2021.05.28.446163, 2021.05.28.446163. https://doi.org/.

Pebesma, E., 2018. Simple features for R: standardized support for spatial vector data. R J. 10, 439–446. https://doi.org/10.32614/RJ-2018-009 https://doi.org/.

Pei, S., Kandula, S., Shaman, J., 2020. Differential effects of intervention timing on COVID-19 spread in the United States. Sci. Adv. 6, eabd6370. https://doi.org/10.1126/sciadv.abd6370 https://doi.org/.

Planas, D., Veyer, D., Baidaliuk, A., Staropoli, I., Guivel-Benhassine, F., Rajah, M.M., Planchais, C., Porrot, F., Robillard, N., Puech, J., Prot, M., Gallais, F., Gantner, P., Velay, A., Le Guen, J., Kassis-Chikhani, N., Edriss, D., Belec, L., Seve, A., Courtellemont, L., Péré, H., Hocqueloux, L., Fafi-Kremer, S., Prazuck, T., Mouquet, H., Bruel, T., Simon-Lorière, E., Rey, F.A., Schwartz, O., 2021. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. Nature 1–7. https://doi.org/10.1038/s41586-021-03777-9 https://doi.org/.

Pybus, O.G., Suchard, M.A., Lemey, P., Bernardin, F.J., Rambaut, A., Crawford, F.W., Gray, R.R., Arinaminpathy, N., Stramer, S.L., Busch, M.P., Delwart, E.L., 2012. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proc. Natl. Acad. Sci. 109, 15066–15071. https://doi.org/10.1073/pnas.1206598109 https://doi.org/.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rambaut, A., 2020. Phylodynamic analysis | 176 genomes | 6 Mar 2020 - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology [WWW Document] Virological. URL. https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356. accessed 2.11.21.

Rambaut, A., Drummond, A.J., Xie, D., Baele, G., Suchard, M.A., 2018. Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst. Biol. 67, 901–904. https://doi.org/10.1093/sysbio/syy032 https://doi.org/.

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2020a. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5, 1403–1407. https://doi.org/10.1038/s41564-020-0770-5 https://doi.org/.

Rambaut, A., Lam, T.T., Max Carvalho,, L., Pybus, O.G., 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2. https://doi.org/10.1093/ve/vew007 https://doi.org/.

Rambaut, A., Loman, N., Pybus, O., Barclay, W., Barrett, J., Carabelli, A., Connor, T., Peacock, T., Robertson, D., Volz, E., 2020b. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations [WWW Document] Virological. URL. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563. accessed 1.4.21.

Resende, P.C., Delatorre, E., Gräf, T., Mir, D., Motta, F.C., Appolinario, L.R., Paixão, A.C.D.da, Mendonça, A.C.da F., Ogrzewalska, M., Caetano, B., Wallau, G.L., Docena, C., Santos, M.C.dos, de Almeida Ferreira, J., Sousa Junior, E.C., Silva, S.P.da, Fernandes, S.B., Vianna, L.A., Souza, L.da C., Ferro, J.F.G., Nardy, V.B., Santos, C.A., Riediger, I., do Carmo Debur, M., Croda, J., Oliveira, W.K., Abreu, A., Bello, G., Siqueira, M.M., 2021. Evolutionary dynamics and dissemination pattern of the SARS-CoV-2 lineage B.1.1.33 during the early pandemic phase in Brazil. Front. Microbiol. 11 https://doi.org/10.3389/fmicb.2020.615280 https://doi.org/.

Ruiu, M.L., 2020. Mismanagement of Covid-19: lessons learned from Italy. J. Risk Res. 23, 1007–1020. https://doi.org/10.1080/13669877.2020.1758755 https://doi.org/.

Sagulenko, P., Puller, V., Neher, R.A., 2018. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. https://doi.org/10.1093/ve/vex042, 4. https://doi.org/.

Seemann, T., Lane, C.R., Sherry, N.L., Duchene, S., Gonçalves da Silva, A., Caly, L., Sait, M., Ballard, S.A., Horan, K., Schultz, M.B., Hoang, T., Easton, M., Dougall, S., Stinear, T.P., Druce, J., Catton, M., Sutton, B., van Diemen, A., Alpren, C., Williamson, D.A., Howden, B.P., 2020. Tracking the COVID-19 pandemic in Australia using genomics. Nat. Commun. 11, 4376. https://doi.org/10.1038/s41467-020-18314-x https://doi.org/.

Shu, Y., McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance 22. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 https://doi.org/.

Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., Tortorici, M.A., Walls, A.C., King, N.P., Veesler, D., Bloom, J.D., 2020. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell 182, 1295–1310. e20. https://doi.org/10.1016/j.cell.2020.08.012.

Su, Y.C.F., Anderson, D.E., Young, B.E., Linster, M., Zhu, F., Jayakumar, J., Zhuang, Y., Kalimuddin, S., Low, J.G.H., Tan, C.W., Chia, W.N., Mak, T.M., Octavia, S., Chavatte, J.M., Lee, R.T.C., Pada, S., Tan, S.Y., Sun, L., Yan, G.Z., Maurer-Stroh, S., Mendenhall, I.H., Leo, Y.S., Lye, D.C., Wang, L.F., Smith, G.J.D., 2020. Discovery and genomic characterization of a 382-nucleotide deletion in ORF7b and ORF8 during the early evolution of SARS-CoV-2. mBio 11. https://doi.org/10.1128/mBio.01610-20 https://doi.org/.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., Rambaut, A., 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. https://doi.org/10.1093/ve/vey016, 4, vey016https://doi.org/.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N., Mlisana, K., von Gottberg, A., Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A.J., Engelbrecht, S., Van Zyl, G., Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G., Hsiao, N., Korsman, S., Davies, M.A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D., Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer, C.K., Sewell, B.T., Lourenço, J., Alcantara, L.C.J., Kosakovsky Pond, S.L., Weaver, S., Martin, D., Lessells, R.J., Bhiman, J.N., Williamson, C., de Oliveira, T., 2021a. Detection of a SARS-CoV-2 variant of concern in South Africa. Nature 592, 438–443. https://doi.org/10.1038/s41586-021-03402-9 https://doi.org/.

Tegally, H., Wilkinson, E., Lessells, R.J., Giandhari, J., Pillay, S., Msomi, N., Mlisana, K., Bhiman, J.N., von Gottberg, A., Walaza, S., Fonseca, V., Allam, M., Ismail, A., Glass, A.J., Engelbrecht, S., Van Zyl, G., Preiser, W., Williamson, C., Petruccione, F., Sigal, A., Gazy, I., Hardie, D., Hsiao, N., Martin, D., York, D., Goedhals, D., San, E.J., Giovanetti, M., Lourenço, J., Alcantara, L.C.J., de Oliveira, T., 2021b. Sixteen novel lineages of SARS-CoV-2 in South Africa. Nat. Med. 1–7. https://doi.org/10.1038/s41591-021-01255-3 https://doi.org/.

Tong, K.J., Duchêne, D.A., Duchêne, S., Geoghegan, J.L., Ho, S.Y.W., 2018. A comparison of methods for estimating substitution rates from ancient DNA sequence data. BMC Evol. Biol. 18, 70. https://doi.org/10.1186/s12862-018-1192-3 https://doi.org/.

Voloch, C.M., Francisco, R.da S., Almeida, L.G.P.de, Cardoso, C.C., Brustolini, O.J., Gerber, A.L., Guimarães, A.P.de C., Mariani, D., Costa, R.M.da, Ferreira, O.C., , Covid19-UFRJ Workgroup, L.W., Frauches, T.S., Mello, C.M.B.de, Leitão, I.de C., Galliez, R.M., Faffe, D.S., Castiñeiras, T.M.P.P., Tanuri, A., Vasconcelos, A.T.R.de, 2021. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. J. Virol. 10 95, e00119–e00121. https://doi.org/10.1128/JVI.00119-21 https://doi.org/.

Weisblum, Y., Schmidt, F., Zhang, F., DaSilva, J., Poston, D., Lorenzi, J.C., Muecksch, F., Rutkowska, M., Hoffmann, H.H., Michailidis, E., Gaebler, C., Agudelo, M., Cho, A., Wang, Z., Gazumyan, A., Cipolla, M., Luchsinger, L., Hillyer, C.D., Caskey, M., Robbiani, D.F., Rice, C.M., Nussenzweig, M.C., Hatziioannou, T., Bieniasz, P.D., 2020. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. eLife 9, e61312. https://doi.org/10.7554/eLife.61312 https://doi.org/.

Wickham, H., 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, 10.1007/978-0-387-98141-3.

World Health Organization, 2020. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 [WWW Document]. URL https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020 accessed 11.10.20.

Pereira, R., Gonçalves, C., De Araujo, P., Carvalho, G., De Arruda, R., Nascimento, I., Da Costa, B., Cavedo, W., Andrade, P., Da Silva, A., Braga, C., Schmertmann, C., Samuel-Rosa, A., Ferreira, D., 2019 Geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil. [WWW Document] URL https://github.com/ipeaGIT/geobr accessed 02.22.21.

Worobey, M., Pekar, J., Larsen, B.B., Nelson, M.I., Hill, V., Joy, J.B., Rambaut, A., Suchard, M.A., Wertheim, J.O., Lemey, P., 2020. The emergence of SARS-CoV-2 in Europe and North America. Science. https://doi.org/10.1126/science.abc8169 https://doi.org/.

Xavier, J., Giovanetti, M., Adelino, T., Fonseca, V., Costa, A.V.B.da, Ribeiro, A.A., Felicio, K.N., Duarte, C.G., Silva, M.V.F., Salgado, Á., Lima, M.T., Jesus, R.de, Fabri, A., Zoboli, C.F.S., Santos, T.G.S., Iani, F., Ciccozzi, M., Filippis, A.M.B.de, Siqueira, M.A.M.T.de, Abreu, A.L.de, Azevedo, V.de, Ramalho, D.B., Albuquerque, C.F.C.de, Oliveira, T.de, Holmes, E.C., Lourenço, J., Alcantara, L.C.J., Oliveira, M.A.A., 2020. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. Emerg. Microbes Infect. 9, 1824–1834. https://doi.org/10.1080/22221751.2020.1803146 https://doi.org/.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.Y., 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36. https://doi.org/10.1111/2041-210X.12628 https://doi.org/.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., Shi, Z.L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273. https://doi.org/10.1038/s41586-020-2012-7 https://doi.org/.

5. CAPÍTULO III

O manuscrito que constitui este capítulo, intitulado “Predominance of the SARS-CoV-2 Lineage P.1 and Its Sublineage P.1.2 in Patients from the Metropolitan Region of Porto Alegre, Southern Brazil in March 2021” buscou identificar se a linhagem P.1, detectada no final de 2020 no estado do Amazonas, já representava grande parte dos casos na região metropolitana de Porto Alegre em Março de 2021, além de caracterizar a evolução e padrões de espalhamento geográfico das amostras sequenciadas. Encontra-se publicado na revista Pathogens (https://www.mdpi.com/journal/pathogens), com fator de impacto JCR 2021 = 3,492 e Qualis/CAPES = A2. O manuscrito e os materiais suplementares estão disponíveis na íntegra (*Open access*) no seguinte *link*: https://www.mdpi.com/2076-0817/10/8/988/html. Todas as análises descritas no manuscrito, assim como a sua redação, foram realizadas pelos alunos Vinícius Bonetti Franceschi e Gabriel Dickin Caldana, sendo os demais autores responsáveis por colaborações na escrita ou análises, bem como na sua orientação e obtenção de fomento. Abaixo, todas as páginas do manuscrito publicado foram anexadas para compor o Capítulo III da presente dissertação.

pathogens

MDPI

Article

Predominance of the SARS-CoV-2 Lineage P.1 and Its Sublineage P.1.2 in Patients from the Metropolitan Region of Porto Alegre, Southern Brazil in March 2021

Vinícius Bonetti Franceschi [1,†], Gabriel Dickin Caldana [2,†], Christiano Perin [3], Alexandre Horn [3], Camila Peter [4], Gabriela Bettella Cybis [5], Patrícia Aline Gröhs Ferrareze [2], Liane Nanci Rotta [2], Flávio Adsuara Cadegiani [6], Ricardo Ariel Zimerman [3,*,‡] and Claudia Elizabeth Thompson [1,2,7,*,‡]

1 Graduate Program in Cell and Molecular Biology (PPGBCM), Center of Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre 91501-970, RS, Brazil; vinicius.franceschi@ufrgs.br
2 Graduate Program in Health Sciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre 90050-170, RS, Brazil; gabrield@ufcspa.edu.br (G.D.C.); p.ferrareze@gmail.com (P.A.G.F.); lnrotta@gmail.com (L.N.R.)
3 Department of Infection Control and Prevention, Hospital da Brigada Militar, Porto Alegre 91900-590, RS, Brazil; drchristianoperin@gmail.com (C.P.); amariantehorn@gmail.com (A.H.)
4 Laboratório Exame, Novo Hamburgo 93510-250, RS, Brazil; camilapptr@gmail.com
5 Department of Statistics, Universidade Federal do Rio Grande do Sul, Porto Alegre 91501-970, RS, Brazil; gcybis@gmail.com
6 Corpometria Institute, Brasília 70390-150, DF, Brazil; flavio.cadegiani@unifesp.br
7 Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre, Porto Alegre 90050-170, RS, Brazil
* Correspondence: ricardoarielzimerman@gmail.com (R.A.Z.); cthompson@ufcspa.edu.br or thompson.ufcspa@gmail.com (C.E.T.); Tel.: +55-(51)-3288-3500 (R.A.Z.); +55-(51)-3303-8889 (C.E.T.)
† These authors contributed equally.
‡ These authors share the senior authorship.

check for updates

Citation: Franceschi, V.B.; Caldana, G.D.; Perin, C.; Horn, A.; Peter, C.; Cybis, G.B.; Ferrareze, P.A.G.; Rotta, L.N.; Cadegiani, F.A.; Zimerman, R.A.; et al. Predominance of the SARS-CoV-2 Lineage P.1 and Its Sublineage P.1.2 in Patients from the Metropolitan Region of Porto Alegre, Southern Brazil in March 2021. *Pathogens* **2021**, *10*, 988. https://doi.org/10.3390/pathogens10080988

Academic Editor: Tomomi Takano

Received: 29 June 2021
Accepted: 29 July 2021
Published: 5 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: Almost a year after the COVID-19 pandemic had begun, new lineages (B.1.1.7, B.1.351, P.1, and B.1.617.2) associated with enhanced transmissibility, immunity evasion, and mortality were identified in the United Kingdom, South Africa, and Brazil. The previous most prevalent lineages in the state of Rio Grande do Sul (RS, Southern Brazil), B.1.1.28 and B.1.1.33, were rapidly replaced by P.1 and P.2, two B.1.1.28-derived lineages harboring the E484K mutation. To perform a genomic characterization from the metropolitan region of Porto Alegre, we sequenced viral samples to: (i) identify the prevalence of SARS-CoV-2 lineages in the region, the state, and bordering countries/regions; (ii) characterize the mutation spectra; (iii) hypothesize viral dispersal routes by using phylogenetic and phylogeographic approaches. We found that 96.4% of the samples belonged to the P.1 lineage and approximately 20% of them were assigned as the novel P.1.2, a P.1-derived sublineage harboring signature substitutions recently described in other Brazilian states and foreign countries. Moreover, sequences from this study were allocated in distinct branches of the P.1 phylogeny, suggesting multiple introductions in RS and placing this state as a potential diffusion core of P.1-derived clades and the emergence of P.1.2. It is uncertain whether the emergence of P.1.2 and other P.1 clades is related to clinical or epidemiological consequences. However, the clear signs of molecular diversity from the recently introduced P.1 warrant further genomic surveillance.

Keywords: COVID-19; severe acute respiratory syndrome coronavirus 2; infectious diseases; high-throughput nucleotide sequencing; molecular evolution; molecular epidemiology; phylogeny

1. Introduction

After its initial emergence in Wuhan (China) in late 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spread rapidly around the world leading to the

COVID-19 pandemic officially recognized in March 2020 [1]. As of May 17, 2021, >163 million cases and >3.3 million deaths have been confirmed. In Brazil, the third most affected country by COVID-19, >15.6 million cases and >435,000 deaths have been reported. From a virological standpoint, this could be related to the continental magnitude of Brazil, leading to multiple viral introductions [2] and the recent emergence of a novel variant of concern (VOC) presenting enhanced infectiousness.

Rio Grande do Sul (RS) is the southernmost state in Brazil. It is bordered southerly by Uruguay, westerly by Argentina, and northerly by the state of Santa Catarina, Brazil. With an estimated population of 11.5 million inhabitants and 39.79 inhabitants per square kilometer, RS is the 6th most populous and the 13th most densely populated state in the country [3]. Since 2017, the Brazilian Institute of Geography and Statistics (IBGE) has divided RS into eight intermediate geographic regions: Porto Alegre, Pelotas, Uruguaiana, Santa Maria, Santa Cruz do Sul/Lajeado, Ijuí, Passo Fundo, and Caxias do Sul [4]. The municipality of Porto Alegre is the state capital, and its metropolitan region comprises 34 municipalities aggregating >4 million inhabitants (~2% of the country's population) and is characterized by intense transit of people. COVID-19 was firstly confirmed in RS on 10 March 2020, in a returning traveler from Italy [5]. The state implemented in May 2020 the "controlled distancing system", which divided the state into 21 regions and 26 areas (R01–R26) and consisted of a flag system establishing restrictions and flexibilizations of non-essential activities based on the weekly occupation of intensive care unit beds and expected deaths. However, due to economic losses, the more amenable "shared management system" allowed mayors to appeal in court and adopt less restrictive flag protocols [6].

Important shifts of COVID-19 epicenters have occurred during 2020, starting with Asia and followed by Europe, North America, and South America. After months of relatively slow evolution, novel VOCs (e.g., B.1.1.7, B.1.351, P.1, and B.1.617.2) harboring a constellation of signature mutations in the spike protein have emerged [7]. More recently, the World Health Organization (WHO) assigned labels for the VOCs based on Greek letters. Therefore, the four VOCs are named, respectively, Alpha, Beta, Gamma, and Delta. These lineages independently arose in the United Kingdom [8], South Africa [9], Brazil [10,11], and India [12,13] and have fueled secondary outbreaks in places where they have emerged, despite previous rates of seroprevalence of up to 75% [14]. The city of Manaus (Amazonas, Brazil), the probable place of origin of the P.1 lineage, faced a major second wave of COVID-19. An explosive resurgence of cases and deaths became evident in mid-December 2020. Since the P.1 variant carries multiple mutations of potential biological significance (especially E484K, K417T, and N501Y in the receptor-binding domain (RBD) of the spike protein): (i) some key substitutions may lead to the immunity evasion; (ii) higher transmissibility when compared with pre-existing lineages has been characterized; (iii) this VOC has been the focus of increased surveillance and deserves being studied in greater detail [15]. After this outbreak, almost all Brazilian states experienced increases in the number of cases, hospitalizations, intensive care unit (ICU) admissions, and deaths, resulting in a reemergence of the public health crisis previously experienced in the first wave of COVID-19 [16].

The diversity of SARS-CoV-2 during the first epidemic wave in Brazil was mainly composed of B.1.1.28 and B.1.1.33 lineages [2,17], although the very low sequencing rate across the country has limited these estimates [17]. However, these previous lineages were rapidly replaced by P.1 and P.2 in late 2020 and early 2021, which are both derived from the common ancestor B.1.1.28 and harbor concerning mutations in the spike protein (e.g., E484K and N501Y) [17,18]. In RS state, the most common lineages identified up to May 2021 are still B.1.1.33 (n = 290) and B.1.1.28 (n = 238) [19,20]. Nevertheless, P.1 has emerged as the most prevalent lineage sequenced in more recent samples [19]. Recently, newer mutations were detected in addition to the original set presented in P.1, giving rise to the sublineage P.1.2 [21]. P.2 probably emerged in Rio de Janeiro state (Southeast) [22], but it was also found in several municipalities of RS state as of October 2020 [23,24]. The first P.1 infection in the state was once thought to be in a patient from Gramado city in

February 2021 [25]. However, in a more recent study, the actual first P.1 was detected on 30 November 2020. This happened in a patient with comorbidities from Campo Bom city, who was reinfected by the P.2 lineage on 11 March 2021 [26].

Even though RS was one of the least affected Brazilian states in the first epidemic wave, it suffered a pronounced increase in cases in late 2020 [16]. In February 2021, the progressive increases in cases and hospitalizations (3.8-fold) led to the collapse of the local state healthcare system. Since recent findings of the widespread dissemination of the SARS-CoV-2 lineage P.1 in Brazil have been confirmed, we sequenced samples from patients from the metropolitan region of Porto Alegre to: (i) identify the prevalence of SARS-CoV-2 lineages in the region, the state, and bordering countries/regions; (ii) characterize the mutation spectra; (iii) hypothesize possible viral dispersal routes by using phylogenetic and phylogeographic approaches.

2. Results

2.1. Epidemiological Information

Of the 56 samples of hospitalized patients between March 9 and 17 2021, 75.0% (n = 42) of them were male, and the mean age was 37.2 years (interquartile range (IQR): 13.5 years). The mean cycle threshold (Ct) value for the first RT-qPCR conducted at Laboratório Exame was 19.12 cycles (median: 18.00; IQR: 6.00 cycles). Forty-seven (83.9%) had contact with a confirmed or suspected case. The majority of them were from the RS state capital, Porto Alegre (n = 32; 57.1%). In total, 51 (91.1%) were from the intermediate geographic region of Porto Alegre and 5 (8.9%) from the intermediate region of Santa Maria (Table 1 and Figure S2C).

2.2. SARS-CoV-2 Mutations and Lineages

Consensus SARS-CoV-2 genomes were obtained with an average coverage depth of 813.2× (median: 820.6×; IQR: 184.7×) (Figure S6). We detected 175 different mutations comprising all samples (Figure 1A). The ORF1ab carried 102 (58.3%) replacements followed by spike (n = 24; 13.7%), nucleocapsid (n = 18; 10.3%), ORF3a (n = 14, 8.0%), ORF7a (n = 6; 3.43%), ORF8 (n = 5; 2.86%), and membrane (n = 3; 1.7%) genes. Remarkably, 50% of the spike substitutions occurred in only one genome, and, of these, nine (75.0%) were missense (Table S1). Fifty-nine (33.7%) mutations were identified in two or more sequences. From these, 36 (61.0%) are non-synonymous (missense), 21 (35.6%) are synonymous, 1 (1.7%) is intergenic at 5′ Untranslated Region (UTR), and 1 (1.7%) is an inframe deletion. Highly frequent (≥10 genomes) mutations were found in 34 genomic positions, 24 (70.5%) being missense and 9 (26.5%) synonymous. Fifteen substitutions (10 in the spike protein: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, and T1027I) are P.1 lineage-defining mutations (Figure 1B and Table 2). The only P.1 defining replacement not found at high frequency in our study was the deletion in ORF1ab (del 11288:11296), called in only four genomes. Deletions overlapping annealing sites of amplicon primers are associated with a strong decrease in the PCR efficiency prior to sequencing, leading to low genomic coverage [27]. Then, after applying a stringent coverage depth filter (DP > 50) for calling the genomic positions in the consensus sequences, this deleted region was flagged as low coverage.

Most importantly, other positions presenting single nucleotide polymorphisms (SNPs) reached the appropriate threshold, since a point mutation is generally unable to cause dropouts.

After comparing the frequency of mutations from the recently sequenced samples and the Brazilian P.1 genomes, we observed a combination of mutations that stood out in a significant proportion (n = 11; 19.6%) compared with previous P.1 sequences from Brazil. This combination was previously described [21] and gave rise to the P.1.2 lineage, which harbors three ORF1ab replacements (synC1912T, D762G, and T1820I), one in ORF3a (D155Y), and one in N protein (synC28789T) (Table 2). Additionally, two of these genomes (18.2%) carry T11296G (ORF1ab nsp6: F3677L) and eight (72.7%) harbor G25641T (ORF3a: L83F) substitutions. Another cluster, made of four local genomes and subsequently named

Clade 2, was also detected. This clade possesses three defining mutations (ORF1ab nsp4: V2862L, synC10507T, and ORF3a: M260K), but it does not fall into a lineage designation at this moment but deserves further monitoring (Figure S1).

Table 1. Epidemiological characteristics of the 56 sequenced samples from Rio Grande do Sul, Southern Brazil.

Study ID (HBM-RS)	GISAID ID (EPI_ISL_)	Cycle Threshold	Municipality of Residence	Gender	Age Group	Lineage	Contact with Confirmed Case
39468	2139494	16	São Leopoldo	Male	30–39	P.1	Yes
39469	2139495	19	Porto Alegre	Female	20–29	P.1.2	Yes
39470	2139496	19	Porto Alegre	Male	60–69	P.1	No
39471	2139497	18	Porto Alegre	Male	20–29	P.1	Yes
39472	2139498	17	Gravataí	Male	30–39	P.1	No
39473	2139499	26	Cachoeira do Sul	Female	20–29	P.1	Yes
39474	2139500	18	Gravataí	Male	30–39	P.1	Yes
39475	2139501	18	Porto Alegre	Female	20–29	P.1	No
39476	2139502	15	Porto Alegre	Male	20–29	P.1	Yes
39477	2139503	21	Porto Alegre	Male	30–39	P.1	Yes
39478	2139504	15	Cachoeira do Sul	Male	30–39	P.1	Yes
39479	2139505	22	Porto Alegre	Male	50–59	P.1	Yes
39480	2139506	17	Novo Hamburgo	Male	40–49	P.1	Yes
39481	2139507	14	Porto Alegre	Female	70–79	P.1	Yes
39482	2139508	14	Porto Alegre	Female	80–89	P.1.2	No
39483	2139509	13	Gravataí	Male	30–39	P.1	Yes
39484	2139510	20	Porto Alegre	Male	20–29	P.1	Yes
39485	2139511	16	Porto Alegre	Male	50–59	P.1	Yes
39486	2139512	27	Porto Alegre	Male	30–39	P.2	No
39487	2139513	14	São Sebastião do Caí	Male	40–49	P.1.2	Yes
39488	2139514	28	Santo Antônio da Patrulha	Male	70–79	P.1	Yes
39489	2139515	27	Porto Alegre	Female	20–29	P.1.2	Yes
39490	2139516	18	Porto Alegre	Male	20–29	P.1	Yes
39491	2139517	15	Alvorada	Female	20–29	B.1.1.28	Yes
39492	2139518	17	Gravataí	Female	30–39	P.1	Yes
39493	2139519	22	Canoas	Male	30–39	P.1.2	Yes
39494	2139520	17	Porto Alegre	Female	30–39	P.1	No
39495	2139521	17	Porto Alegre	Male	30–39	P.1	Yes
39496	2139522	17	Canoas	Female	30–39	P.1	Yes
39497	2139523	21	Porto Alegre	Male	40–49	P.1.2	Yes
39498	2139524	20	Porto Alegre	Female	40–49	P.1	Yes
39499	2139525	22	Portão	Male	30–39	P.1	Yes
39500	2139526	11	Porto Alegre	Male	20–29	P.1.2	Yes
39501	2139527	14	Santa Maria	Male	20–29	P.1.2	Yes
39502	2139528	21	Porto Alegre	Male	30–39	P.1	Yes
39503	2139529	16	Porto Alegre	Male	30–39	P.1	No
39504	2139530	21	Gravataí	Male	40–49	P.1	Yes
39505	2139531	13	Porto Alegre	Male	30–39	P.1.2	Yes
39506	2139532	23	Porto Alegre	Female	40–49	P.1	Yes
39507	2139533	28	Canoas	Female	30–39	P.1	Yes
39508	2139534	22	Porto Alegre	Male	20–29	P.1	Yes
39509	2139535	23	Alvorada	Male	20–29	P.1	Yes
39510	2139536	19	Canoas	Male	50–59	P.1.2	Yes
39511	2139537	22	Porto Alegre	Male	30–39	P.1	No
39512	2139538	25	Cachoeira do Sul	Female	40–49	P.1	Yes
39513	2139539	23	Santa Maria	Male	40–49	P.1	Yes
39514	2139540	15	Porto Alegre	Male	30–39	P.1	No
39515	2139541	21	Porto Alegre	Male	20–29	P.1	Yes
39516	2139542	28	Porto Alegre	Male	50–59	P.1	Yes
39517	2139543	17	Sapiranga	Male	30–39	P.1	Yes
39518	2139544	17	Porto Alegre	Male	30–39	P.1.2	Yes
39519	2139545	23	Porto Alegre	Male	20–29	P.1	Yes
39520	2139546	15	Campo Bom	Male	20–29	P.1	Yes
39521	2139547	15	Porto Alegre	Male	20–29	P.1	Yes
39522	2139548	21	Porto Alegre	Male	50–59	P.1	Yes
39523	2139549	18	Porto Alegre	Male	20–29	P.1	Yes

All samples were nasopharyngeal swabs collected during 9–17 March 2021, from residents of RS state. Study ID: Study identifier only known by study investigators. The Ct values are related to the first RT-qPCR conducted at Laboratório Exame.

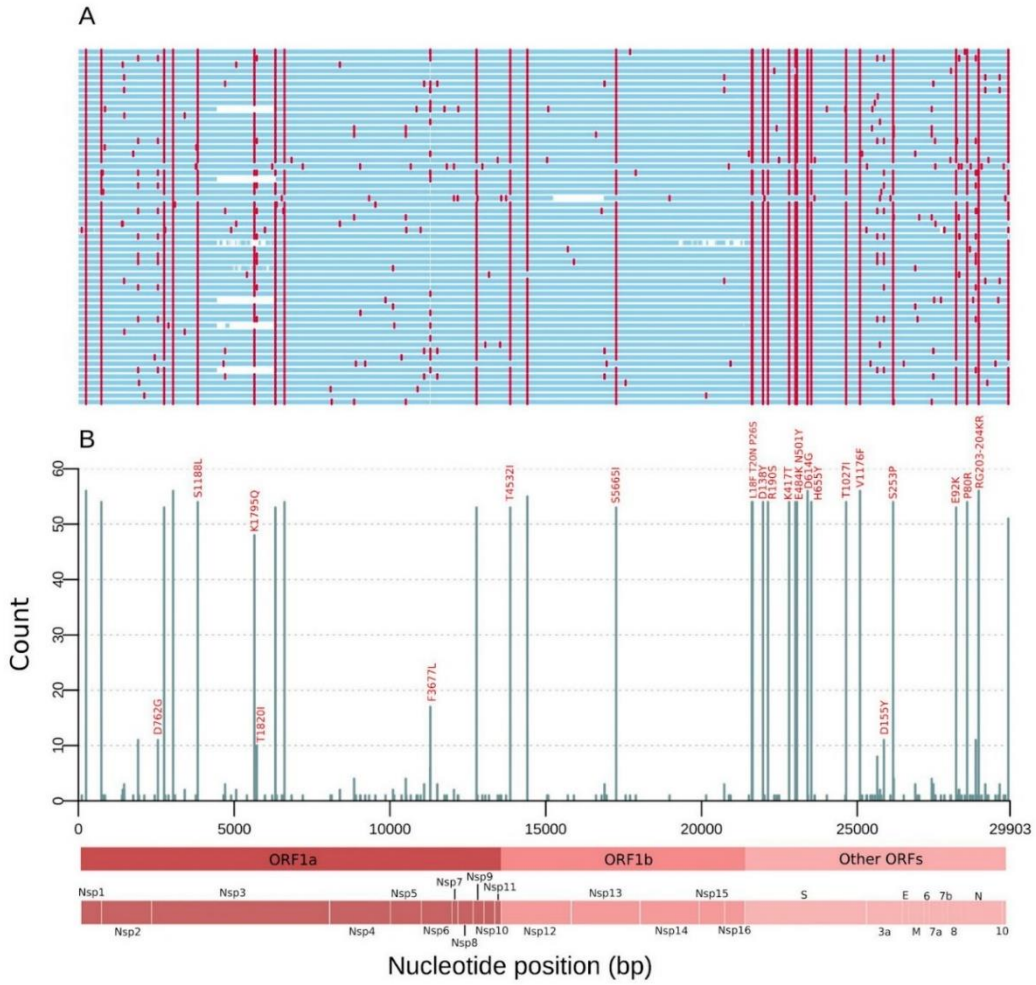

Figure 1. Mutations of the SARS-CoV-2 genomes from RS state, Southern Brazil sampled in March 2021. (**A**) Genome map for the 56 genomes sequenced. Nucleotide substitutions are colored in red and blank regions represent low sequencing coverage. (**B**) Count of single nucleotide polymorphisms (SNPs) per SARS-CoV-2 genome position along the 56 genomes. These mutations are corresponding to the red lines in (**A**), and only missense substitutions represented by >10 sequences have their respective amino acid changes indicated above the bars. Main open reading frames (ORFs) and SARS-CoV-2 proteins are indicated at the bottom to allow a rapid visualization of the viral proteins affected.

Considering PANGO lineages, 54 genomes (96.4%) were designated as P.1, one (1.8%) as P.2, and one (1.8%) as B.1.1.28. Even without being classified according to the Pango-designation's most updated version, the P.1.2 lineage was present in 11/54 (20.4%) of the P.1 sequences (https://github.com/cov-lineages/pango-designation/issues/56; accessed on 6 May 2021) (Figure S1).

2.3. Lineage Distribution in Neighboring Countries and Brazilian Regions

RS state shares borders with Argentina to its west (Figure S2), leading to the transit of people at the frontiers. From March to April 2020, B.1 was the most prevalent lineage in this bordering country. B.1.499 and N.3 were abundant from May to July when the N.5 started to rise and surpassed B.1.499 in November 2020 (Figure 2A). Importantly, N.3 and N.5 are derived from the B.1.1.33 lineage widespread in Brazil (https://cov-lineages.org/lineages.html; accessed on 4 May 2021). The P.2 lineage, which initially emerged in

Brazil [22] and derived from another Brazilian disseminated lineage (B.1.1.28), was firstly found in November 2020, and, by January and February 2021, it had already outnumbered the other lineages in Argentina (Figure 2A).

Table 2. Detailed description and frequency of mutations found in our 56 sequences compared with all Brazilian P.1 sequences until 26 April 2021.

Genomic Position	Effect	Amino Acid Change	Gene/Region	Product	Frequency Our Study (%)	Frequency in Brazilian's P.1 (%)
C241T	Intergenic	NA	5′ UTR	NA	100.0	97.2
T733C	Synonymous	D156D	ORF1ab	Leader Protein	96.4	99.9
C1912T	Synonymous	S549S		nsp2	19.6	1.4
A2550G	Missense	D762G			19.6	1.5
C2749T	Synonymous	D828D			94.6	99.7
C3037T	Synonymous	F924F			100.0	99.9
C3828T	**Missense**	**S1188L**			96.4	95.3
A5648C	**Missense**	**K1795Q**		nsp3	85.7	100.0
C5724T	Missense	T1820I			17.9	2.3
A6319G	Synonymous	P2018P			87.5	99.5
A6613G	Synonymous	V2116V			96.4	99.8
T11296G	Missense	F3677L		nsp6	30.4	8.2
C12778T	Synonymous	Y4171Y		nsp9	94.6	98.9
C13860T	Missense	T4532I		RdRp	94.6	99.8
C14408T	Synonymous	L4715L			98.2	96.8
G17259T	Missense	S5665I		Helicase	94.6	99.7
C21614T	**Missense**	**L18F**	S	Surface Glycoprotein	96.4	99.9
C21621A	**Missense**	**T20N**			94.6	99.8
C21638T	**Missense**	**P26S**			96.4	99.1
G21974T	**Missense**	**D138Y**			96.4	100.0
G22132T	**Missense**	**R190S**			96.4	98.4
A22812C	**Missense**	**K417T**			96.4	83.4
G23012A	**Missense**	**E484K**			96.4	99.9
A23063T	**Missense**	**N501Y**			96.4	99.8
A23403G	Missense	D614G			100.0	97.7
C23525T	**Missense**	**H655Y**			96.4	100.0
C24642T	**Missense**	**T1027I**			96.4	99.9
G25088T	Missense	V1176F			100.0	99.9
G25855T	Missense	D155Y	ORF3a	ORF3a Protein	19.6	1.6
T26149C	**Missense**	**S253P**			94.6	98.7
G28167A	**Missense**	**E92K**	ORF8	ORF8 Protein	94.6	99.8
C28512G	**Missense**	**P80R**	N	Nucleocapsid Phosphoprotein	96.4	98.3
C28789T	Synonymous	Y172Y			19.6	1.3
AGTAGGG 28877–28883 TCTAAAC	Missense	RG203-204KR			96.4	99.8

Original bases or amino acids are represented before the genomic coordinate, while the mutated ones are presented after. Only mutations observed in more than 10 genomes from this study are shown. P.1 lineage-defining mutations are highlighted in bold. P.1.2 (new lineage) defining replacements are underlined and marked with gray background color. UTR, untranslated region; ORF, open reading frame; S, spike; N, nucleocapsid; nsp, nonstructural protein; RdRp, RNA-dependent RNA polymerase.

In the entire Brazil, despite early introductions of B.1 and B.1.1, lineages B.1.1.28 and B.1.1.33 were most abundant from March to October 2020. In October, P.2 already represented an important portion of the sequences, and, by November, it had already surpassed B.1.1.33. In December 2020 and January 2021, with the emergence of P.1, this lineage and P.2 already became the most prevalent, while, between February and April, P.1 replaced all other lineages (Figure 2A,B). Some fluctuations evidently occurred in different Brazilian regions, such as a prevalence of more local lineages (e.g., B.1.195 and B.1.1.378 in the Northern region, B.1.1 and N.9 in the Northeast, and B.1.1.7 in the Southeast and

Centre-West regions) (Figure S3). In RS, a similar landscape was observed compared to the Brazilian scenario. B.1.1.28 and B.1.1.33 were most prevalent until October 2020, when P.2 emerged and remained until January 2021 along with B.1.1.28 as the most prevalent. After the introduction of P.1 (January 2021), this lineage practically supplanted the others in February and March 2021 (Figure 2A,B).

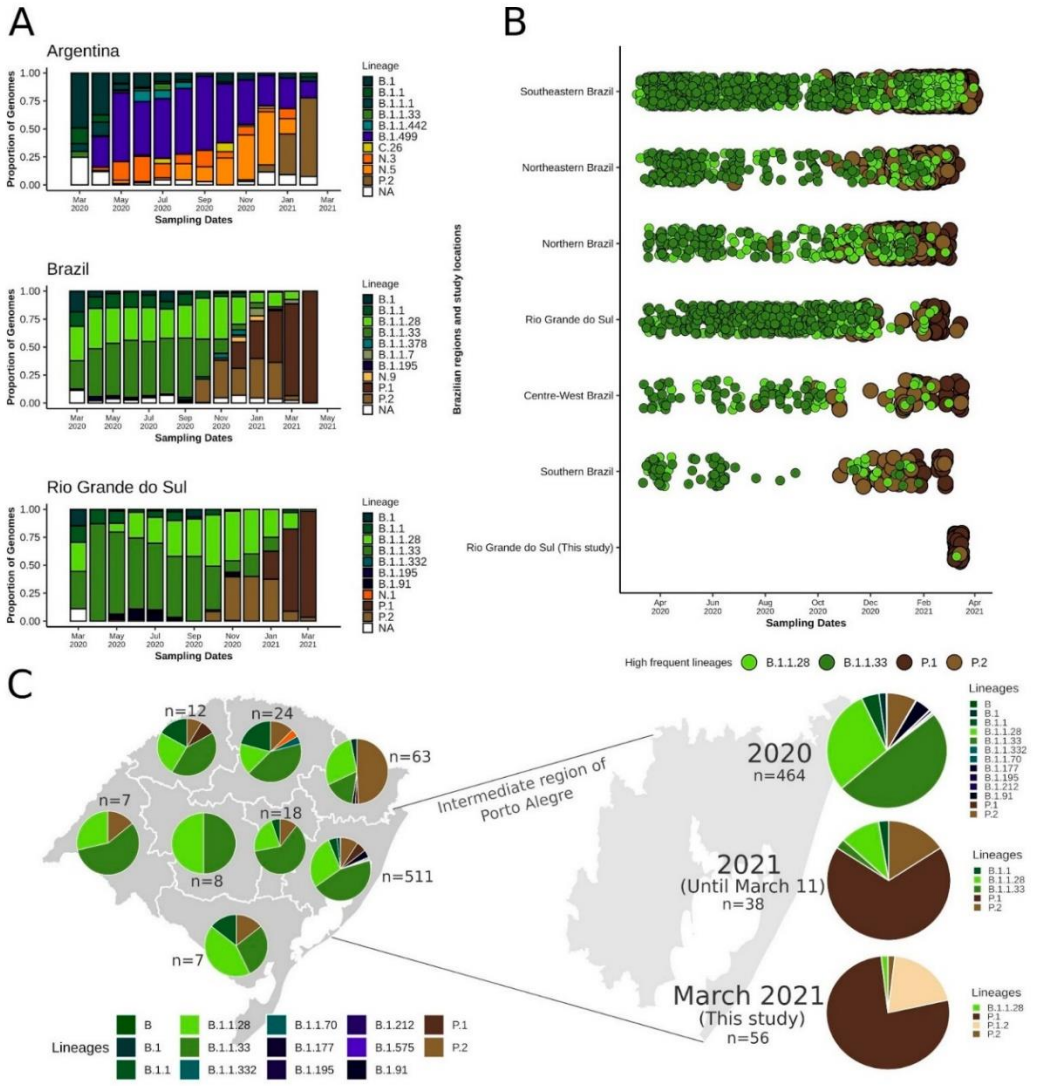

Figure 2. Distribution of SARS-CoV-2 lineages across time in Argentina, other Brazilian regions, and RS state. The other bordering country (Uruguay) was not included due to the limited number of samples available (n = 135). (**A**) Per month lineage distribution in Argentina, all of Brazil, and RS state. Only the 10 most prevalent lineages were considered. (**B**) Timeline showing the distribution of the most prevalent lineages until the end of 2020 (B.1.1.28 and B.1.1.33) and from the end of 2020 onward (P.1 and P.2) for the five Brazilian regions, RS state, and considering only this study (March 2021). (**C**) Map of RS state divided into eight intermediate regions, as defined by IBGE, displaying the proportion of the lineages in each area. The region of Porto Alegre was amplified and the lineage frequencies from 2020, 2021 and the present study are presented on the right. NA, other lineages.

After dividing RS into the intermediate regions proposed by IBGE (Figure S2), it was possible to gain insights into the dynamics of the lineages in the state, despite the low sample size of some regions (Figure 2C). In most regions, the lineages B.1.1.28 and B.1.1.33 were more prevalent, but P.2 was also detected. In fact, in the Caxias do Sul region, more P.2 (n = 31; 49.2%) were sequenced in relation to other lineages. Since the Porto Alegre region has a larger sample size, we divided its results by year to check the most recent (2021) evolutionary abundance. In 2020, B.1.1.33 (n = 229; 49.3%) and B.1.1.28 (n = 137, 29.5%) were the most abundant, followed by P.2 (n = 37; 8.0%). In 2021, P.1 (n = 26, 68.4%) and P.2 (n = 6, 15.8%) have already outperformed the other lineages (Figure 2C). In our study from March 2021, 96.4% of the samples were classified as P.1. We were able to identify a new P.1 sublineage (P.1.2) in 11 (20.4%) genomes from four different municipalities (Porto Alegre, Canoas, São Sebastião do Caí, and Santa Maria), demonstrating the possible diversification of P.1 and its spread within RS (Figure 2C and Figure S1).

2.4. Maximum Likelihood Phylogenomic Analysis

After running the Nextstrain workflow using quality control and subsampling approaches, we obtained a dataset of 8635 time- and geographical-representative genomes. From these, 861 were from Africa, 1370 from Asia, 2219 from Europe, 481 from North America, 218 from Oceania, and 3486 from South America. Brazil was represented by 2608 sequences and RS state by 730 sequences (56 from this study and 674 available in GISAID) (Table S2).

The time-resolved ML phylogenetic tree confirmed the PANGO lineages assigned, since 54 genomes (96.4%) grouped with P.1 representatives, 1 (1.8%) with B.1.1.28, and 1 (1.8%) with P.2 sequences. We also observed a strong correlation between genetic distances and sampling dates ($R^2 = 0.71$). The P.1 sequences were grouped above the regression line, showing higher evolutionary rates than the other lineages in the SARS-CoV-2 phylogeny, as observed in other studies [11,14]. We highlighted the most abundant global lineages present in RS state that passed the quality control criteria (B.1.1 (n = 32), P.1 (n = 83), P.2 (n = 83), B.1.1.28 (n = 203), and B.1.1.33 (n = 286)). We also noticed the high abundance of B.1.1.28 and B.1.1.33 lineages until October and November 2020, followed by the rise and establishment of P.2 and P.1, respectively (Figure 3A).

The only B.1.1.28 sequence identified in this study (RS-HBM-39491) branched in a clade represented by 30 sequences, mostly represented by Southeastern Brazilian (n = 17; 56.6%) genomes. This clade is supported by the ORF1ab:synC15810T mutation and includes a subclade characterized by the ORF1ab:L4182F mutation, where the local sequence is placed together with four samples from São Paulo (SP), one from Portugal, and one from Chile (Figure S4). Most importantly, this local genome harbors seven other mutations: ORF1ab:T2087I (nsp3), D3022N (nsp4), N3970S (nsp8), V4436A (RNA-dependent RNA polymerase), synC13724T, synG18973A, and intergenic:G29736T. Additionally, the only P.2 sequence from this study (RS-HBM-39486) formed a separate clade composed of 20 sequences from several Brazilian states (10 from RS, 1 from Paraná, 1 from SP, and 1 from Maranhão), 5 from Argentina, 1 from USA, and 1 from Norway (Figure S5). This clade is characterized by the ORF1ab:synT6218C (nsp3) mutation. Moreover, this local sequence accrued seven specific mutations: ORF1ab:synA7201G, S2926F (nsp4), V6871A (2′-O-ribose methyltransferase), S:G1251V, ORF7a:G38V, N:synC28333T, and intergenic:G29688T.

To get a more detailed understanding of the P.1 diffusion throughout Rio Grande do Sul, other Brazilian regions, and worldwide, we built an ML tree of 4499 genomes belonging to this lineage (Table S3). P.1 sequences from this study were allocated into several distinct branches, suggesting multiple introductions and the formation of different P.1-derived clades and clusters.

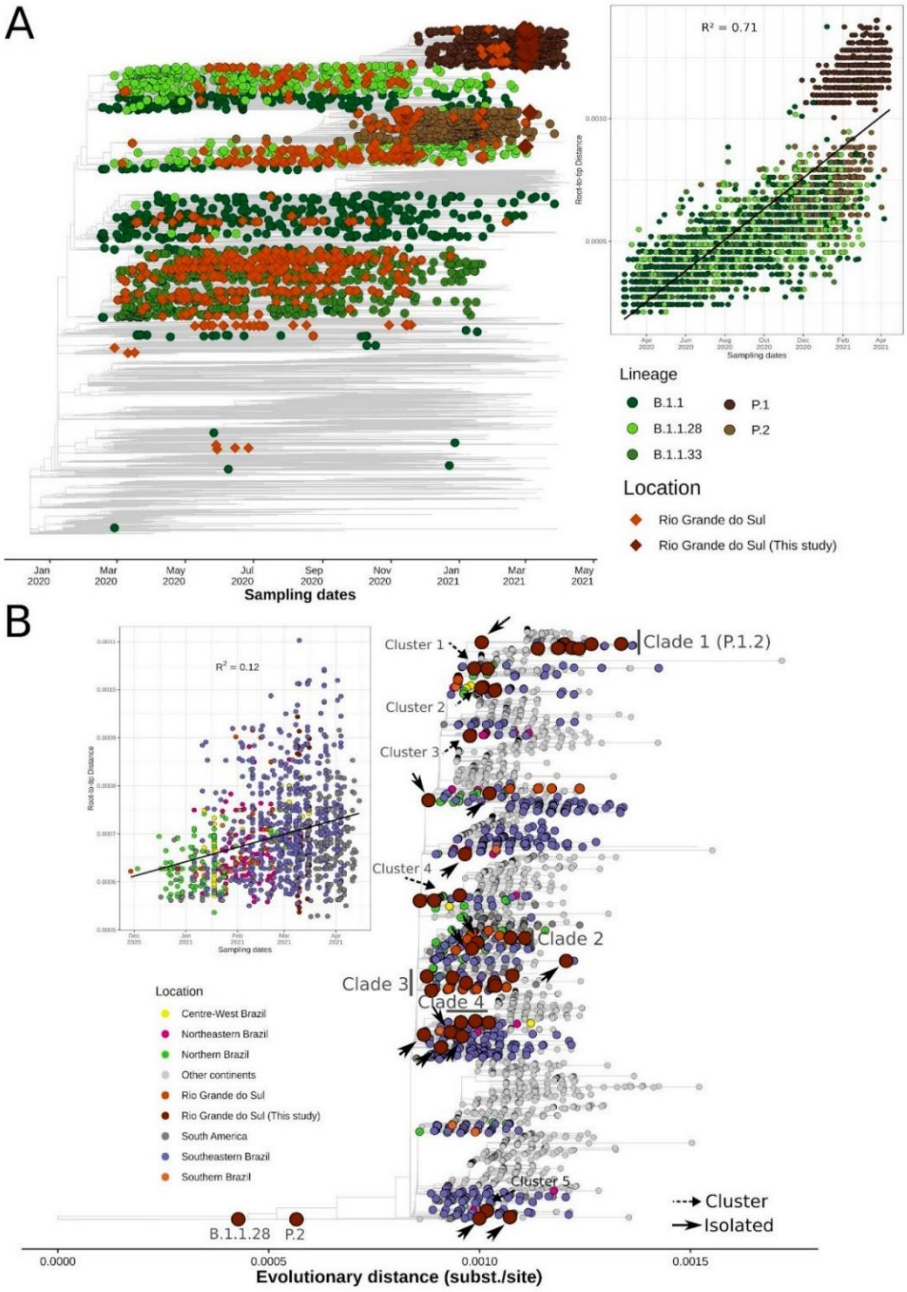

Figure 3. Phylogenetic analysis of genomes sequences in RS state in a global context. (**A**) Time-resolved ML tree of 8635 global representative SARS-CoV-2 genomes. Circles represent global sequences belonging to the five most abundant lineages in RS state that passed quality control criteria: B.1.1 (n = 32), P.1 (n = 83), P.2 (n = 83), B.1.1.28 (n = 203), and B.1.1.33 (n = 286). Diamonds represent RS genomes (available in GISAID and sequenced in this study). Root-to-tip regression is represented on the right of the tree. (**B**) ML tree of 4499 SARS-CoV-2 genomes belonging to the P.1 lineage. Tips are colored by Brazilian regions, South America, or other continents. Introductions, clusters, and clades are annotated in the tree (see Methods). Root-to-tip regression is depicted on the left of the tree and sequences from "other continents" were dropped to improve visualization.

We identified 4 clades, 5 clusters, and 13 isolated sequences (Figure 3B and Figure S7). Most importantly, Clade 1 had high branch-support (SH-aLRT = 76.3) and was composed of 11 sequences originated in this study that shared five lineage-defining mutations as previously described (Table 2) and were recently attributed to the P.1.2 sublineage (https://github.com/cov-lineages/pango-designation/issues/56; accessed on 6 May 2021). As of April 26, 2021, this sublineage is already distributed worldwide in 93 sequences (the Netherlands, Spain, England, and USA) and in other Brazilian states (Rio de Janeiro (RJ) and SP) [21]. Clade 2 sequences harbored two mutations in ORF1ab:V2862L (nsp4) and synC10507T and one in ORF3a:M260K, and it comprised 81 genomes. Four samples are from this study. The majority are from Amazonas (n = 15), São Paulo (n = 11), RS (n = 8), and Bahia (BA) (n = 4), and worldwide sequences are mainly represented by French Guiana, USA, Spain, Japan, and Jordan. Clade 3 is represented by three ORF1ab mutations (synC1420T, D1600N [nsp3], and synT8392A) in three of the seven local genomes. It is composed of sequences from RS (n = 25), SP (n = 15), Maranhão (n = 10), and RJ (n = 8), as well as other countries (mainly Spain, French Guiana, and USA). Clade 4 is characterized by two ORF1ab substitutions (G400S (nsp2) and S6822I (2′-O-ribose methyltransferase)), one N:synT26861C in three genomes, and carries other additional mutations (ORF1ab: synG10096A, G3676S (nsp6), F3677L (nsp6)) and M:synT26861C. This clade is mainly found in SP (n = 11), RS (n = 7), Santa Catarina (n = 5), BA (n = 4), and Goiás (n = 4), as well as other countries (mainly USA, Chile, and England).

Clusters 1 and 3 have, respectively, one (ORF1ab: G3676S (nsp6)) and two (ORF1ab: synC1471T and A1049V (nsp3)) shared mutations. Among all identified clusters, the most diverse was Cluster 5, which contains three samples from this study and has five defining mutations: four in ORF1ab (synT4705C, synC11095T, syn11518, and T5541I (helicase)) and one in ORF7a: E16D. Moreover, two sequences share one distinct mutation (ORF1ab: F3677L (nsp6)).

2.5. Bayesian Molecular Clock and Phylogeographic Analysis

To date the time of the most recent common ancestor (TMRCA) and the diffusion of the four P.1 clades identified in our ML analysis, we used coalescent and phylogeographic methods. For Clade 1, which is correspondent to the recently labeled P.1.2 lineage, sequences showed a moderate correlation of genetic distances and sampling dates (correlation coefficient: 0.59, $R^2 = 0.34$) (Figure 4A). We estimated a median evolutionary rate of 7.68×10^{-4} (95% highest posterior density interval [HPD]: 4.18×10^{-4} to 1.14×10^{-3} subst/site/year) and the TMRCA on 18 December 2020 (95% HPD: 29 October 2020 to 31 January 2021). Interestingly, the tree's root was placed in RS, between a sequence from RS (the oldest sequence from this clade: EPI_ISL_983865) and a subclade from USA. The divergence of these subclades was dated on 15 January 2021 (95% HPD: 15 January to 26 March 2021). The subclade composed of the RS sequences formed two separate clusters, one with three sequences from this study and one Australian genome and another composed of sequences from RS, SP, UK, Portugal, USA, and transmission clusters from RJ and Netherlands (Figure 4B). The emergence of an important cluster in RJ carrying additional mutations [21] was dated here on 11 March 2021 (95% HPD: 11 March to 6 April 2021). As American sequences formed a separate subclade, local transmission is probably occurring in the country. The divergence of the American subclade was dated 7 February 2021 (95% HPD: 1 February to 11 May 2021). In accordance with the root being placed in RS, the BSSVS procedure identified well-supported rates of diffusion from RS to other Brazilian states such as São Paulo (Bayes Factor (BF): 6.82; posterior probability (PP): 0.52), Rio de Janeiro (BF: 39.18; PP: 0.86), and other countries such as USA (BF: 31.94; PP: 0.84) and Netherlands (BF: 80.16; PP: 0.93).

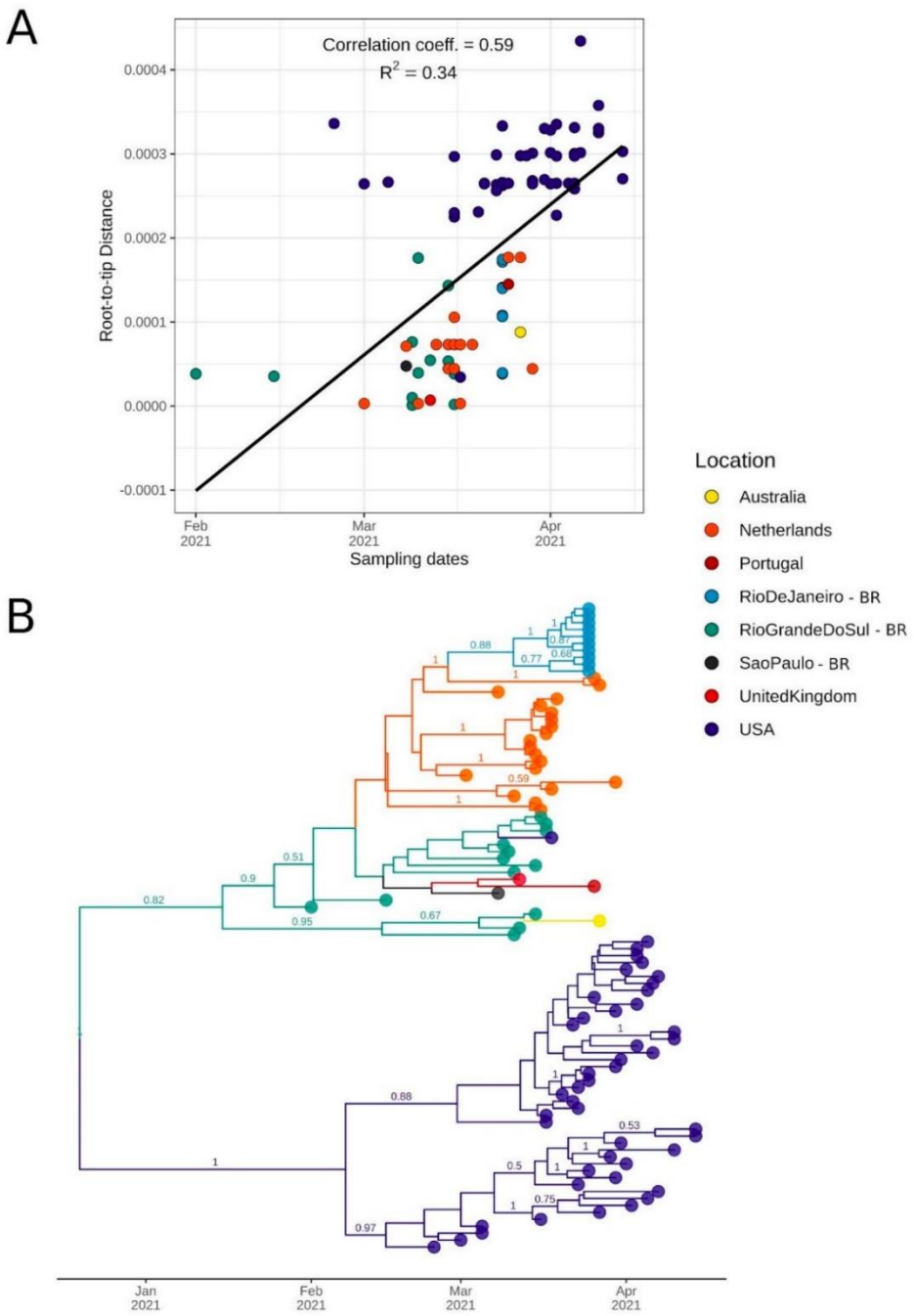

Figure 4. Bayesian discrete asymmetric phylogeographic analysis of the identified Clade 1 (lineage P.1.2). (**A**) Root-to-tip regression of genetic distances and sampling dates for Clade 1. Correlation coefficient and R squared are depicted above the graph. (**B**) MCC tree of the 93 sequences included in this analysis up to 26 April 2021 (82 from GISAID and 11 from this study). Numbers above branches represent the posterior probability of each branch. Only posteriors > 0.5 are shown. Circles indicate countries outside Brazil and Brazilian states (BR suffix).

However, it is possible that this lineage emerged in another Brazilian state, but its earlier representatives were not sampled. This is a strong hypothesis since this sequence is associated with community transmission after contact with tourists in a city of RS (Gramado) that receives numerous visitors annually [25].

For Clade 2, we estimated a median evolutionary rate of 5.85×10^{-4} (95% HPD: 4.18×10^{-4} to 7.71×10^{-4} subst/site/year), and the TMRCA was dated 30 November 2020 (95% HPD: 2 November to 21 December 2020). This clade includes sequences from 11 Brazilian states from all 5 regions and 9 other countries. We were able to detect at least five introductions from Amazonas, where this clade probably emerged. These introductions ranged from December 28, 2020 (95% HPD: 28 December 2020 to 5 January 2021) to 28 January 2021 (95% HPD: 28 January to 7 March 2021). Importantly, we identified a well-supported subclade (PP = 1) of four genomes from this study (Figure 5A).

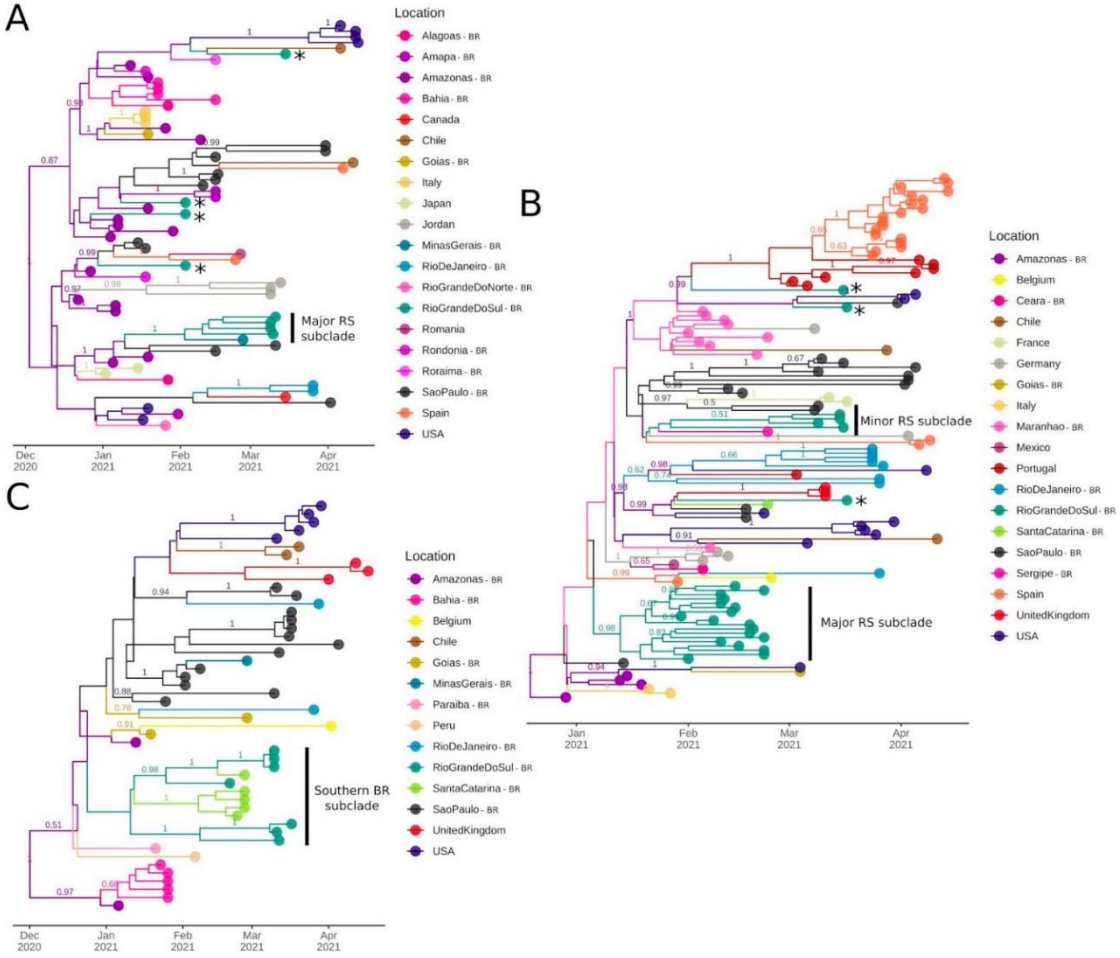

Figure 5. Bayesian discrete asymmetric phylogeographic analysis of the identified Clades 2–4 up to 26 April, 2021. (**A**) Clade 2: MCC tree of the 71 sequences included in this analysis. (**B**) Clade 3: MCC tree of the 122 genomes included in this analysis. (**C**) Clade 4: MCC tree of the 50 sequences included in this analysis. For all MCC trees, numbers above branches represent the posterior probability of each branch. Only posteriors > 0.5 are shown. Asterisks represent potential introductions in RS state and subclades cited in the text are indicated. Circles indicate countries outside Brazil and Brazilian states (BR suffix).

For Clade 3, the TMRCA was estimated on 20 December 2020 (95% HPD: November 25 to 29 December 2020) and the median evolutionary rate was 7.85×10^{-4} (95% HPD: 6.06×10^{-4} to 1.02×10^{-3} subst/site/year). This clade harbors sequences from 9 Brazilian states and 10 other countries. Amazonas is the most probable source of its emergence. From then onwards, multiple transmission clusters were established in foreign countries (e.g., Spain, Portugal, and USA) and Brazilian states (especially Maranhão, SP, and RS). This clade was introduced at least 5 times in RS, leading to 2 major subclades represented by 18 and 4 sequences, respectively. The major subclade (n = 18, PP = 0.98) was dated 11 January 2021 (95% HPD: 11 January to 1 February 2021) (Figure 5B).

For Clade 4, the TMRCA was dated on 2 December 2020 (95% HPD: 7 October 2020 to 3 January 2021), and the median evolutionary rate was 6.26×10^{-4} (95% HPD: 3.51×10^{-4} to 1.01×10^{-3}). This clade comprises nine Brazilian states and five foreign countries. After its initial emergence and spread in Amazonas, it had already formed transmission clusters in SP, BA, United Kingdom, and USA. Most importantly, a subclade containing sequences from two neighboring states from Southern Brazil (seven from RS and five from Santa Catarina (SC)) indicates its diffusion from RS to SC, probably leading to two separate introductions. The divergence of this subclade was estimated on 16 December 2020 (95% HPD: 16 December 2020 to 19 January 2021) (Figure 5C).

Phylogenetic and molecular clock approaches suggest the wide circulation of the VOC P.1 both nationally and internationally between late 2020 and early 2021. This lineage has already diversified into some clades that bear characteristic mutations, although they exhibit similar evolutionary rates. We inferred that P.1 (and its derived clades) was introduced multiple times in the southernmost Brazilian state (RS) between mid-December 2020 and January 2021. Remarkably, this date is close to the first P.1 detection in Manaus, which is located ~4000 km away. These early introductions led to the formation of local subclades that could be identified even using a reduced set of sequenced samples.

3. Discussion

In this study, the analysis of 56 samples from the state of Rio Grande do Sul (RS), Southern Brazil, confirmed that the P.1 lineage was already highly prevalent. Interestingly, we demonstrated that P.1 is already showing signs of diversification and has originated a new sublineage (P.1.2). Herein, we indicate the likely origin and the first clusters of this novel lineage. This sublineage was detected in three Brazilian states, and other countries, and its most recent common ancestor was dated on mid-December, 2020 (95% HPD: 29 October 2020 to 31 January 2021). In accordance with the majority of the states from Brazil, this state experienced significant increases in hospitalizations in early 2021. This scenario was related to the emergence and rapid spread of the P.1 variant across the country.

After almost one year of relatively slow SARS-CoV-2 evolution, the emergence of multiple and convergent lineages harboring a constellation of mutations in the spike protein raised concern in the scientific community. This protein is responsible for mediating interaction with the human Angiotensin-Converting Enzyme 2 receptor (hACE2) and is a primary target of neutralizing antibodies and vaccines [28]. The variants harboring different mutational signatures, including spike protein substitutions, were classified as VOCs and "variants of interest" (VOIs), depending on their growing relevance in the current pandemic. The first three VOCs emerged in England (B.1.1.7) [8], South Africa (B.1.351) [9], and Brazil (P.1) [10]. More recently, B.1.617.2 (India) [12,13] also was characterized as a VOC. By May 2020, B.1.427/429 [29], B.1.526 (New York, USA), B.1.617 (India), and P.2 (Brazil) [22] were categorized as VOIs. B.1.1.7, B.1.351, and P.1, the most studied VOCs, have the D614G and N501Y mutations in common. B.1.351 and P.1 share a mutation in the K417 site (K417N and K417T, respectively) and the E484K replacement, which is also observed in the P.2 lineage. Additionally, B.1.1.7 carries the P681H substitution in the furin-cleavage site and multiple VOIs bear the L452R substitution [7]. The presence of common substitutions in different SARS-CoV-2 lineages suggests co-evolutionary and convergent mutational processes [8–10,30].

In the present study, we noticed that B.1.1.33 and B.1.1.28 lineages, detected at the beginning of the pandemic in Brazil [2], had been similarly prevalent in different regions until September 2020, before the appearance of P.2 (in October) and P.1 (in December 2020). The B.1.1.33 lineage shows variable abundance in different Brazilian states (ranging from 2% in Pernambuco to 80% in Rio de Janeiro), with moderate prevalence in South American countries (5–18%). Surprisingly, this lineage was firstly detected in early March 2020 in other American countries (e.g., Argentina and USA). Apparently, an intermediate strain probably emerged in Europe and subsequently spread to Brazil, where its spread gave rise to B.1.1.33 [31] and possibly triggered secondary outbreaks in Argentina and Uruguay [31,32]. We found that N.3 and N.5, both derived from B.1.1.33, represented an important proportion of the sequences from Argentina from May to December 2020, when it was replaced by the P.2 lineage, which probably emerged in Rio de Janeiro (Southeastern Brazil). The B.1.1.28 lineage, despite apparently being less abundant than B.1.1.33 in several Brazilian regions, quickly diversified into two variants: VOC P.1 and VOI P.2 [33]. Since the end of 2020, these two lineages have lead the diversity of SARS-CoV-2 in Brazil [17] and have caused concern in other countries after several introductions. Regarding the distribution of sequenced samples across RS state, the cumulative frequency of B.1.1.28 and B.1.1.33 was higher until mid-April 2021 [16]. However, since the end of 2020 and beginning of 2021, a rise in P.1 and P.2 sequences was observed. Our study supported that P.1 outperformed other lineages in RS as of March 2021, although the collection of samples in hospitalized patients and low geographic representativeness does not allow the extrapolation of these findings.

The emergence of a B.1.1.28 derived lineage carrying the S:E484K mutation (P.2) was dated, in a retrospective study, late February 2020 in the Southeast (São Paulo and Rio de Janeiro), followed by transmission to the South (especially RS). Since then, multiple dispersion routes were observed between Brazilian states, especially in late 2020 and early 2021 [18]. However, this lineage went unreported until October 2020, when it was first detected in the state of Rio de Janeiro [22] and in the small municipality of Esteio in RS [23]. The increased frequency of B.1.1.28 and derived lineages was corroborated by another study that included samples from several municipalities of RS in November 2020. This study found that 86% of the genomes could be classified as B.1.1.28 and ~50% of these, in fact, belong to the new lineage P.2 [24]. Nonetheless, our current study suggests that P.2 has already been nearly entirely replaced by the P.1 lineage or is not particularly well represented among the analyzed patients seeking emergency consultation or requiring hospitalization.

Between June and October 2020, an extremely high seroprevalence (44–76%) was observed in Manaus (Amazonas, Brazil) in a study from blood donors [14]. However, despite these numbers, Manaus faced a resurgence of cases and a six-fold increase in hospitalizations between December 2020 and January 2021. The most plausible hypotheses that would justify this condition are: (i) the previous overestimation of seroprevalence in Manaus; (ii) the immune evasion property of some SARS-CoV-2 mutations found in VOCs; (iii) higher transmissibility and pathogenicity of SARS-CoV-2 lineages circulating in the second wave compared with pre-existing lineages [15].

A genomic epidemiology study that used 250 SARS-CoV-2 genomes from 25 different municipalities from Amazonas sampled between March 2020 and January 2021 shows that the first exponential phase in the state was driven mainly by the spread of lineage B.1.195, which was gradually replaced by B.1.1.28. The second wave coincided with the emergence of P.1 in November, which rapidly replaced the parental lineage (<2 months) [11] and whose emergence was preceded by a period of rapid molecular evolution [10]. Importantly, rapid accumulation of mutations over short timeframes have been reported in chronically infected or immunocompromised hosts [34,35]. However, preliminary findings pointed to the existence of P.1 intermediate lineages, suggesting that the constellation of mutations defining P.1 were acquired at sequential steps during multiple rounds of infections instead of within a single long-term infected individual [36]. The VOC P.1 carries three deletions,

four synonymous substitutions, a four base-pair nucleotide insertion, and at least 17 other lineage-defining replacements, including 10 missense mutations in the spike protein (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, and T1027I), 8 of which are subjected to positive selection [10].

Regarding infectiousness, transmissibility, and case fatality, the viral load was ~10-fold higher in P.1 infections than in non-P.1 infections [11]. Although another study points to uncertainties regarding viral load and duration of infection after accounting for confounding effects [10]. Moreover, it was estimated to be 1.7–2.4-fold more transmissible, raising the probability that reinfections would be caused more frequently in hosts infected by P.1 rather than by older lineages. Remarkably, infections were 1.2–1.9 times more likely to result in death in the period following the emergence of P.1 compared to previous time frames [10]. These findings support that successive lineage replacements in Amazonas were driven by a complex combination of factors, including the emergence of the more transmissible VOC P.1 virus [11].

A study conducted in RS described a P.1 lineage infection on 30 November 2020 followed by a P.2 lineage reinfection on 11 March 2021 in a patient with comorbidities. This report was the first detected P.1 in the state [26]. Other analyses suggest that the P.1 lineage presumably emerged in Manaus, Brazil, in mid-November 2020 [10,11]. Therefore, the P.1 lineage was present in Southern Brazil about a few days after its emergence in Manaus, Northern Brazil. Our molecular clock analysis supported this scenario. Another study, once thought to be the first P.1 report in RS, documented local transmission of P.1 from a person who had close contact with tourists and was positive for COVID-19 in early February 2021 [25]. This happened in the city of Gramado, a town in the mountains that receives around 6.5 million tourists every year and belongs to the Caxias do Sul intermediate region. Interestingly, this sample from Gramado was the earliest representative of a new P.1-derived lineage (P.1.2), described in 11 patients from our study and found in transmission clusters from the RJ state in Southeastern Brazil, USA, and the Netherlands. Remarkably, our local sequences are more similar to genomes from other countries compared to the RJ cluster, which acquired at least four additional mutations (including S:A262S) [21].

Whether P.1.2 has worse clinical outcomes than its prior variant (P.1) is unknown. However, as described above, the missense mutations characteristic of the new sublineage are located at nsp2 and nsp3 (ORF1ab), ORF3a, and nucleocapsid. These sites are known for their interaction with human proteome, potentially influencing the immunological and inflammatory response against SARS-CoV-2 infection [37]. The ORF3a:D155Y substitution is located near SARS-CoV caveolin-binding Domain IV. The binding interaction of viral ORF3a protein to host caveolin-1 is essential for entry and endomembrane trafficking of SARS-CoV-2. Since this mutation breaks the salt bridge formation between Asp155 and Arg134, it can interfere with the binding affinity of ORF3a to host caveolin-1 and change virulence properties. Most importantly, this disrupted interaction may be associated with improved viral fitness, since it can avoid the induction of host cell apoptosis or extend the asymptomatic phase of infection [38]. We hypothesize that these new substitutions could, therefore, influence epidemiological and clinical outcomes favoring P.1.2 evolution. This is elusive at best at this time, however, and further sublineage characterization is needed to further explore its real relevance.

Some limitations should be considered. Firstly, the sample size is low and not necessarily representative of RS state. Considering the number of sequences from each intermediate region in RS available in GISAID, it is very likely that the distribution seen on the map (Figure 2C) is a consequence of sampling at different times in these localities or simple randomness. Thus, it should not be assumed as a true representation of the spatial diversity in the state. Since publicly available genomes are a result of episodic sequencing efforts, especially in Brazil, more precise inferences about introductions and diffusion processes in regional and worldwide contexts are restricted due to the lack of proper geographical and temporal distribution of the samples. Therefore, more research and surveillance are

essential to unravel a more precise genomic characterization of SARS-CoV-2 in Brazil, promptly identifying novel variants to better respond and control its spread.

In summary, our study corroborates the total virtual substitution of previous lineages by P.1 in Southern Brazil in COVID-19 cases sequenced in March 2020. Moreover, we confirmed various cases caused by the novel P.1.2 sublineage and placed its origin in the state of Rio Grande do Sul. The continuous evolution of the VOC P.1 is concerning, considering its clinical and epidemiological impact, and warrants enhanced genomic surveillance.

4. Materials and Methods

4.1. Sample Collection and Clinical Testing

Samples were obtained from Hospital da Brigada Militar patients, both admitted or visiting the emergency ward, from Porto Alegre, RS, Brazil. Nasopharyngeal swabs were collected and placed in saline solution. Samples were transported to the clinical laboratory (Laboratório Exame) and tested on the same day for SARS-CoV-2 using Real-Time Reverse-transcriptase Polymerase Chain Reaction (Charité RT-qPCR assays). The RTq-PCR assay used primers and probes recommended by the World Health Organization targeting the nucleocapsid (N1 and N2) genes [39]. Remnant samples were stored at −20 °C.

Between 9 March and 17 March 2021, all routinely tested samples of the clinical laboratory provenient of the Hospital da Brigada Militar patients and yielded positive RT-qPCR were selected. Subsequently, those positive clinical samples were submitted to a second RT-qPCR performed by BiomeHub (Florianópolis, Santa Catarina, Brazil), using the same protocol (charite-berlin). Only samples with quantification cycle (Cq) below 30 for at least one primer were submitted to the SARS-CoV-2 genome sequencing. In total, 56 patients who presented symptoms such as fever, cough, sore throat, dyspnea, anosmia, fatigue, diarrhea, and vomiting (moderate and severe clinical status) [40] were included in the study.

4.2. RNA Extraction, Library Preparation, and Sequencing

Total RNAs were prepared as in the reference protocol [41] using SuperScript IV (Invitrogen, Carlsbad, CA, USA) for cDNA synthesis and Platinum Taq High Fidelity (Invitrogen, Carlsbad, CA, USA) for specific viral amplicons. Subsequently, cDNA was used for the library preparation with Nextera Flex (Illumina, San Diego, CA, USA) and quantified with Picogreen and Collibri Library Quantification Kit (Invitrogen, Carlsbad, CA, USA). The sequencing was performed on the Illumina MiSeq (Illumina, San Diego, CA, USA) 150 × 150 runs with 500xSARS-CoV-2 coverage (50–100 thousand reads/sample).

4.3. Quality Control and Consensus Calling

Quality control, reference mapping, and consensus calling were performed using an in-house pipeline developed by BiomeHub (Florianópolis, Santa Catarina, Brazil). Briefly, adapters were removed, and reads were trimmed by size = 150. Reads were mapped to the reference SARS-CoV-2 genome (GenBank accession number NC_045512.2) using Bowtie v2.4.2 (end-to-end and very-sensitive parameters) [42]. Mapping coverage and depth were retrieved using samtools v1.11 [43] (minimum base quality per base (Q) $\geq$ 30). Consensus sequences were generated using bcftools mpileup (Q $\geq$ 30; depth (d) $\leq$ 1000) combined with bcftools filter (DP > 50) and bcftools consensus v1.11 [44]. Coverage values for each genome were plotted using the karyoploteR v1.12.4 R package [45]. Finally, we assessed the consensus sequences quality using Nextclade v0.14.2 (https://clades.nextstrain.org/; accessed on 4 May 2021).

4.4. Mutation Analysis

SNPs and insertions/deletions in each sample were identified using snippy variant calling and core genome alignment pipeline v4.6.0 (https://github.com/tseemann/snippy; accessed on 4 May 2021), which uses FreeBayes v1.3.2 [46] to call variants and snpEff v5.0 [47] to annotate and predict their effects on genes and proteins. Genome map and

SNP histogram were generated after running MAFFT v7.475 [48] alignment using the msastats.py script, and plotAlignment and plotSNPHist functions [49]. Sequence positions refer to GenBank RefSeq sequence (NC_045512.2), isolated and sequenced from an early case from Wuhan (China) in 2019.

We identified global virus lineages using the dynamic nomenclature implemented in Pangolin v2.3.8 [50] (https://github.com/cov-lineages/pangolin; accessed on 4 May 2021) and global clades and mutations using Nextclade v0.14.2 (https://clades.nextstrain.org/; accessed on 4 May 2021). We also used Pathogenwatch (https://pathogen.watch/; accessed on 4 May 2021) and Microreact [51] to explore mutations and lineages across time and geography initially.

4.5. Maximum Likelihood Phylogenomic Analysis

All available SARS-CoV-2 genomes (1,048,519 sequences) were obtained from GISAID on April 26, 2021 and combined with our 56 sequences to obtain a global representative dataset. These sequences were subjected to analysis inside the NextStrain ncov pipeline [52] (https://github.com/nextstrain/ncov; accessed on 4 May 2021). In this workflow, sequences were aligned using Nextalign v0.1.6 (https://github.com/neherlab/nextalign; accessed on 4 May 2021). In the initial filtering step, short and low-quality sequences or those with incomplete sampling dates were excluded. Uninformative sites and ends (100 positions in the beginning and 50 in the end) were also masked from the alignment. Genetically closely related genomes to our focal subset were selected, prioritizing sequences geographically closer to Brazil's RS state. The maximum likelihood (ML) phylogenetic tree was built using IQ-TREE v2.1.2 [53], employing the general time-reversible (GTR) model with unequal rates and base frequencies [54]. The tree's root was placed between lineage A and B (Wuhan/WH01/2019 and Wuhan/Hu-1/2019 representatives), and sequences that deviate more than four interquartile ranges from the root-to-tip regression of genetic distances against sampling dates were removed from the analysis. A time-scaled ML tree was generated with TreeTime v0.8.1 [55] under a strict clock and a skyline coalescent prior with a mean rate of 8×10^{-4} substitutions per site per year. Finally, clades and mutations were assigned and geographic movements inferred. The results were exported to JSON format to enable interactive visualization through Auspice.

Additionally, as P.1 sequences mostly represent our dataset, we downloaded all complete and high-quality global genomes assigned to P.1 PANGO lineage (4499 sequences) submitted until 26 April 2021. These sequences were aligned using MAFFT v7.475, the ends of the alignment (300 in the beginning and 500 in the end) were masked, and the ML tree was built with IQ-TREE v2.0.3 using the GTR + F + R3 nucleotide substitution model as selected by the ModelFinder [56]. Branch support was calculated using the Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT) [57] with 1000 replicates.

Local sequences were classified according to the following scheme: monophyletic clades composed by one local genome were classified as "isolated", while clades composed by 2 < genomes < 4 were considered "clusters", and, if $\geq$ 4 local genomes were represented, we assigned a "clade" designation.

ML trees were inspected in TempEst v1.5.3 [58] to investigate the temporal signal through regression of root-to-tip genetic divergence against sampling dates. For the P.1 ML tree, samples with missing days of the collection were filled with the 15th day of the month. ML and time-resolved trees were visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/; accessed on 4 May 2021) and ggtree R package v2.0.4 [59].

4.6. Discrete Bayesian Phylogeographic and Phylodynamic Analysis

Considering the four identified clades composed of $\geq$4 sequences from this study, we extracted the clade members using the caper R package v1.0.1 [60]. Clade-specific ML trees and root-to-tip regression assignments were generated as described above. Evolutionary parameter estimates and spatial diffusion were assessed separately for each

clade using a Bayesian Markov Chain Monte-Carlo (MCMC) approach implemented in BEAST v10.4 [61]. The BEAGLE library [62] was used to enhance computational time. Time-stamped Bayesian trees were generated using the HKY + Γ nucleotide model [63], a strict molecular clock model with a Continuous Time Markov Chain (CTMC) rate reference prior [64] (mean rate = 8×10^{-4}) and a non-parametric skygrid tree prior [65] with grid points defined by the approximate number of weeks spanned by the duration of the phylogeny.

The MCMC chains were run in duplicates for at least 50 million generations, and convergence was checked using Tracer v1.7.1 [66]. Log and tree files were combined using LogCombiner v1.10.4 to ensure stationarity and good mixing after removing 10% as burn-in. Maximum clade credibility (MCC) was generated using TreeAnnotator v1.10.4 [61]. Viral migrations were reconstructed using a reversible discrete asymmetric phylogeographic model [67] to infer the locations of the internal nodes of the tree. A discretization scheme with a resolution of different Brazilian states and other countries was applied. Location diffusion rates were estimated using the Bayesian stochastic search variable selection (BSSVS) [67] procedure employing Bayes factors to identify well-supported rates.

4.7. Geoplotting

Geographical maps and general plots were generated using R v3.6.1 [68], and the ggplot2 v3.3.2 [69], geobr v.1.4 [70], and sf v0.9.8 [71] packages. For the discrete phylogeographic analysis, SpreaD3 v0.9.7.1 software [72] was used to map spatiotemporal information embedded in MCC trees.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/pathogens10080988/s1. Figure S1. Spatiotemporal distribution of the 56 sequenced samples from RS. Figure S2. Neighbouring countries, Brazilian divisions and RS intermediate regions. Figure S3. Proportion of the 10 most frequent lineages of SARS-CoV-2 across time in four Brazilian regions. Figure S4. ML tree augmenting the B.1.1.28 clade where the local sequence (RS-HBM-39491) was placed. Figure S5. ML tree augmenting the P.2 clade where the local sequence (RS-HBM-39486) was placed. Figure S6. Coverage depth plots for each genome sequenced. Figure S7. Expanded ML tree built using 4499 P.1 sequences deposited in GISAID until 26 April 2021. Tips represented by sequences sequenced in this study are augmented to enable visualization of the different introductions, clusters, and clades reported. Table S1. Collection of all mutations (n = 175) found in the 56 sequenced genomes, including effects on SARS-CoV-2 genes and proteins. The table is ordered by genomic position. Table S2. GISAID acknowledgment table of the 8635 global sequences used in the Nextstrain workflow. Table S3. GISAID acknowledgment table of the 4499 P.1 sequences analyzed. Supplementary File S1. Data and code used to reproduce the results presented.

Author Contributions: Conceptualization, R.A.Z. and C.E.T.; Data curation, V.B.F., C.P. (Christiano Perin), A.H. and C.P. (Camila Peter); Formal analysis, V.B.F., G.D.C., G.B.C. and C.E.T.; Investigation, V.B.F., G.D.C., C.P. (Christiano Perin), A.H., G.B.C., R.A.Z. and C.E.T.; Methodology, V.B.F., G.D.C., C.P. (Camila Peter), G.B.C., R.A.Z. and C.E.T.; Project administration, R.A.Z. and C.E.T.; Resources, C.P. (Christiano Perin), A.H., C.P. (Camila Peter), L.N.R., F.A.C., R.A.Z. and C.E.T.; Software, V.B.F.; Supervision, G.B.C. and C.E.T.; Validation, V.B.F., G.D.C. and G.B.C.; Visualization, V.B.F. and G.D.C.; Writing—original draft, V.B.F., G.D.C., R.A.Z. and C.E.T.; Writing—review and editing, V.B.F., G.D.C., C.P. (Christiano Perin), A.H., G.B.C., P.A.G.F., L.N.R., F.A.C., R.A.Z. and C.E.T. All authors have read and agreed to the published version of the manuscript.

Funding: The sequencing was supported by donations from Beppler & Puppi Advogados, Smellbox Produtos de Higiene Ltd.a., and Dr. Leonardo Mestre Negri. Scholarships and Fellowships were supplied by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES)—Finance Code 001 and Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA).

Institutional Review Board Statement: Ethical approval was obtained from the Comitê de Ética em Pesquisa em Seres Humanos da Universidade Federal de Ciências da Saúde de Porto Alegre (CEP-UFCSPA) under process number CAAE 35083220.2.0000.5345. The study was performed in accordance with the Declaration of Helsinki. All samples belonging to the Hospital da Brigada Militar patients that yielded positive RT-qPCR had their laboratory electronic records reviewed to compile metadata such as date of collection, sex, age, symptoms, exposure history, and clinical status, when available. Samples were anonymized before being received by the study investigators, following Brazilian and international ethical standards.

Informed Consent Statement: This study obtained a waiver of informed consent approved by Comitê de Ética em Pesquisa em Seres Humanos da Universidade Federal de Ciências da Saúde de Porto Alegre (CEP-UFCSPA) under process number CAAE 35083220.2.0000.5345.

Data Availability Statement: Full tables acknowledging the authors and corresponding labs submitting sequencing data used in this study can be found in Files S3 and S4. Consensus genomes generated in this study were deposited in the GISAID database under Accession IDs: EPI_ISL_2139494 to EPI_ISL_2139549. Data and code used to reproduce the results presented are available in Supplementary Materials.

Acknowledgments: We thank the administrators of the GISAID database and research groups across the world for supporting the rapid and transparent sharing of genomic data during the COVID-19 pandemic. We also thank the staff of Hospital da Brigada Militar, Laboratório Exame, Beppler & Puppi Advogados, Smellbox Produtos de Higiene Ltd.a., Leonardo Mestre Negri, and BiomeHub Pesquisa e Desenvolvimento who directly contributed to this study.

Conflicts of Interest: Dr. Cadegiani has served as a clinical director for Applied Biology, Inc. The other authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. World Health Organization. WHO Director-General's Opening Remarks at the Media Briefing on COVID-19—11 March 2020. Available online: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed on 27 May 2021).
2. Candido, D.; Claro, I.M.; de Jesus, J.G.; Souza, W.M.; Moreira, F.R.R.; Dellicour, S.; Mellan, T.A.; du Plessis, L.; Pereira, R.H.M.; Sales, F.C.S.; et al. Evolution and Epidemic Spread of SARS-CoV-2 in Brazil. *Science* **2020**, *369*, 1255–1260. [CrossRef] [PubMed]
3. IBGE (Brazilian Institute of Geography and Statistics) Rio Grande do Sul—Cidades e Estados. Available online: https://www.ibge.gov.br/cidades-e-estados/rs.html (accessed on 17 May 2021).
4. IBGE (Brazilian Institute of Geography and Statistics) Regiões Geográficas. Available online: https://www.ibge.gov.br/apps/regioes_geograficas/ (accessed on 17 May 2021).
5. Rio Grande do Sul Department of Health—SES-RS Confirmado o Primeiro Caso de Novo Coronavírus no Rio Grande do Sul. Available online: https://saude.rs.gov.br/confirmado-o-primeiro-caso-de-novo-coronavirus-no-rio-grande-do-sul (accessed on 24 November 2020).
6. Secretaria de Planejamento, Governança e Gestão—Governo do Estado do Rio Grande do Sul Cogestão Regional—Distanciamento Controlado. Available online: https://distanciamentocontrolado.rs.gov.br/ (accessed on 17 May 2021).
7. Mullen, J.L.; Tsueng, G.; Latif, A.A.; Alkuzweny, M.; Cano, M.; Haag, E.; Zhou, J.; Zeller, M.; Matteson, N.; Andersen, K.G.; et al. Outbreak.Info. Available online: https://outbreak.info (accessed on 19 July 2021).
8. Rambaut, A.; Loman, N.; Pybus, O.; Barclay, W.; Barrett, J.; Carabelli, A.; Connor, T.; Peacock, T.; Robertson, D.; Volz, E.; et al. Preliminary Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations. Available online: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (accessed on 4 January 2021).
9. Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E.J.; Msomi, N.; et al. Detection of a SARS-CoV-2 Variant of Concern in South Africa. *Nature* **2021**, *592*, 438–443. [CrossRef]
10. Faria, N.; Mellan, T.A.; Whittaker, C.; Claro, I.M.; Candido, D.D.S.; Mishra, S.; Crispim, M.A.E.; Sales, F.C.S.; Hawryluk, I.; McCrone, J.T.; et al. Genomics and Epidemiology of the P.1 SARS-CoV-2 Lineage in Manaus, Brazil. *Science* **2021**, *372*, 815–821. [CrossRef]
11. Naveca, F.; Nascimento, V.; Souza, V.; Corado, A.; Nascimento, F.; Silva, G.; Costa, Á.; Duarte, D.; Pessoa, K.; Mejía, M.; et al. COVID-19 Epidemic in the Brazilian State of Amazonas Was Driven by Long-Term Persistence of Endemic SARS-CoV-2 Lineages and the Recent Emergence of the New Variant of Concern P.1. Available online: https://www.researchsquare.com/article/rs-275494/v1 (accessed on 1 March 2021).

12. Dhar, M.S.; Marwal, R.; Radhakrishnan, V.S.; Ponnusamy, K.; Jolly, B.; Bhoyar, R.C.; Sardana, V.; Naushin, S.; Rophina, M.; Mellan, T.A.; et al. Genomic Characterization and Epidemiology of an Emerging SARS-CoV-2 Variant in Delhi, India. *medRxiv* **2021**, *6*, 21258076. [CrossRef]
13. Peacock, T.P.; Sheppard, C.M.; Brown, J.C.; Goonawardane, N.; Zhou, J.; Whiteley, M.; Consortium, P.V.; de Silva, T.I.; Barclay, W.S. The SARS-CoV-2 Variants Associated with Infections in India, B.1.617, Show Enhanced Spike Cleavage by Furin. *bioRxiv* **2021**, *5*, 446163. [CrossRef]
14. Buss, L.F.; Prete, C.A.; Abrahim, C.M.M.; Mendrone, A.; Salomon, T.; de Almeida-Neto, C.; França, R.F.O.; Belotti, M.C.; Carvalho, M.P.S.S.; Costa, A.G.; et al. Three-Quarters Attack Rate of SARS-CoV-2 in the Brazilian Amazon during a Largely Unmitigated Epidemic. *Science* **2021**, *371*, 288–292. [CrossRef]
15. Sabino, E.C.; Buss, L.F.; Carvalho, M.P.S.; Prete, C.A.; Crispim, M.A.E.; Fraiji, N.A.; Pereira, R.H.M.; Parag, K.V.; da Silva Peixoto, P.; Kraemer, M.U.; et al. Resurgence of COVID-19 in Manaus, Brazil, despite High Seroprevalence. *Lancet* **2021**, *397*, 452–455. [CrossRef]
16. Brazilian Ministry of Health Painel Coronavírus Brasil. Available online: https://covid.saude.gov.br/ (accessed on 17 May 2021).
17. Franceschi, V.B.; Ferrareze, P.A.G.; Zimerman, R.A.; Cybis, G.B.; Thompson, C.E. Mutation Hotspots, Geographical and Temporal Distribution of SARS-CoV-2 Lineages in Brazil, February 2020 to February 2021: Insights and Limitations from Uneven Sequencing Efforts. *medRxiv* **2021**, *8*, 21253152. [CrossRef]
18. Lamarca, A.P.; de Almeida, L.G.P.; Francisco, R. da S.; Lima, L.F.A.; Scortecci, K.C.; Perez, V.P.; Brustolini, O.J.; Sousa, E.S.S.; Secco, D.A.; Santos, A.M.G.; et al. Genomic Surveillance of SARS-CoV-2 Tracks Early Interstate Transmission of P.1 Lineage and Diversification within P.2 Clade in Brazil. *medRxiv* **2021**, *3*, 21253418. [CrossRef]
19. Rio Grande do Sul Health Surveillance Center Genomic Bulletin 5 (16/04/2021). Available online: https://coronavirus.rs.gov.br/upload/arquivos/202104/16173629-vigilancia-genomica-rs-boletim05-compactado.pdf (accessed on 20 April 2021).
20. Shu, Y.; McCauley, J. GISAID: Global Initiative on Sharing All Influenza Data—From Vision to Reality. *Eurosurveillance* **2017**, *22*, 30494. [CrossRef] [PubMed]
21. De Almeira, L.G.; Lamarca, A.P.; Francisco Junior, R.d.S.; Cavalcante, L.; Gerber, A.L.; Guimarães, A.P.d.C.; Machado, D.T.; Alves, C.; Mariani, D.; Cruz, T.F.; et al. Genomic Surveillance of SARS-CoV-2 in the State of Rio de Janeiro, Brazil: Technical Briefing—SARS-CoV-2 Coronavirus/NCoV-2019 Genomic Epidemiology. Available online: https://virological.org/t/genomic-surveillance-of-sars-cov-2-in-the-state-of-rio-de-janeiro-brazil-technical-briefing/683 (accessed on 4 May 2021).
22. Voloch, C.M.; Francisco, R.D.S.; de Almeida, L.G.P.; Cardoso, C.C.; Brustolini, O.J.; Gerber, A.L.; Guimarães, A.P.D.C.; Mariani, D.; da Costa, R.M.; Ferreira, O.C.; et al. Genomic Characterization of a Novel SARS-CoV-2 Lineage from Rio de Janeiro, Brazil. *J. Virol.* **2021**, *95*. [CrossRef] [PubMed]
23. Franceschi, V.B.; Caldana, G.D.; de Menezes Mayer, A.; Cybis, G.B.; Neves, C.A.M.; Ferrareze, P.A.G.; Demoliner, M.; de Almeida, P.R.; Gularte, J.S.; Hansen, A.W.; et al. Genomic Epidemiology of SARS-CoV-2 in Esteio, Rio Grande Do Sul, Brazil. *BMC Genom.* **2021**, *22*, 371. [CrossRef]
24. Francisco, R.D.S., Jr.; Benites, L.F.; Lamarca, A.P.; de Almeida, L.G.P.; Hansen, A.W.; Gularte, J.S.; Demoliner, M.; Gerber, A.L.; Guimarães, A.P.D.C.; Antunes, A.K.E.; et al. Pervasive Transmission of E484K and Emergence of VUI-NP13L with Evidence of SARS-CoV-2 Co-Infection Events by Two Different Lineages in Rio Grande Do Sul, Brazil. *Virus Res.* **2021**, *296*, 198345. [CrossRef] [PubMed]
25. Salvato, R.S.; Gregianini, T.S.; Campos, A.A.S.; Crescente, L.V.; Vallandro, M.J.; Ranieri, T.M.S.; Vizeu, S.; Martins, L.G.; da Silva, E.V.; Pedroso, E.R.; et al. Epidemiological Investigation Reveals Local Transmission of SARS-CoV-2 Lineage P.1 in Southern Brazil. *Rev. Epidemiol. E Controle Infecção* **2021**, *1*, 1–6. [CrossRef]
26. Soares da Silva, M.; Demoliner, M.; Hansen, A.; Gularte, J.; Silveira, F.; Heldt, F.; Filippi, M.; Pereira da Silva, F.; Mallmann, L.; Fink, P.; et al. Early Detection of SARS-CoV-2 P.1 Variant in Southern Brazil and Reinfection of the Same Patient by P.2. Available online: https://www.researchsquare.com (accessed on 14 May 2021).
27. Kubik, S.; Marques, A.C.; Xing, X.; Silvery, J.; Bertelli, C.; De Maio, F.; Pournaras, S.; Burr, T.; Duffourd, Y.; Siemens, H.; et al. Recommendations for Accurate Genotyping of SARS-CoV-2 Using Amplicon-Based Sequencing of Clinical Samples. *Clin. Microbiol. Infect.* **2021**, *27*, 1036. [CrossRef] [PubMed]
28. Walls, A.C.; Park, Y.-J.; Tortorici, M.A.; Wall, A.; McGuire, A.T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181*, 281–292. [CrossRef] [PubMed]
29. Deng, X.; Garcia-Knight, M.A.; Khalid, M.M.; Servellita, V.; Wang, C.; Morris, M.K.; Sotomayor-González, A.; Glasner, D.R.; Reyes, K.R.; Gliwa, A.S.; et al. Transmission, Infectivity, and Neutralization of a Spike L452R SARS-CoV-2 Variant. *Cell* **2021**, *184*, 3426–3437. [CrossRef]
30. Martin, D.P.; Weaver, S.; Tegally, H.; San, E.J.; Shank, S.D.; Wilkinson, E.; Giandhari, J.; Naidoo, S.; Pillay, Y.; Singh, L.; et al. The Emergence and Ongoing Convergent Evolution of the N501Y Lineages Coincides with a Major Global Shift in the SARS-CoV-2 Selective Landscape. *medRxiv* **2021**, *2*, 21252268. [CrossRef]
31. Resende, P.C.; Delatorre, E.; Gräf, T.; Mir, D.; Motta, F.C.; Appolinario, L.R.; da Paixão, A.C.D.; Mendonça, A.C.D.F.; Ogrzewalska, M.; Caetano, B.; et al. Evolutionary Dynamics and Dissemination Pattern of the SARS-CoV-2 Lineage B.1.1.33 During the Early Pandemic Phase in Brazil. *Front. Microbiol.* **2021**, *11*, 615280. [CrossRef]
32. Mir, D.; Rego, N.; Resende, P.C.; López-Tort, F.; Fernandez-Calero, T.; Noya, V.; Brandes, M.; Possi, T.; Arleo, M.; Reyes, N.; et al. Recurrent Dissemination of SARS-CoV-2 through the Uruguayan-Brazilian Border. *medRxiv* **2021**, *1*, 20249026. [CrossRef]

33. Naveca, F.; Nascimento, V.; Souza, V.; Corado, A.; Nascimento, F.; Silva, G.; Costa, Á.; Duarte, D.; Pessoa, K.; Gonçalves, L.; et al. Phylogenetic Relationship of SARS-CoV-2 Sequences from Amazonas with Emerging Brazilian Variants Harboring Mutations E484K and N501Y in the Spike Protein. Available online: https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585 (accessed on 24 February 2021).
34. Choi, B.; Choudhary, M.C.; Regan, J.; Sparks, J.A.; Padera, R.F.; Qiu, X.; Solomon, I.H.; Kuo, H.-H.; Boucau, J.; Bowman, K.; et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* **2020**, *383*, 2291–2293. [CrossRef]
35. Kemp, S.A.; Collier, D.A.; Datir, R.P.; Ferreira, I.A.T.M.; Gayed, S.; Jahun, A.; Hosmillo, M.; Rees-Spear, C.; Mlcochova, P.; Lumb, I.U.; et al. SARS-CoV-2 Evolution during Treatment of Chronic Infection. *Nature* **2021**, *592*, 277–282. [CrossRef] [PubMed]
36. Gräf, T.; Bello, G.; Venas, T.M.M.; Pereira, E.C.; Paixão, A.C.D.; Appolinario, L.R.; Lopes, R.S.; Mendonça, A.C.d.F.; da Rocha, A.S.B.; Motta, F.C.; et al. Identification of SARS-CoV-2 P.1-Related Lineages in Brazil Provides New Insights about the Mechanisms of Emergence of Variants of Concern—SARS-CoV-2 Coronavirus / NCoV-2019 Genomic Epidemiology. Available online: https://virological.org/t/identification-of-sars-cov-2-p-1-related-lineages-in-brazil-provides-new-insights-about-the-mechanisms-of-emergence-of-variants-of-concern/694/1 (accessed on 17 May 2021).
37. Plante, J.A.; Mitchell, B.M.; Plante, K.S.; Debbink, K.; Weaver, S.C.; Menachery, V.D. The Variant Gambit: COVID-19's next Move. *Cell Host Microbe* **2021**, *29*, 508–515. [CrossRef] [PubMed]
38. Gupta, S.; Mallick, D.; Banerjee, K.; Sarkar, S.; Lee, S.T.M.; Basuchowdhuri, P.; Jana, S.S. D155Y Substitution of SARS-CoV-2 ORF3a Weakens Binding with Caveolin-1: An in Silico Study. *bioRxiv* **2021**, *3*, 437194. [CrossRef]
39. Corman, V.M.; Landt, O.; Kaiser, M.; Molenkamp, R.; Meijer, A.; Chu, D.K.; Bleicker, T.; Brünink, S.; Schneider, J.; Schmidt, M.L.; et al. Detection of 2019 Novel Coronavirus (2019-NCoV) by Real-Time RT-PCR. *Eurosurveillance* **2020**, *25*, 2000045. [CrossRef]
40. World Health Organization COVID-19 Clinical Management: Living Guidance. Available online: https://www.who.int/publications-detail-redirect/WHO-2019-nCoV-clinical-2021-1 (accessed on 1 May 2021).
41. Eden, J.-S. SARS-CoV-2 Genome Sequencing Using Long Pooled Amplicons on Illumina Platforms. *bioRxiv* **2020**. [CrossRef]
42. Langmead, B.; Salzberg, S.L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
43. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
44. Li, H. A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics* **2011**, *27*, 2987–2993. [CrossRef]
45. Gel, B.; Serra, E. KaryoploteR: An R/Bioconductor Package to Plot Customizable Genomes Displaying Arbitrary Data. *Bioinforma. Oxf. Engl.* **2017**, *33*, 3088–3090. [CrossRef]
46. Garrison, E.; Marth, G. Haplotype-Based Variant Detection from Short-Read Sequencing. *arXiv* **2012**, arXiv:1207.3907. Available online: http://arxiv.org/abs/1207.3907 (accessed on 4 May 2021).
47. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff. *Fly* **2012**, *6*, 80–92. [CrossRef] [PubMed]
48. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]
49. Du Plessis, L. Laduplessis/SARS-CoV-2_Guangdong_Genomic_Epidemiology; Initial Release; Zenodo. Available online: https://zenodo.org/record/3922606 (accessed on 4 May 2021).
50. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G. A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology. *Nat. Microbiol.* **2020**, *5*, 1403–1407. [CrossRef] [PubMed]
51. Argimón, S.; Abudahab, K.; Goater, R.J.E.; Fedosejev, A.; Bhai, J.; Glasner, C.; Feil, E.J.; Holden, M.T.G.; Yeats, C.A.; Grundmann, H.; et al. Microreact: Visualizing and Sharing Data for Genomic Epidemiology and Phylogeography. *Microb. Genom.* **2016**, *2*, e000093. [CrossRef]
52. Hadfield, J.; Megill, C.; Bell, S.M.; Huddleston, J.; Potter, B.; Callender, C.; Sagulenko, P.; Bedford, T.; Neher, R.A. Nextstrain: Real-Time Tracking of Pathogen Evolution. *Bioinformatics* **2018**, *34*, 4121–4123. [CrossRef]
53. 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. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef]
54. Tavare, S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lect. Math. Life Sci.* **1986**, *17*, 57–86.
55. Sagulenko, P.; Puller, V.; Neher, R.A. TreeTime: Maximum-Likelihood Phylodynamic Analysis. *Virus Evol.* **2018**, *4*. [CrossRef]
56. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nat. Methods* **2017**, *14*, 587–589. [CrossRef]
57. Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321. [CrossRef] [PubMed]
58. Rambaut, A.; Lam, T.T.; Max Carvalho, L.; Pybus, O.G. Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (Formerly Path-O-Gen). *Virus Evol.* **2016**, *2*, vew007. [CrossRef]
59. Yu, G.; Smith, D.K.; Zhu, H.; Guan, Y.; Lam, T.T.-Y. Ggtree: An r Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data. *Methods Ecol. Evol.* **2017**, *8*, 28–36. [CrossRef]

60. Orme, D.; Freckleton, R.; Thomas, G.; Petzoldt, T.; Fritz, S.; Isaac, N.; Pearse, W. Caper: Comparative Analyses of Phylogenetics and Evolution in R (R Package Version 1.0.1). Available online: https://www.scienceopen.com/document?vid=d750bff2-a400-41dd-aa1e-728bb7aaf4d5 (accessed on 10 May 2021).
61. Suchard, M.A.; Lemey, P.; Baele, G.; Ayres, D.L.; Drummond, A.J.; Rambaut, A. Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10. *Virus Evol.* **2018**, *4*, vey016. [CrossRef]
62. Ayres, D.L.; Darling, A.; Zwickl, D.J.; Beerli, P.; Holder, M.T.; Lewis, P.O.; Huelsenbeck, J.P.; Ronquist, F.; Swofford, D.L.; Cummings, M.P.; et al. BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. *Syst. Biol.* **2012**, *61*, 170–173. [CrossRef] [PubMed]
63. Hasegawa, M.; Kishino, H.; Yano, T. Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. *J. Mol. Evol.* **1985**, *22*, 160–174. [CrossRef]
64. Ferreira, M.A.R.; Suchard, M.A. Bayesian Analysis of Elapsed Times in Continuous-Time Markov Chains. *Can. J. Stat.* **2008**, *36*, 355–368. [CrossRef]
65. Gill, M.S.; Lemey, P.; Faria, N.R.; Rambaut, A.; Shapiro, B.; Suchard, M.A. Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci. *Mol. Biol. Evol.* **2013**, *30*, 713–724. [CrossRef]
66. Rambaut, A.; Drummond, A.J.; Xie, D.; Baele, G.; Suchard, M.A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **2018**, *67*, 901–904. [CrossRef] [PubMed]
67. Lemey, P.; Rambaut, A.; Drummond, A.J.; Suchard, M.A. Bayesian Phylogeography Finds Its Roots. *PLoS Comput. Biol.* **2009**, *5*, e1000520. [CrossRef] [PubMed]
68. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
69. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2009; ISBN 978-0-387-98141-3.
70. Pereira, R.; Gonçalves, C.; De Araujo, P.; Carvalho, G.; De Arruda, R.; Nascimento, I.; Da Costa, B.; Cavedo, W.; Andrade, P.; Da Silva, A.; et al. Geobr: Loads Shapefiles of Official Spatial Data Sets of Brazil. Available online: https://github.com/ipeaGIT/geobr (accessed on 4 May 2021).
71. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector Data. *R. J.* **2018**, *10*, 439–446. [CrossRef]
72. Bielejec, F.; Baele, G.; Vrancken, B.; Suchard, M.A.; Rambaut, A.; Lemey, P. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Mol. Biol. Evol.* **2016**, *33*, 2167–2169. [CrossRef] [PubMed]

6. DISCUSSÃO

No presente estudo, realizamos a análise de genomas isolados de pacientes acometidos pela COVID-19 para a compreensão da distribuição de linhagens e padrão de espalhamento geográfico em nível municipal (Esteio, RS, Brasil), estadual (RS) e nacional (Brasil).

A seguir, serão apresentados estudos relevantes conduzidos na primeira fase da pandemia no Brasil. Em nosso estudo realizado em Esteio, entre maio e outubro de 2020, verificamos a provável dominância das linhagens B.1.1.33 e B.1.1.28, assim como ocorreu em território nacional (FRANCESCHI et al., 2021c; REDE GENÔMICA FIOCRUZ, 2021). Também evidenciamos a grande contribuição da região Sudeste na difusão de ambas as linhagens para outros estados e países, o que é consistente com a observação de que as metrópoles Rio de Janeiro e São Paulo eram os locais com maior risco de importações virais devido a sua conectividade aérea internacional e nacional (CANDIDO et al., 2020a). Nesse sentido, uma modelagem matemática sugeriu que 17 cidades foram responsáveis por 98-99% dos casos iniciais e 26 rodovias federais foram responsáveis por cerca de 30% da propagação do SARS-CoV-2 (NICOLELIS et al., 2021). Adicionalmente, trajetórias do centro geográfico da epidemia ao longo de 2020 demonstraram que, após a introdução em São Paulo, tanto os casos quanto as mortes se deslocaram progressivamente para o norte até meados de junho de 2020, quando voltaram a se deslocar para o sudeste (CASTRO et al., 2021).

Importantemente, conseguimos realizar, à época, a primeira detecção da linhagem P.2 no RS em outubro, em data próxima à sua detecção no Rio de Janeiro (FRANCESCHI et al., 2021a; FRANCISCO JUNIOR et al., 2021a). Por ter sido realizado em um pequeno município, fomos capazes de rastrear dois conjuntos de

mutações virais em pacientes epidemiologicamente relacionados, destacando a importância do rastreamento da disseminação viral em pequenas áreas da comunidade e a formação de sublinhagens.

Em uma investigação retrospectiva dos primeiros casos de cada um das 15 Coordenadorias Regionais de Saúde do RS, foi evidenciado o papel das introduções nacionais e internacionais para o estabelecimento da pandemia no estado. Foram detectadas as linhagens A (n=1), B.1 (33,3%) e B.1.1 (13,3%) relacionadas à diversidade viral inicial em outros países, e as linhagens B.1.1.28 (26,7%) e B.1.1.33 (10,0%), que, possivelmente, se originaram no Brasil. Adicionalmente, foram investigadas 75 amostras das quatro cidades com maior incidência cumulativa (Santo Ângelo, Passo Fundo, Caxias do Sul e Porto Alegre) durante os três picos epidêmicos do RS. Esta análise, embora com representatividade espacial e temporal limitada, indicou a predominância de B.1.1.33 (50%) e B.1.1.28 (35%) durante o primeiro pico epidêmico (julho a agosto de 2020), o surgimento e ampla distribuição de P.2 (55,6%) no segundo pico (novembro a dezembro de 2020) e disseminação massiva de P.1 e derivadas (P.1-*like*-II, P.1.1 e P.1.2; 78,4%) no terceiro pico (fevereiro a abril de 2021) (VARELA et al., 2021).

Em um estudo mais abrangente, que sequenciou 340 genomas de 33 cidades do RS, entre abril e novembro de 2020, foi constatada a presença de cinco linhagens principais: B.1.1.28 (n=117; 27,9%), P.2 (n = 14; 4,1%), B.1.1.33 (n = 183; 53,8%), B.1.91 (n = 9; 2,6%), e B.1.195 (n = 3; 0,9%). A linhagem B.1.1.33 teve maior frequência até o mês de outubro, seguida da B.1.1.28. Contudo, a partir de então as linhagens P.2 e VUI-NP13L (P.7) passaram a representar uma importante fonte de diversidade genética no território gaúcho. A VUI-NP13L foi identificada

dentro do *cluster* de B.1.1.28, sendo representada por 22 genomas com a substituição N:P13L, além de ORF3a:T151I e ORF9b:P10S. Apesar das primeiras amostras dessa nova linhagem no RS terem sido identificadas no final de agosto de 2020, análises filodinâmicas dataram a sua emergência em junho de 2020, seguida de espalhamento para outras regiões brasileiras e demais continentes (SANT'ANNA et al., 2021).

A linhagem B.1.1.33 foi estudada em maiores detalhes utilizando 190 genomas de 13 estados brasileiros, mostrando uma abundância variável em diferentes estados (variando de 2% em Pernambuco a 80% no Rio de Janeiro) e uma frequência moderada em países da América do Sul (5-18%). Surpreendentemente, esta linhagem foi primeiramente detectada no início de março em outros países americanos (por exemplo, Argentina, Canadá e EUA). Análises mais detalhadas sugerem a existência de uma linhagem intermediária (B.1.1.33-*like*) que, muito provavelmente, surgiu na Europa e foi posteriormente disseminada para o Brasil, onde sua disseminação deu origem à linhagem B.1.1.33 (RESENDE et al., 2021a), a qual possivelmente desencadeou surtos secundários na Argentina e no Uruguai (MIR et al., 2021; RESENDE et al., 2021a). A linhagem B.1.1.28, apesar de aparentemente menos abundante que B.1.1.33 em várias regiões brasileiras (FRANCESCHI et al., 2021c), rapidamente se diversificou em duas novas variantes: a VOC P.1 e a VOI P.2.

O surgimento da VOI P.2, derivada de B.1.1.28, carregando a mutação S:E484K foi datado, em um estudo retrospectivo, no final de fevereiro de 2020 no Sudeste (São Paulo e Rio de Janeiro), seguida pela transmissão para o Sul (especialmente RS). A partir de então, foram observadas múltiplas rotas de dispersão entre os estados brasileiros, especialmente no final de 2020 e início de

2021 (LAMARCA et al., 2021b). Entretanto, esta linhagem não foi relatada até outubro de 2020, quando foi detectada pela primeira vez no estado do Rio de Janeiro (VOLOCH et al., 2021) e no pequeno município de Esteio, no RS (FRANCESCHI et al., 2021a). Além disso, o aumento da frequência de B.1.1.28 e linhagens derivadas foi corroborado por outro estudo que incluiu amostras de vários municípios do RS em novembro de 2020 e constatou que 86% dos genomas foram classificados como B.1.1.28 e ~50% destes pertenciam à nova linhagem P.2 (FRANCISCO JUNIOR et al., 2021a).

Após a primeira onda de casos em Manaus (Amazonas), dominada pelas linhagens B.1.195 e B.1.1.28, a segunda onda coincidiu com o surgimento da VOC P.1. Esta variante, provavelmente, evoluiu de um clado local composto por sequências B.1.1.28 no final de novembro de 2020 e substituiu sua linhagem parental em menos de dois meses (NAVECA et al., 2021a). Intermediários evolutivos entre B.1.1.28 e P.1 (denominados P.1-*like I* e P.1-*like*-II) foram identificados, sugerindo que a diversidade de variantes do SARS-CoV-2 que abrigam mutações da proteína *Spike* em Manaus pode ser maior do que inicialmente esperado e que essas variantes provavelmente circularam e se diversificaram por algum tempo antes do surgimento e expansão da P.1 (GRÄF et al., 2021a; NAVECA et al., 2021a).

Outro estudo que acompanhou a evolução da linhagem P.1 também estimou seu surgimento em novembro de 2020, precedido por um período de evolução molecular acelerada. Além disso, demonstrou que as migrações do vírus entre o Amazonas e as metrópoles urbanas do Sudeste do Brasil seguem padrões de mobilidade aérea nacional, uma vez que os estados que reportaram a identificação da P.1 até o final de fevereiro de 2021 receberam cerca de 100.000 passageiros

aéreos de Manaus em novembro. Dentre as 10 novas mutações não-sinônimas na proteína *Spike* (L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, T1027I) comparadas a seu ancestral imediato (B.1.1.28), as análises de seleção positiva encontraram evidências de que oito delas estejam sob seleção positiva (FARIA et al., 2021).

No início de 2021, o Brasil se tornou novamente o epicentro[52] da pandemia de COVID-19 no mundo. Por exemplo, no Rio Grande do Sul, um crescimento exponencial de casos, hospitalizações e mortes por COVID-19 ocorreu em fevereiro de 2021. Das 27 amostras sequenciadas em fevereiro de 2021 no Hospital de Clínicas de Porto Alegre, 24 (88,9%) foram classificadas como linhagem P.1, contra uma (6,7%) de 15 amostras obtidas em janeiro de 2021, gerando preocupações quanto a uma possível associação entre a linhagem P.1 e o rápido crescimento dos casos e internações no estado (MARTINS et al., 2021).

Em nosso estudo realizado em Março de 2021 – período de altas dramáticas no número de casos, hospitalizações e mortes por COVID-19 no estado – com 56 amostras de 11 municípios da região metropolitana (n=51) e dois da região intermediária de Santa Maria (n=5) foi possível confirmar a altíssima frequência da P.1. Além de descrever a frequência das principais linhagens no Brasil, no RS e na Argentina durante a pandemia, verificamos que 96,4% (n=54) das amostras recém geradas pertenciam à linhagem P.1 e aproximadamente 20% (n=11) dessas caracterizavam uma nova sublinhagem derivada da P.1 que abrigava três mutações na ORF1ab (synC1912T, D762G e T1820I), uma na ORF3a (D155Y) e uma na proteína N (synC28789T), a qual havia sido recentemente detectada no RJ e países estrangeiros (DE ALMEIDA et al., 2021). As análises filogenéticas indicaram a

[52] Centro da pandemia, lugar com maior número de casos em determinado momento.

presença de 4 clados, 5 *clusters* e 13 sequências isoladas (*singletons*). Além disso, como as sequências deste estudo foram acomodadas em ramos distintos da filogenia com mutações definidoras, supôs-se que múltiplas introduções no RS ocorreram, o que foi confirmado pela análise filogeográfica. Destaca-se o surgimento da P.1.2 (clado 1), possivelmente, no RS em meados de dezembro de 2020 e a difusão dos três demais clados principalmente em janeiro de 2020.

Inicialmente, verificamos que nossas sequências eram mais proximamente relacionadas àquelas de fora do *cluster* do RJ, pois não carregavam a mutação S:A262S. Quando realizamos a análise, havia apenas 93 sequências de três estados brasileiros (RJ, RS e SP) e cinco outros países com esse conjunto de mutações disponíveis no GISAID, sendo a mais antiga do RS (município turístico de Gramado), onde a emergência dessa linhagem foi estimada em meados de dezembro de 2020 (FRANCESCHI et al., 2021b). Contudo, uma análise filogeográfica mais recente com mais genomas disponíveis (11 estados brasileiros e 10 países), estimou sua primeira divergência em SP entre o fim de 2020 e início de 2021, muito embora essa linhagem tenha se espalhado rapidamente, alcançando boa parte dos estados brasileiros e demais países amostrados no final de janeiro de 2021 (FRANCISCO JUNIOR et al., 2021b). Essas observações demonstram que a ausência de coordenação nacional de vigilância genômica e atrasos em sequenciamento e compartilhamento de dados podem dificultar inferências precisas sobre datas e locais de emergência de novas linhagens.

Em um estudo semelhante, 71 amostras de 16 municípios de três regiões do RS (especialmente região metropolitana) no período de meados de dezembro de 2020 a final de abril de 2021 foram sequenciadas. Como resultado, a VOC P.1 (Gama) foi a mais frequente (67,14%, n=47), seguida pela VOI P.2 (Zeta) (27,14%,

n=20) e B.1.1.28 (5,71%, n=4) considerando todo o período. Realizando uma estratificação mensal, a P.2 predominou durante janeiro e fevereiro de 2021, enquanto a P.1 passou a se estabelecer fortemente entre março e abril (DEMOLINER et al., 2021), corroborando os dados de vigilância genômica (CENTRO ESTADUAL DE VIGILÂNCIA EM SAÚDE, 2021) e nosso estudo.

Uma análise filogenômica realizada na segunda maior metrópole brasileira (Rio de Janeiro) entre dezembro de 2020 a maio de 2021 (MOREIRA et al., 2021), demonstrou, utilizando 244 isolados de todas as semanas epidemiológicas deste período, que a P.1 (Gama) tem circulado localmente desde as primeiras semanas de 2021 e apenas sete semanas foram necessárias para que ela atingisse uma frequência acima de 70%, de modo similar ao detectado em Manaus. Utilizando métodos filogeográficos Bayesianos, foram sugeridos, pelo menos, 13 introduções no município vindas de várias regiões brasileiras, bem como se demonstrou maior carga viral em pacientes infectados pela variante, reforçando seu aumento de transmissibilidade e distinto comportamento epidemiológico anteriormente demonstrado (FARIA et al., 2021; NAVECA et al., 2021a).

Ao realizar o sequenciamento de 185 amostras de três das cinco regiões brasileiras, incluindo Amazonas (região Norte), Rio Grande do Norte, Paraíba e Bahia (região Nordeste), e Rio de Janeiro (região Sudeste), um estudo demonstrou a ampla dispersão das linhagens P.1 e P.2 pelas regiões brasileiras. P.2 foi a linhagem predominante identificada no país, exceto em Manaus. Esse estudo também identificou amostras mais recentes da linhagem P.2, estimando sua origem em fevereiro de 2020 e sua divisão em novos clados. A transmissão interestadual da P.2 foi detectada desde março, mas atingiu seu pico entre dezembro de 2020 e janeiro de 2021 (LAMARCA et al., 2021b).

Finalmente, utilizando 1.188 genomas (0,6% dos casos diagnosticados) do estado do Amazonas entre 01 de janeiro e 06 de julho de 2021, observou-se um aumento acentuado na frequência de sublinhagens de P.1 que abrigam dois tipos de substituições adicionais na proteína *Spike*: deleções no domínio N-terminal (NTD) (particularmente Δ141-144) ou mutações na junção S1/S2 (N679K e P681H/R). Estas novas linhagens derivadas da P.1 (designadas como P.1.3 a P.1.8) somaram quase a totalidade de casos positivos sequenciados no estado do Amazonas em julho de 2021 e algumas surgiram ou se espalharam para outros estados brasileiros. Portanto, esses resultados sugerem que a VOC P.1 não teria atingido seu pico evolutivo e que seu espalhamento está relacionado à contínua evolução viral para sublinhagens que podem ser ainda mais transmissíveis, uma vez que a carga viral foi em média 6 vezes maior. Apesar de tais deleções estarem associadas ao escape de anticorpos anti-NTD ou aumento de afinidade ou clivagem pelo sítio polibásico de furinas, a ausência de superrepresentação em vacinados infectados (*breakthrough*) sugere que essas novas sublinhagens não sejam mais eficientes para evadir a resposta imunológica elicitada por vacinas quando comparadas à linhagem parental (NAVECA et al., 2021b).

Análises filogeográficas demonstraram o papel essencial do Sudeste para a difusão viral no Brasil, contribuindo com 40% dos movimentos virais, seguido pela região Norte (25%) (GIOVANETTI et al., 2021). Em relação à Gama, a difusão se deu a partir da região Norte no final de 2020, com amplificação das transições no início de 2021 a partir da região Sudeste. Já a P.2 teve seu fluxo concentrado principalmente nas regiões Sul e Sudeste, com *clusters* locais ocorrendo a partir de outubro de 2020 (GIOVANETTI et al., 2021; GRÄF et al., 2021b). Importantemente, o potencial de aumento de transmissibilidade foi confirmado pela divisão das

estimativas de número reprodutivo efetivo (Re) de Gama e P.2 em seus epicentros (estados do Amazonas e Rio de Janeiro, respectivamente). Como o epicentro da P.2 é a região sudeste mais bem conectada e esta linhagem demonstrou menores taxas de difusão do que a Gama, que surgiu na região norte historicamente mais isolada, é provável que a dinâmica espaço-temporal diferente entre estas linhagens tenha sido moldada, principalmente, pela maior transmissibilidade intrínseca da Gama (GRÄF et al., 2021b).

Em nosso estudo realizado com a utilização de cerca de 2700 genomas do SARS-CoV-2 durante o primeiro ano da pandemia no Brasil (Fevereiro 2020-2021), revelou heterogeneidades temporais e geográficas no sequenciamento do SARS-CoV-2 no território brasileiro. Temporalmente, os esforços foram concentrados especialmente na primeira fase da epidemia (maio e abril de 2020), ficando abaixo de 1% dos casos até o fim do período investigado. Geograficamente, houve representatividade desproporcionalmente maior do Sudeste, Nordeste e Sul, enquanto as regiões Norte e Centro-Oeste contribuíram com cerca de 12,5% das sequências somadas. Ainda, estados com maior taxa de sequenciamento, identificaram maior número de linhagens circulantes. Adicionalmente, foram confirmadas as sucessivas substituições das linhagens B.1.1.28 e B.1.1.33 por P.2 e P.1 que caracterizaram as duas primeiras ondas epidêmicas. Finalmente, demonstramos padrões filogeográficos complexos de espalhamento viral, de modo que alguns clados identificados permaneceram mais restritos geograficamente, demonstrando maior papel da difusão intra-estadual, e outros estiveram mais difusos em várias regiões por espalhamento interestadual e inter-regional, provavelmente por rodovias e aeroportos. Muito embora a quantidade de amostras tenha crescido para 78.366 sequências brasileiras no início de dezembro de 2021,

sendo 91,9% (n=72.040) com data de coleta de 2021, os dois maiores estados do Sudeste (SP e RJ) sequenciaram juntos ~65% desses genomas, e registraram ~26% dos casos brasileiros de COVID-19 do país (SHU; MCCAULEY, 2017). Portanto, o sequenciamento no Brasil continua a ser temporalmente e geograficamente enviesado.

Além do surgimento e espalhamento de novas linhagens do SARS-CoV-2 no território nacional, a introdução de VOCs oriundas de outras nações representaram um capítulo importante na história evolutiva do SARS-CoV-2 no Brasil. Em dezembro de 2020, a linhagem B.1.1.7 (Alfa) foi detectada em dois pacientes no estado de São Paulo, e após o rastreio de contatos e análises filogenéticas, verificou-se que se tratavam de duas introduções separadas, bem como foi confirmada a hipótese de transmissão comunitária em curso (CLARO et al., 2021). Também em São Paulo, 217 sequências genômicas completas foram obtidas dos maiores departamentos regionais de saúde em março, documentando a primeira introdução da linhagem B.1.351 (Beta) no Brasil e demonstrando a maior frequência da P.1 (Gama) (64,05%), seguida por B.1.1.28 (25,34%) e B.1.1.7 (5,99%) (SLAVOV et al., 2021). A lenta progressão das variantes Alfa e Beta após suas introduções no território brasileiro sugere, portanto, uma vantagem competitiva da VOC Gama.

Nos primeiros meses de 2021, a linhagem B.1.617.2 (Delta), originária da Índia, espalhou-se rapidamente pelo mundo e substituiu VOCs já amplamente distribuídas, como a Alfa e Beta, em alguns países em poucos meses (MULLEN et al., 2021b). No Brasil, a VOC Delta foi detectada inicialmente em membros da tripulação de um navio que viajou da Malásia ao Maranhão (Nordeste brasileiro), passando pela África do Sul, em maio de 2021 (DOS SANTOS et al., 2021).

Subsequentemente, o primeiro aumento de casos associado à variante Delta foi detectado no estado do Rio de Janeiro, o qual possui um programa de vigilância genômica bem estabelecido. Em meados de junho de 2021, 0,57% dos genomas (n=2) sequenciados foram classificados como Delta. Cerca de um mês após, entre junho e julho, a VOC Delta alcançou uma frequência de 61,8%, enquanto em meados de agosto já representava 89,2% das sequências do estado (DE ALMEIDA et al., 2021). Análises filogenéticas e filodinâmicas indicam a ocorrência de pelo menos seis introduções e a formação de três clados principais dessa variante no Brasil (LAMARCA et al., 2021a). Em meados de dezembro de 2021, apenas um mês após sua possível emergência no continente africano, a variante Ômicron já registrava 19 casos (regiões Sudeste, Sul e Centro-Oeste) (REDE GENÔMICA FIOCRUZ, 2021), mas acreditava-se que já havia transmissão comunitária no país. Em meados de janeiro de 2022, apesar da defasagem de cerca de um mês entre a coleta e o sequenciamento das amostras, a Ômicron já representa cerca de 20% dos genomas depositados de dezembro de 2021 (REDE GENÔMICA FIOCRUZ, 2021), e os casos confirmados sobem de modo expressivo no Brasil, muito embora o patamar de hospitalizações mantenha-se sob controle devido ao avanço da vacinação, aos altos níveis de infecção natural e à potencial menor severidade da nova variante.

Muito embora o estado do Amazonas tenha presenciado o surgimento e a diversificação da variante P.1 (Gama) em novas sublinhagens com mutações adicionais no RBD e NTD, um estudo mais recente, incluindo amostras entre 01 de julho a 15 de outubro de 2021 e utilizando 1132 genomas do estado (4,5% dos casos confirmados), demonstrou uma redução acentuada da frequência de Gama em detrimento do aumento de Delta. Essa VOC dominante mundialmente

representava 1% dos casos sequenciados no Amazonas em julho de 2021, enquanto em outubro já correspondia a 89%. Contudo, essa substituição foi acompanhada de uma redução do número de casos, muito provavelmente devido aos altos níveis de imunidade natural e ao avanço da vacinação no estado (NAVECA et al., 2021c).

Em dois estudos independentes (FERRAREZE et al., 2021; RESENDE et al., 2021c), foram detectadas, entre os meses de novembro de 2020 e fevereiro de 2021, uma nova variante de interesse (N.9) descendente da linhagem B.1.1.33 que também abriga a mutação S:E484K, bem como três mutações adicionais na ORF1ab (NSP3:A1711V, NSP6:F36L e NS7b:E33A). Sua emergência foi estimada em agosto de 2020, já sendo encontrada no início de 2021 em quatro regiões brasileiras. Esses achados demonstram que a mutação E484K emergiu quase simultaneamente e de modo independente nas duas linhagens brasileiras mais frequentes na primeira fase da epidemia (B.1.1.28 e B.1.1.33) (RESENDE et al., 2021c). Adicionalmente, foi detectada em dois estados brasileiros (Amapá e Maranhão) outra linhagem derivada de B.1.1.33 (N.10), porém com 14 mudanças genéticas definidoras, incluindo V445A e E484K no RBD e várias mutações não-sinônimas (P9L, I210V e L212I), e três deleções (∆141-144, ∆211 e ∆256-258) no NTD da proteína *Spike.* Também foi identificada a truncagem da ORF7b devido a uma deleção de *frame-shifting* (RESENDE et al., 2021b). Além das mutações de preocupação no RBD, as deleções encontradas na linhagem N.10 estão localizadas dentro ou próximas a regiões de deleções recorrentes que compõem o super sítio antigênico do NTD, podendo conferir resistência à neutralização por anticorpos anti-NTD (CERUTTI et al., 2021; RESENDE et al., 2021b).

Uma característica marcante das VOCs é que muitas delas adquirem um alto número de mutações definidoras em um curto período de tempo. Embora as estimativas filogenéticas atuais da taxa evolutiva do SARS-CoV-2 sugerem que seu genoma acumula cerca de 2 mutações por mês, as VOCs podem possuir até 15 mutações definidoras. Portanto, podem emergir ao longo de alguns meses, implicando que a taxa evolutiva seria algumas vezes maior.

Ao analisar desde modelos de relógio molecular mais simples (*e. g., strict clock*) até mais complexos (*e. g. uncorrelated relaxed clock* e *fixed local clock*), e verificar os mais adequados para a estimativa da taxa evolutiva das VOCs, foi evidenciado que o surgimento dessas variantes é impulsionado por um aumento episódico na taxa evolutiva de cerca de quatro vezes em relação à taxa filogenética de ramos externos (*background*) (TAY et al., 2021). Atualmente, cinco cenários podem ser considerados plausíveis para que tal evolução molecular acelerada ocorra: (i) evolução intra-hospedeiro[53] em pacientes imunocomprometidos cronicamente infectados; (ii) adaptação a novos hospedeiros animais; (iii) recombinação entre diferentes linhagens circulantes; (iv) evolução intra-hospedeiro em *superspreads*; e (v) evolução gradual em regiões com baixa vigilância genômica.

Um primeiro cenário, já comprovado por uma série de estudos (AVANZATO et al., 2020; BAZYKIN et al., 2021; CHOI et al., 2020; KEMP et al., 2021; TRUONG et al., 2021), demonstra que infecções crônicas podem acelerar a evolução viral e reduzir a sensibilidade aos anticorpos neutralizantes em indivíduos imunocomprometidos. Isso ocorre pois pacientes imunocomprometidos possuem deficiências no sistema imunológico que o tornam incapazes de eliminar a infecção,

[53] Acúmulo de mutações virais que ocorrem durante a replicação viral dentro do hospedeiro.

fornecendo espaço para o surgimento de mutações associadas a escape imunológico[54], especialmente as localizadas na proteína *Spike*. Além disso, o tratamento com plasma convalescente[55] foi associado a forte pressão seletiva sobre o SARS-CoV-2, sendo associado ao surgimento de variantes virais que apresentam menor susceptibilidade à neutralização por anticorpos nestes indivíduos. Mutações identificadas nestes pacientes durante sucessivos pontos no tempo da sua infecção persistente e tratamento com plasma convalescente incluem as regiões do NTD (Δ69-70, ΔY144, e Δ157-158), do super sítio da NTD, do RBD (K417N, E484K, N501Y) e do sítio de clivagem de furinas (P681H/R) (COREY et al., 2021), os quais possuem vários epítopos antigênicos[56] e são alvos das principais classes de anticorpos. Como estas substituições são associadas à VOCs e VOIs, e como tais variantes adquirem mutações na *Spike* a uma taxa muito mais rápida do que o esperado, especula-se que este fenômeno seja determinante para a emergência de linhagens mais transmissíveis e capazes de evadir o sistema imune.

Outro aspecto importante foi relatado em uma análise genômica de um paciente com linfoma[57] sofrendo de COVID-19 crônica. Durante os quatro meses da doença, foram observadas 18 novas mutações, incluindo S:Y453F e Δ69-70 (combinação ΔF), anteriormente associadas a *clusters* detectados em *minks* (EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL, 2020;

[54] Ocorre quando o o sistema imunológico do hospedeiro não é mais capaz de reconhecer e eliminar um patógeno.

[55] Utilização do sangue possivelmente rico em anticorpos de pacientes recuperados de uma doença para tentar ajudar outras pessoas a se recuperarem.

[56] Área da molécula do antígeno (substância que é capaz de estimular uma resposta imunológica) que se liga aos receptores celulares e aos anticorpos.

[57] Tipo de câncer que se origina no sistema linfático, o qual é composto por um conjunto de órgãos (linfonodos ou gânglios) e tecidos que produzem as células responsáveis pela imunidade.

KOOPMANS, 2021; ORESHKOVA et al., 2020). Contudo, a análise filogenética indica que a linhagem deste paciente não está relacionada com tais *clusters*, indicando que estas mutações foram adquiridas independentemente. Como ambas as mutações são encontradas em frequências intermediárias no paciente, destaca-se o papel da evolução intra-hospedeiro. Portanto, a aquisição independente de um par idêntico de mutações em um *mink* e em um paciente com linfoma, bem como entre múltiplos pacientes imunossuprimidos, sugere convergência evolutiva (BAZYKIN et al., 2021). Além disso, como a transmissão do SARS-CoV-2 entre *minks* tem sido associada à aquisição recorrente de uma série de mutações, esta adaptação a novos hospedeiros pode estar vinculada ao surgimento de novas variantes.

Uma terceira, e menos provável, via para o surgimento das VOCs está relacionada a eventos de recombinação entre as linhagens circulantes do SARS-CoV-2. Embora este seja um fenômeno bastante recorrente entre os *Betacoronavirus* (DUDAS; RAMBAUT, 2016; HON et al., 2008; LAI et al., 1985), apenas um estudo até o momento demonstrou a existência de múltiplos recombinantes amostrados no Reino Unido entre o final de 2020 e início de 2021. Tais eventos estiveram relacionados à linhagem B.1.1.7 e outras linhagens circulantes no mesmo período. Em quatro casos houve transmissão comunitária dos vírus recombinantes, incluindo um *cluster* de 45 casos sequenciados ao longo de dois meses. Contudo, acredita-se que estes vírus foram extintos e não continuam circulando local ou globalmente (JACKSON et al., 2021). Em uma análise de novembro de 2021, demonstrou-se que um *cluster* de B.1.628 originou-se por recombinação entre as linhagens B.1.631 e B.1.634. Importantemente, esse evento de recombinação é suportado pela distribuição espaço-temporal destas três

linhagens, as quais co-circularam no México e nos Estados Unidos por cerca de nove meses (dezembro de 2020 a agosto de 2021) (GUTIERREZ et al., 2021).

Outro fenômeno relevante é a presença de *superspreads*, que foi evidenciada em estudo de vigilância genômica do Rio de Janeiro (FRANCISCO JUNIOR et al., 2021b), consistente com achados anteriores (YANG et al., 2021). Estes representavam a grande maioria dos vírus circulantes (>90%) entre os indivíduos sintomáticos no estado, podendo desempenhar um papel importante na dinâmica de disseminação do vírus. O aumento da taxa de replicação viral nestes indivíduos pode, portanto, colaborar para o acelerado surgimento de novas mutações e a consequente diversificação em novas linhagens (FRANCISCO JUNIOR et al., 2021b).

Uma última possibilidade é a de que tais variantes podem ter evoluído gradativamente em partes do mundo onde há pouca ou inexistente vigilância genômica, mas circulação viral generalizada, fator que pode maximizar a ação da seleção natural gerando variantes com alta capacidade de escape imunológico (OUDE MUNNINK et al., 2021).

Estruturalmente, a proteína *Spike* (S) possui 1273 aminoácidos. Apresenta-se na forma de homotrímero[58], estando presente na superfície viral do SARS-CoV-2 e sendo composta de duas subunidades funcionais, S1 e S2. Cada protômero[59] de S possui um domínio S1 (resíduos 1-686), o qual abriga principalmente o domínio de ligação do receptor (RBD) na porção distal – responsável pela ligação do vírus na superfície celular – e o domínio N-Terminal (NTD) na região proximal, também crítico para as propriedades de ligação. Complementarmente, o domínio

[58] Produto de reação de três moléculas idênticas cujas subunidades também são iguais.

[59] Unidade estrutural de uma proteína oligomérica (com número finito de pequenas moléculas).

S2 (resíduos 687-1273) desempenha a função de fusão das membranas celulares e virais (HUANG et al., 2020b; WALLS et al., 2020; WRAPP et al., 2020).

A estrutura da proteína S foi determinada por microscopia crioeletrônica[60] a nível atômico, revelando diferentes conformações do domínio RBD nos estados aberto (para cima) e fechado (para baixo), bem como suas funções correspondentes (WALLS et al., 2020; WRAPP et al., 2020). No estado nativo, a proteína existe como um precursor inativo. Durante a infecção viral, as proteases da célula-alvo ativam a proteína S, clivando-a nas subunidades S1 e S2 utilizando a serinoprotease TMPRSS2 como *primer* (BERTRAM et al., 2013). Adicionalmente, o RBD adquire uma conformação aberta (para cima) acessível ao receptor, o que é necessário para ativar o domínio de fusão da membrana após a entrada viral nas células alvo (HOFFMANN et al., 2020; LAN et al., 2020; V'KOVSKI et al., 2021).

Substituições individuais constantemente sofrem pressão seletiva, podendo apresentar maior ou menor aptidão (*fitness*) para interação com a célula hospedeira por meio de alterações conformacionais, replicação, transmissão, evasão imune, entre outras características. Utilizando a metodologia de cristalografia de raio-X[61] para resolver a estrutura do RBD da proteína S de SARS-CoV-2 ligada ao receptor celular hACE2, demonstrou-se que o modo de ligação entre ambos é quase idêntico em relação ao RBD de SARS-CoV. Adicionalmente, a maioria dos resíduos essenciais para a ligação à hACE2 são altamente conservados ou compartilham propriedades semelhantes de cadeia lateral ao SARS-CoV, indicando a

[60] Tipo de microscopia eletrônica de transmissão na qual a amostra biológica é estudada a temperaturas muito baixas, permitindo desde a análise da estrutura de proteínas de membrana até biomoléculas maiores e mais complexas.

[61] Técnica que consiste em fazer passar um feixe de raios-X através de um cristal da substância em estudo, permitindo a determinação da estrutura dos cristais ou moléculas por meio do fenômeno da difração.

possibilidade de evolução convergente entre os RBDs do SARS-CoV-2 e do SARS-CoV para uma melhor ligação à hACE2 (LAN et al., 2020).

Devido ao seu papel chave para a ligação do vírus à célula hospedeira, acredita-se que mutações na proteína S, especialmente no RBD, possam conferir mudanças conformacionais que aumentem a afinidade ao receptor, confiram incompatibilidade com epítopos antigênicos (evasão imune) e possam conferir maior transmissibilidade e letalidade. Tratam-se de forças seletivas independentes que impulsionam a diversidade genética viral. Muito embora as substituições mais preocupantes na proteína S seriam aquelas que simultaneamente aumentassem a transmissão, a gravidade da doença e conferissem evasão imune, análises estruturais apontaram para a ausência de tal combinação nas variantes B.1.1.7 e B.1.351 (CAI et al., 2021).

A seguir, são apresentadas as principais mutações e deleções observadas na proteína *spike* do SARS-CoV-2, bem como suas principais características (Tabela 4). A primeira mutação de relevância na proteína S do SARS-CoV-2 (D614G) surgiu no final de janeiro de 2020, e rapidamente se tornou dominante em praticamente todas as sequências mundiais derivadas da linhagem B.1 (KORBER et al., 2020). D614G é uma substituição de ácido aspártico por glicina no aminoácido 614 da subunidade S1. Apesar de não ocorrer perto do RBD e não modificar diretamente a afinidade de ligação ao hACE2, tal mutação interrompe um ou mais contatos interprotoméricos[62], resultando em uma maior probabilidade de que um (ou mais) dos três RBDs estejam em um aberto (para cima) em detrimento à posição fechada (para baixo) e, portanto, compatível com uma ligação aprimorada à hACE2 (YURKOVETSKIY et al., 2020) e podendo resultar em taxas de replicação

[62] Entre os protômeros (ver [59]) de uma proteína oligomérica.

Tabela 4. Resumo das principais mutações e deleções observadas na proteína *spike* do SARS-CoV-2, representando as principais linhagens que as apresentam, sua localização e relevância, principais características e perfis de resistência à neutralização.

Mutação *spike* / Características	**D614G**	**E484K**	**N501Y**	**L452R**	**P681H/R**	**Deleções no NTD**
Linhagens que a apresentam	Linhagens derivadas da B.1 que começaram a predominar desde o início de 2020 globalmente	VOCs (Beta e Gama) e VOIs (Eta, Iota, Mu, Theta e Zeta)	VOCs (Alfa, Beta, Gama e Omicron) e VOIs (Theta e Mu)	VOC Delta and VOIs (Kappa e Épsilon)	VOCs: Alfa e Omicron (P681H) e Delta (P681R)	Comum em VOCs, VOIs e pacientes com infecções crônicas
Localização e relevância	Fora da RBD, mas aumenta a probabilidade de que um ou mais protômeros de S estejam em estado aberto	Troca de carga em um *loop* flexível, aumentando afinidade com hACE2	Um dos seis principais resíduos do RBD que interagem com hACE2	Sem contato direto com hACE2, mas forma uma mancha hidrofóbica na superfície do RBD	Região importante para a infecciosidade, virulência e potencial pandêmico	Essencialmente localizadas no supersítio antigênico do NTD
Principais características	Aumento na expressão em células pulmonares e nas cargas virais	Importante escape imune em indivíduos vacinados e convalescentes	Aumento na afinidade ao receptor e replicação viral	Estabiliza a ligação com hACE2 e pode promover a entrada de um maior número de partículas virais	Aumento nas cargas positivas na interface RBD-ACE2 e na clivagem na junção S1/S2	Associadas com maiores cargas virais e infectividade (Δ69-70), mas especialmente confere evasão imune
Perfil de resistência	Igualmente suscetível à neutralização por anticorpos neutralizantes, não sendo determinante para gravidade da doença	Adquirida de forma convergente, e localizado dentro de um epítopo reconhecido por muitos anticorpos	Incapaz de alterar significativamente a ligação e a neutralização	Resistência moderada à neutralização	Não há indícios de que modifique a neutralização de modo significativo	Elevada resistência à neutralização

Fonte: Adaptado de TAO et al. (2021).

em células epiteliais pulmonares humanas (PLANTE et al., 2020) e cargas virais mais elevadas (KORBER et al., 2020; VOLZ et al., 2021a). D614G também parece estar associada com um aumento no número de proteínas *Spike* por vírion (ZHANG et al., 2020) e na taxa de clivagem na região S1/S2 (GOBEIL et al., 2021b). Entretanto, os mutantes G614 são similarmente (ou até mais) suscetíveis à neutralização imunológica do que a variante original D614 (WEISSMAN et al., 2021; YURKOVETSKIY et al., 2020), de modo que tal substituição não foi considerada uma determinante para a gravidade/severidade da COVID-19 (KORBER et al., 2020; VOLZ et al., 2021a).

Atualmente, VOCs e VOIs do SARS-CoV-2 em circulação compartilham várias mutações que permitem que o vírus se espalhe mesmo com o aumento da imunidade da população, enquanto mantém ou aumenta sua capacidade de replicação (TAO et al., 2021). A substituição do ácido glutâmico (E) por lisina (K) na posição 484 do RBD (E484K) resulta em uma mudança de carga em um *loop* flexível, formando um par iônico favorável contactando o aminoácido 75 de hACE2 (NELSON et al., 2021). Além de potencialmente aumentar a afinidade ao receptor, essa mutação tem sido considerada a mais preocupante, uma vez que está associada a escape imunológico a anticorpos neutralizantes de indivíduos recuperados e vacinados (BAUM et al., 2020; GREANEY et al., 2021a; WANG et al., 2021a) pela sua localização dentro de um epítopo imunodominante[63] reconhecido por muitos anticorpos neutralizantes (GOBEIL et al., 2021a). Sua vantagem evolutiva e imunológica fica evidente à luz da evolução convergente, uma

[63] Subunidades do antígeno que são mais facilmente reconhecidas pelo sistema imunológico e, portanto, influenciam mais a especificidade do anticorpo induzido.

vez que grande parte das VOCs (Beta e Gama) e VOIs (Eta, Iota, Mu, Theta e Zeta) carrega tal substituição (MULLEN et al., 2021c). Neste trabalho, fomos capazes de detectar a substituição E484K em sua fase inicial de espalhamento no RS carregada pela linhagem P.2 (Zeta) (FRANCESCHI et al., 2021a), bem como amplamente distribuída no início de 2021 na linhagem P.1 (Gama) (FRANCESCHI et al., 2021b).

A substituição de asparagina para uma tirosina na posição 501 (N501Y) localizada no RBD da proteína S começou a causar preocupação após sua emergência nas VOCs Alfa e Beta. Trata-se de um dos seis principais resíduos de contato do RBD interagindo com a hACE2 e foi associado com o aumento da afinidade ao receptor (GU et al., 2020; STARR et al., 2020), bem como na replicação viral nas células do trato respiratório superior de humanos e *hamsters* (LIU et al., 2021d). Utilizando métodos de dinâmica molecular, verificou-se um aumento nas interações eletrostáticas devido à formação de uma forte ligação de hidrogênio entre a S:T500 e o ACE2-D355 próximo ao local da mutação (ALI; KASRY; AMIN, 2021). Adicionalmente, a convergência evolutiva de múltiplas VOCs (Alfa, Beta, Gama e Ômicron) e VOIs (Theta e Mu) (MULLEN et al., 2021c) reforça que esta substituição possa estar relacionada a um aumento de transmissibilidade (LEUNG et al., 2021). Contudo, não é capaz de alterar significativamente a ligação e a neutralização pela maioria dos anticorpos (WANG et al., 2021a; WEISBLUM et al., 2020).

Apesar do resíduo L452 não contatar diretamente com o receptor hACE2, a substituição por arginina (R) forma juntamente com os resíduos F490 e L492 uma mancha hidrofóbica[64] na superfície do RBD, estabilizando a interação entre a

[64] Um grupo acessível de átomos apolares vizinhos

proteína S e o receptor hACE2 e podendo promover a entrada de um maior número de partículas virais em organóides[65] de pulmão (DENG et al., 2021). Além disso, as linhagens portadoras da mutação L452R parecem apresentar uma resistência moderada à neutralização por anticorpos elicitados por infecção prévia (4 a 6,7 vezes) ou vacinação (2 vezes) (DENG et al., 2021; GREANEY et al., 2021b). Importantemente, a mutação L452R está presente nas VOIs Kappa e Epsilon, bem como na VOC Delta (MULLEN et al., 2021c), sendo esta última a variante mais transmissível até o momento, capaz de suplantar outras VOCs e alcançar dominância global rapidamente (MULLEN et al., 2021b). Contudo, acredita-se que a combinação de mutações presente, e não somente L452R, confira tal vantagem evolutiva a esta VOC.

Diferentes mutações no sítio K417 surgiram em diferentes VOCs, sendo K417N em Beta e Ômicron e K417T em Gama. Contudo, K417N/T raramente ocorre isoladamente, sendo majoritariamente acompanhada por outras substituições no RBD (TAO et al., 2021). Uma possível causa para este fenômeno é seu efeito na redução da ligação à hACE2 (GREANEY et al., 2020; WANG et al., 2021c), necessitando de mutações compensatórias que aumentem o *fitness* viral. K417N é capaz de reduzir significativamente a susceptibilidade a alguns anticorpos monoclonais[66] (etesevimab e casirivimab), mas se mantém suscetível a outros (bamlanivimab, imdevimab e sotrovimab). Importantemente, K417N/T retém suscetibilidade completa a amostras de plasma de pacientes previamente

[65] Versão miniaturizada e simplificada de um órgão produzido *in vitro* em três dimensões e com anatomia realista.

[66] Anticorpos produzidos por um único clone de um único linfócito B parental, que é clonado e imortalizado, produzindo sempre os mesmos anticorpos em resposta a um agente patogênico.

infectados ou imunizados com pelo menos uma dose de vacinas de mRNA (TAO et al., 2021; WANG et al., 2021a, 2021c).

N439K é uma substituição que causou preocupação no Reino Unido devido ao seu aumento significativo de frequência em duas linhagens independentes (B.1.141 e B.1.258) até setembro de 2020, o que foi rapidamente dissipado pela emergência da VOC Alfa (THOMSON et al., 2021). Foi demonstrado que esta mutação aumenta a afinidade da proteína S ao receptor hACE2 (STARR et al., 2020; THOMSON et al., 2021) e adiciona uma ponte salina na interface RBD-ACE2, muito embora não aumente o *fitness* viral e a severidade da doença (THOMSON et al., 2021). Apesar de N439K conferir resistência a alguns anticorpos monoclonais e respostas policlonais[67] de indivíduos recuperados, os níveis de redução de neutralização são baixos (~2 vezes) (GREANEY et al., 2021a; THOMSON et al., 2021).

Múltiplas mutações localizadas em posições adjacentes ao sítio de clivagem de furinas – região importante para a infectividade, virulência e potencial pandêmico do SARS-CoV-2 (HOFFMANN; KLEINE-WEBER; PÖHLMANN, 2020) –, incluindo Q675H/R, Q677H/P, N679K e P681H/R, emergiram independentemente em variantes do SARS-CoV-2 (HODCROFT et al., 2021). A mutação P681H surgiu inicialmente na VOC Alfa no Reino Unido, mas também de modo independente em Ômicron e em diferentes VOIs (*e.g.,* Theta e Mu) (LASEK-NESSELQUIST et al., 2021b; MULLEN et al., 2021c). Outra substituição no mesmo sítio (P681R) emergiu na VOC Delta e na VOI Kappa (MULLEN et al., 2021c). O aumento de cargas positivas associadas na interface RBD-ACE2 tanto em P681H quanto em P681R

[67] Respostas provenientes de anticorpos que são originados de diferentes linfócitos B, cada um reconhecendo um epítopo (ver [56]) diferente.

parece influenciar o tropismo do vírus aumentando a clivagem na região S1/S2 em células epiteliais das vias aéreas humanas (JOHNSON et al., 2021a; PEACOCK et al., 2021b; SAITO et al., 2021). De um modo especial, P681R parece diminuir a infectividade viral ao passo que confere maior resistência a anticorpos neutralizantes (SAITO et al., 2021). Além de facilitar e acelerar a clivagem da proteína S por furinas, está associada a um aumento e aceleração na fusão célula-célula por TMPRSS2 e no espalhamento viral *in vitro*, os quais podem estar associados com maior transmissibilidade e patogenicidade (PEACOCK et al., 2021b; SAITO et al., 2021).

Deleções na região NTD da proteína S emergiram em inúmeras VOCs e VOIs, bem como em pacientes com infecções prolongadas (AVANZATO et al., 2020; BAZYKIN et al., 2021; CHOI et al., 2020; KEMP et al., 2021; TRUONG et al., 2021). A principal delas detectada até o momento é Δ69-70 —presente em Alfa e Ômicron—, a qual parece alterar um *loop* exposto na região do NTD e está associada com aumento na replicação viral e infectividade (KEMP et al., 2021; MENG et al., 2021a), enquanto Δ144 (presente em Alfa), Δ241-243 (Beta) e Δ157-158 (Delta) conferem resistência à neutralização por anticorpos direcionados ao super sítio antigênico do NTD (CHI et al., 2020; PLANAS et al., 2021b; WANG et al., 2021a). De um modo geral, as maiores evidências de evasão imune por deleções e mutações em NTD estão localizadas na região de um epítopo conformacional[68] constituído pelos resíduos 140-156 (*loop* N3) e 246-260 (*loop* N5) (CHI et al., 2020).

[68] Sequência de aminoácidos que forma uma estrutura tridimensional e compõe um antígeno que entra em contato direto com um receptor do sistema imunológico.

A região S1, onde ficam localizados o RBD e o NTD, exibe maior variabilidade de aminoácidos em relação a S2 entre os coronavírus relacionados à SARS. Até junho de 2021, 42 mutações na proteína S tinham uma frequência global de ≥1%, incluindo 15 na NTD, seis no RBD, cinco no domínio carboxi-terminal[69] de S1 (CTD) e nove na região S2 (TAO et al., 2021). Sendo assim, torna-se importante avaliar sítios em S1 associados a eventos de seleção positiva.

Contudo, métodos tradicionais para captura de sinais de seleção positiva não são direcionados para a análise em um curto período evolutivo e em populações virais densamente amostradas, uma vez que requerem a fixação de mutações não sinônimas ao longo de um tempo evolutivo de anos ou décadas. Portanto, ao aplicar um novo método que correlaciona o sucesso de clados utilizando regressão logística[70] com a acumulação de modificações não sinônimas em certos genes, demonstrou-se que a região S1 da proteína S apresenta a mais forte correlação entre estas características em relação a outras regiões do genoma do SARS-CoV-2. Adicionalmente, observou-se que a taxa dN/dS dentro de S1 evoluiu de 0,76 em 2020 para 1,85 em 2021, alcançando 2,07 em meados de 2021. Ao avaliar homoplasias que expandiram-se em clados bem sucedidos, algumas mutações foram associadas a taxas de crescimento aceleradas, como S:95I e S:452R (presentes na VOC Delta) e ORF1a:Δ3675-3677 (presente em Alfa, Beta, Gama e Lambda). Esses resultados demonstram que S1 é o principal *locus* de adaptação, mas outras mutações positivamente selecionadas em outros genes também influenciam a evolução do SARS-CoV-2 (KISTLER; HUDDLESTON;

[69] Uma das extremidades da cadeia polipeptídica, a qual apresenta um grupamento carboxila (COOH).

[70] Técnica estatística que objetiva produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores assumidos por uma variável categórica.

BEDFORD, 2021). Importantemente, mutações na proteína N, como R203K e G204R foram associadas a um aumento na fosforilação[71] do nucleocapsídeo, sugerindo que a troca do motivo RG pode estar associada a maior replicação e patogênese do SARS-CoV-2 (JOHNSON et al., 2021b). Sendo assim, acredita-se que mutações fora da proteína S — por exemplo nos genes N, ORF3a e ORF6 —, estejam envolvidas em diferentes estágios do ciclo de vida viral e sejam componentes importantes para a interação e contínua adaptação do SARS-CoV-2 ao hospedeiro humano (TIMMERS et al., 2021).

Variantes são classificadas de acordo com sua combinação de mutações e características epidemiológicas particulares. A designação de VOCs e VOIs foi criada para categorizar tais variantes baseado em sua transmissibilidade, capacidade de causar doença mais severa ou evasão do sistema imunológico (WORLD HEALTH ORGANIZATION, 2021d). Por aumento de transmissibilidade entende-se a capacidade de suplantar outras linhagens virais e apresentar maiores taxas de reprodução efetiva e de ataque secundário[72] quando comparada com outras variantes (KORBER et al., 2020; PENG et al., 2021; VOLZ et al., 2021b). Mudanças na severidade têm sido investigadas utilizando dados de morbidade ou hospitalização devido à variante (DAVIES et al., 2021b). Finalmente, a evasão às respostas imunes é realizada avaliando a susceptibilidade da variante a anticorpos monoclonais, plasma convalescente e plasma de vacinadas quando comparada a outras linhagens (CHEN et al., 2021; HARVEY et al., 2021).

[71] Adição de um grupo fosfato (PO_4) a uma proteína ou outra molécula, compondo um dos principais mecanismos de regulação das proteínas.

[72] Probabilidade de que uma infecção ocorra entre pessoas suscetíveis dentro de um período de incubação razoável após contato conhecido com uma pessoa infecciosa em ambientes domésticos ou de contato próximo.

A seguir, são apresentadas as principais características moleculares, epidemiológicas e imunológicas das VOCs (Tabela 5). A linhagem B.1.1.7 (VOC Alfa), a qual apresenta especialmente as mutações N50Y e P681H, também carrega as deleções Δ69-70 e Δ144 (RAMBAUT et al., 2020b), bem como substituições características em outros genes (N:R203K e G204R). Tal variante expandiu-se rapidamente, correspondendo à maioria das infecções na Europa e nos EUA no final de 2021 (MULLEN et al., 2021a). O aumento de transmissibilidade e letalidade em relação às demais variantes previamente circulantes foi estimado em 50% por estudos epidemiológicos e filodinâmicos (DAVIES et al., 2021b; VOLZ et al., 2021b). Apesar das alterações epidemiológicas, essa VOC mantém a susceptibilidade a anticorpos monoclonais e plasma de indivíduos recuperados (CHEN et al., 2021; PLANAS et al., 2021a; WANG et al., 2021a), não sendo associada a risco superior de reinfecções[73] (GRAHAM et al., 2021). Importantemente, a vacina Pfizer/Biontech (BNT162b) demonstrou cerca de 90% de eficácia contra esta variante (ABU-RADDAD; CHEMAITELLY; BUTT, 2021; HAAS et al., 2021).

A linhagem B.1.351 (VOC Beta) apresenta três mutações na região do RBD (K417N, E484K and N501Y), bem como cinco substituições no NTD (principalmente Δ242-244). Entre o final de 2020 e início de 2021, os casos de COVID-19 na África do Sul aumentaram cerca de 10 vezes, apesar da alta exposição prévia da população ao vírus. Nesse contexto, estimou-se que a VOC Beta seria 50% mais transmissível do que suas antecessoras (TEGALLY et al., 2021). Ao contrário de Alfa, essa VOC está fortemente associada a eventos de reinfecção (SHINDE et al., 2021) e apresenta suscetibilidade reduzida (de cerca de 10 vezes) a anticorpos

[73] Nova infecção ocorrida após recuperação da anterior e causada pelo mesmo agente infeccioso.

Tabela 5. Resumo das principais mutações e impactos epidemiológicos (transmissibilidade e morbidade) e imunológicos (perfil de resistência e eficácia vacinal) das VOCs.

VOC / Características	B.1.1.7 (Alfa)	B.1.351 (Beta)	P.1 (Gama)	B.1.617.2 (Delta)	B.1.1.529 (Ômicron)
Principais mutações	N50Y, P681H, Δ69-70 e Δ144	K417N, E484K, N501Y e Δ242-244	E484K, N501Y, K417T e deleções no NTD	L452R, T478K e P681R	Δ69-70, T95I, G142D, Δ143–145, K417N, T478K, N501Y, H655Y, N679K e P681H
Transmissibilidade e morbidade	Aumento de 50% na transmissibilidade e letalidade quando comparada a variantes anteriores	Aumento de 50% na transmissibilidade quando comparada a variantes anteriores	1,7 a 2,4 vezes mais transmissível, 21 a 46% capaz de reinfectar e 1,2 a 1,6 vezes mais letal	Suplantou as demais VOCs entre o início e meio de 2021, provavelmente devido a seu aumento de transmissibilidade	Risco de reinfecção 5 vezes maior e maior transmissibilidade; severidade reduzida entre 20 e 45% (em comparação à Delta) e menor replicação em células pulmorares
Perfil de resistência	Permaneceu altamente suscetível ao plasma de vacinados e convalescentes	Mais reinfecções, e 10 vezes menor susceptibilidade à neutralização por anticorpos e plasma convalescente	Perfil de resistência semelhante à VOC Beta	Perfil de resistência relevante: Neutralização de 3 a 10 vezes reduzida em 45% dos indivíduos; e mais de 10 vezes em 5% dos pacientes	Redução dramática na neutralização de vacinados e convalescentes, parcialmente recuperada com doses de reforço
Eficácia vacinal contra infecção	~90% (BNT162b)	Variável em diferentes estudos: 10% (AZD1222; n=39) e 75% (BNT162b)	Varia entre 54% (CoronaVac) e 70% (AZD1222)	Varia entre 60% (AZD1222) e 85% (BNT162b)	Baixa eficácia contra doença sintomática após duas doses (BNT162b2), porém proteção significativa (~75%) após dose de reforço

Fonte: Adaptado de TAO et al. (2021).

neutralizantes e plasma convalescente, especialmente devido à interferência das mutações K417N e E484K na ligação a algumas classes de anticorpos monoclonais (HOFFMANN et al., 2021; WANG et al., 2021a). Estimação da eficácia vacinal variou amplamente, alcançando o mínimo de 10% para AZD1222 em um pequeno estudo (n=39) (MADHI et al., 2021) e o máximo de 75% para BNT162b (ABU-RADDAD; CHEMAITELLY; BUTT, 2021).

A linhagem P.1 (VOC Gama) apresenta três mutações nos mesmos sítios de Beta (E484K, N501Y e K417T) e outras cinco mutações na NTD (FARIA et al., 2021). Sua emergência na região Amazônica, cuja soroprevalência[74] estimada alcançou ~75% anteriormente (BUSS et al., 2020; SABINO et al., 2021), demonstrou suas características epidemiológicas alteradas. Estimou-se que seja 1,7 a 2,4 vezes mais transmissível, 21 a 46% mais capaz de evadir o sistema imune de indivíduos previamente infectados (causar reinfecção) e 1,2 a 1,6 vezes mais letal (FARIA et al., 2021). Esta VOC apresenta perfil de resistência similar à Beta, com uma importante parcela (20 a 60%) dos indivíduos recuperados ou vacinados (com vacinas de mRNA ou AZD1222) apresentando redução de neutralização de 3 a 10 vezes, enquanto 5 a 10% englobam declínio maior que 10 vezes (CHEN et al., 2021; DEJNIRATTISAI et al., 2021; HOFFMANN et al., 2021; SOUZA et al., 2021; WANG et al., 2021b).

A linhagem B.1.617.2 (VOC Delta) estabeleceu-se na Índia, possuindo principalmente as mutações L452R, T478K, P681R, assim como outras na ORF3, ORF7a e N (MULLEN et al., 2021b; TAO et al., 2021). Caracterizou-se por ser a linhagem mais transmissível detectada até o final de 2021, uma vez que suplantou

[74] Número total de pessoas (ou porcentagem) em uma população que apresenta resultados positivos para uma doença em determinado momento com base em amostras de soro sanguíneo.

as variantes Alfa, Beta e Gama em seus países de emergência e predominância (MISHRA et al., 2021; MULLEN et al., 2021b; REDE GENÔMICA FIOCRUZ, 2021), e tornou-se predominante mundialmente. Apresenta perfil de resistência relevante, alcançando redução de 3 a 10 vezes em 45% dos indivíduos e 10 vezes em 5% deles (LIU et al., 2021a, 2021c; PLANAS et al., 2021b). Dados do Reino Unido apontam para eficácias vacinais de 60% (AZD1222) a 85% (BNT162b) (LOPEZ BERNAL et al., 2021; SHEIKH et al., 2021).

A linhagem B.1.1.529/BA.1 (Ômicron) apresenta cerca de 30 mutações na proteína S, muitas delas compartilhadas com outras VOCs e VOIs (*e. g.*, Δ69-70, T95I, G142D, Δ143–145, K417N, T478K, N501Y, H655Y, N679K, e P681H) e muitas novas (MULLEN et al., 2021c). Somada a este perfil mutacional preocupante, esteve associada a aumentos explosivos no número de casos na África do Sul, Reino Unido, entre outros países. Uma análise de mais de 65 mil casos de reinfecção na África do Sul sugere que tal variante possui uma capacidade substancial de causar maiores taxas de reinfecção quando comparada à Beta e Delta (PULLIAM et al., 2021). Na Inglaterra, o aumento no risco de reinfecção foi de 5,41 vezes (intervalo de confiança de 95%: 4,87-6,00) (FERGUSON et al., 2021).

Na Inglaterra, foi observado um rápido crescimento da variante Ômicron em relação à Delta entre o fim de novembro e início de dezembro de 2021 com alta taxa de crescimento exponencial e duplicação no número estimado de casos a cada dois dias. Além disso, pessoas jovens (8 a 29 anos), residentes de Londres e com etnia africana apresentaram taxas mais elevadas de infecção pela nova variante (FERGUSON et al., 2021). Contudo, estimou-se uma redução na severidade da

Ômicron em relação à Delta de 20-45% no país (FERGUSON et al., 2021b), consistente com dados da África do Sul (WOLTER et al., 2021).

Apesar da variante Ômicron possuir mutações que poderiam favorecer a clivagem na região S1/S2, a eficiência de clivagem observada foi substancialmente menor que em Delta (MENG et al., 2021b). Nesse sentido, utilizando cultivo de vírus vivos e pseudovírus[75], identificou-se uma via de entrada alterada que favorece a fusão endossomal[76] dependente de catepsina[77] no endossomo em oposição à via de fusão da superfície celular TMPRSS2-dependente (WILLETT et al., 2021). Outro fator que pode explicar a reduzida severidade da infecção por Ômicron é a menor replicação desta em células de pulmão expressando TMPRSS2 em comparação à Delta (MENG et al., 2021b).

A redução na neutralização foi confirmada em ensaios sorológicos, com redução de eficácia em convalescentes e recipientes de duas doses de vacina, a qual foi parcialmente recuperada com doses de reforço (*booster*) (CARREÑO et al., 2021; MENG et al., 2021b; WILLETT et al., 2021). Importantemente, apesar do extensivo escape imunológico, este foi incompleto em pacientes previamente infectados e vacinados (CELE et al., 2021). Muito embora apresente susceptibilidade reduzida a anticorpos neutralizantes, outros componentes da resposta imune adaptativa como células T CD4 e CD8 contribuem para a proteção contra doença grave. A resposta às células T foi mantida em 70-80% dos casos de

[75] Modelos de vírus engenheirados que não contêm o genoma do vírus, mas são capazes de mimetizar a forma como o vírus infecta as células, por meio da substituição das estruturas virais de interesse.

[76] Compartimento formado a partir do processo de endocitose (pelo qual células vivas ativamente absorvem moléculas e outras células), por meio de fusão de vesículas provenientes de organelas celulares como membrana plasmática, complexo de Golgi e lisossomos.

[77] Família de proteases (ver [12]) de cisteína que são ativadas sob o pH ácido dos lisossomos.

Ômicron investigados, demonstrando que tais respostas induzidas por vacinação ou infecção natural reconhecem essa variante (KEETON et al., 2021).

Uma taxa inicialmente alta de mutações é consistente com a ideia de que, logo após um evento de *spillover*, há muitas substituições acessíveis que são vantajosas evolutivamente no novo hospedeiro (KISTLER; HUDDLESTON; BEDFORD, 2021). Isto foi observado no vírus da pandemia de influenza H1N1 (H1N1pdm). Dado que durante dois anos após seu surgimento o H1N1pdm apresentou altas taxas de dN/dS em todo o genoma, acredita-se que sua evolução tenha sido amplamente direcionada pelo aumento de transmissibilidade e adaptação ao novo hospedeiro. Depois desse período inicial, sua evolução passou a ser dominada por mudanças antigênicas (SU et al., 2015). É possível que o SARS-CoV-2 esteja seguindo uma trajetória evolutiva semelhante, com adaptação inicial ao hospedeiro seguida por deriva antigênica[78] (KISTLER; HUDDLESTON; BEDFORD, 2021), principalmente após a introdução das vacinas, que limitam a circulação viral, mas ao mesmo tempo aumentam a probabilidade de aquisição de mutações de escape ao sistema imunológico. Sendo assim, torna-se essencial a coordenação entre a detecção e a caracterização fenotípica do SARS-CoV-2 para reduzir os casos e mortes mundialmente, bem como fornecer atualizações para as estratégias terapêuticas existentes.

Apesar dos avanços científicos sem precedentes para a compreensão da evolução do SARS-CoV-2, alguns problemas e desafios são recorrentemente observados nas análises genômicas que utilizam este vírus como objeto de estudo.

[78] Acúmulo de substituições de aminoácidos em proteínas virais selecionadas pelo sistema imune adaptativo do hospedeiro à medida que o vírus circula em uma população, o qual pode limitar substancialmente a duração da imunidade conferida pela infecção e pela vacinação.

Em um estudo que avaliou as dificuldades intrínsecas de inferir e pós-processar árvores filogenéticas do SARS-CoV-2 baseadas em ML, demonstraram-se inúmeras nuances advindas da baixa diversidade genética do vírus (poucas mutações) aliada à grande quantidade de sequências, o que caracteriza um fraco sinal filogenético. Com relação ao algoritmo de busca da árvore com maior ML, foram observadas diferenças topológicas importantes, de modo que seria mais adequado computar estatísticas sumárias em relação a um “conjunto mais plausível de árvores”, o que é realizado de modo natural por métodos Bayesianos, do que apresentar uma única árvore resultante da análise. Adicionalmente, foram observados problemas numéricos referentes à otimização do comprimento dos ramos e ao modelo de taxas livres. Com relação ao pós-processamento, enraizar árvores com *outgroups* de morcegos ou pangolins ou usando modelos de evolução não produziram uma posição razoável para a raiz. Portanto, muito embora a aplicação de métodos filogenéticos para a compreensão da evolução e propagação da COVID-19 fornece *insights* importantes, seus resultados devem ser interpretados com extrema cautela (MOREL et al., 2021).

Além da investigação da origem do SARS-CoV-2 a partir dos seus possíveis reservatórios animais e dados epidemiológicos do início da pandemia, outros métodos têm sido usados para estimar confiavelmente a raiz da filogenia e o primeiro caso (KUMAR et al., 2021; PEKAR et al., 2021a; PIPES et al., 2021). Devido ao rápido acúmulo de mutações nos vírus, torna-se desafiador identificar uma sequência *outgroup* que esteja suficientemente relacionada com as sequências em análise para permitir um enraizamento confiável (PIPES et al., 2021). Uma alternativa e esta prática é o enraizamento por relógio molecular, que se baseia na suposição de que as mutações ocorrem a uma taxa aproximadamente

constante (ZUCKERKANDL; PAULING, 1962), ou variam em diferentes ramos da filogenia (DRUMMOND et al., 2006).

Ao investigar seis diferentes variações destas duas estratégias de enraizamento de árvores e apresentar métricas relacionadas à incerteza, demonstrou-se que métodos baseados em grupo externo (*outgroup)* tendem a posicionar a raiz em sequências da linhagem A, enquanto abordagens de relógio molecular localizam a raiz na linhagem B. A explicação mais provável para a discrepância observada entre as raízes seria a hipermutabilidade[79] nos sítios que definem estas linhagens, resultando em um excesso de “mutações para trás” e sugerindo que o enraizamento por relógio molecular é mais confiável. Entretanto, não é possível excluir uma maior taxa de mutação (ou erros de sequenciamento) no clado A que atrairia a raiz para si. O enraizamento no clado B é compatível com uma origem em Wuhan, enquanto que um enraizamento na clado A sugere origens alternativas do vírus, possivelmente fora do leste asiático. Portanto, dadas as dificuldades de conciliação entre métodos, recomenda-se evitar inferências exageradas sobre a divergência inicial do SARS-CoV-2 com base em um enraizamento fixo em A ou B, e que análises baseadas no enraizamento por *outgroup* devem ser evitadas até que sejam descobertas sequências mais estreitamente relacionados. Dessa forma, é improvável que as evidências filogenéticas disponíveis identifiquem a origem do vírus SARS-CoV-2 de modo independente de outras fontes.

Outra questão primordial para o entendimento da origem e dinâmica inicial do SARS-CoV-2 é a datação do MRCA da filogenia e do primeiro caso (*index*). Ao combinar inferências retrospectivas baseadas em relógio molecular com

[79] Capacidade de exibir um número excessivo de mutações.

simulações epidemiológicas para determinar quanto tempo o SARS-CoV-2 teria circulado antes do tempo do MRCA, estimou-se que o intervalo mais plausível seria entre meados de outubro e meados de novembro de 2019 na província de Hubei. Além disso, conjecturou-se que dois terços dos eventos zoonóticos seriam autolimitados e incapazes de desencadear uma pandemia, bem como teriam sido necessárias 2,2 mutações (0,5 a 3,9) antes de dar origem aos padrões observados de diversidade genética amostrados (PEKAR et al., 2021a). Esta estimativa é coerente com a identificação do provável MRCA do SARS-CoV-2 (proCoV2), o qual foi estimado com base em uma modelagem do histórico mutacional do vírus. O proCoV2 diferiria dos primeiros genomas amostrados na China por três substituições, o que implica que nenhum dos primeiros pacientes amostrados representaria o caso índice ou teria dado origem a todas as infecções humanas observadas (KUMAR et al., 2021).

Um estudo que investigou quando, onde e como o SARS-CoV-2 estabeleceu suas primeiras cadeias de transmissão na Europa e nos Estados Unidos, demonstrou a necessidade de utilizar múltiplas fontes de informação (fluxo de passageiros de avião, incidência da doença na província de Hubei e outros locais, bem como modelagem dos possíveis caminhos evolutivos virais) e integrá-las para a testagem de hipóteses epidemiológicas. Isso ocorre pois, muito embora genomas virais possam fornecer informações críticas sobre a ligação epidemiológica de vírus separados no espaço e no tempo, o SARS-CoV-2 evolui a uma taxa considerada lenta de $\cong 1\times10^{-3}$ substituições por sítio por ano (≅2 mutações por mês). Consequentemente, toda a população global do SARS-CoV-2 até março de 2020 divergiu por apenas 0 a 12 mutações nucleotídicas em comparação ao ancestral inferido de toda a pandemia, o que faz com que as cadeias de transmissão sejam

definidas por 1 a 3 nucleotídeos diferentes. As inferências filogeográficas foram ainda mais confundidas pela disponibilidade baixa de dados de sequências de locais que sofreram surtos iniciais, incluindo Itália, Irã e o epicentro original em Hubei. Portanto, a combinação da taxa relativamente lenta de evolução da SARS-CoV-2, sua rápida disseminação dentro e entre locais, e a amostragem não representativa apresenta riscos para interpretações errôneas (WOROBEY et al., 2020).

Outro desafio imposto em um contexto pandêmico em que a vigilância genômica torna-se essencial é a desigualdade de sequências genômicas do SARS-CoV-2 depositadas por países desenvolvidos em comparação a países subdesenvolvidos e em desenvolvimento. Em um estudo que analisou essas disparidades, verificou-se que os países de alta renda contribuíram com 94% dos genomas depositados. Por exemplo, 100 dos 167 países analisados sequenciaram menos de 0,5% dos casos confirmados, enquanto 16 nações genotiparam mais de 5% dos casos. Adicionalmente, causa preocupação o fato de que 20 países, principalmente da África, não foram capazes de sequenciar nenhum genoma. Tais heterogeneidades espaço-temporais significativas apresentam correlação com fatores socioeconômicos, como: gastos em pesquisa e desenvolvimento *per capita*, produto interno bruto (PIB) *per capita* e índice de desenvolvimento humano (IDH) (BRITO et al., 2021).

Apesar do desenvolvimento de algumas ferramentas interativas que constroem filogenias com a maioria dos genomas disponíveis nos bancos de dados públicos (*e.g.*, https://cov2tree.org/ e http://pando.tools/), bem como outras que utilizam escores baseados em máxima parcimônia para adicionar novas sequências a grandes filogenias previamente construídas (TURAKHIA et al., 2021),

muitas vezes estamos interessados em avaliar a dinâmica espaço-temporal do SARS-CoV-2 em escalas mais locais, o que torna necessária uma amostragem capaz de incluir genomas representativos da diversidade genética, temporal e geográfica do vírus. Muito embora existam algumas ferramentas disponíveis publicamente para realizar tais procedimentos (ALPERT et al., 2021; BOLYEN et al., 2020; HADFIELD et al., 2018), não há um consenso sobre a utilização destas, de modo que outras abordagens também são usadas para necessidades específicas (DU PLESSIS et al., 2021; LEMEY et al., 2021). Portanto, a comparabilidade entre as diferentes formas de amostragem e seu impacto nos resultados observados ainda é pouco explorada, mas necessária para o entendimento das vantagens e desvantagens de cada metodologia.

Algumas limitações devem ser consideradas quanto às conclusões dos resultados apresentados nestre trabalho. Primeiramente, não foi possível analisar uma amostra maior de genomas devido aos recursos disponíveis. Além disso, a baixa quantidade e representatividade espacial das sequências do RS e demais estados limitou as inferências de eventos de introdução e movimento do vírus, especialmente em resolução municipal e estadual. Nesse sentido, observamos heterogeneidades espaciais e temporais nos esforços de sequenciamento do Brasil, dificultando uma acurada inferência das principais linhagens circulantes no país e podendo introduzir fatores de confusão. Uma distribuição desigual e sem proporcionalidade em relação ao número de casos por localidade, implica que as conclusões podem ser desproporcionalmente afetadas por eventos em períodos muito e pouco amostrados.

Portanto, o estabelecimento de um programa de vigilância nacional centralizado, estabelecendo: (i) regras claras de randomização e

representatividade, (ii) padronização de protocolos de sequenciamento e (iii) colaboração entre diferentes grupos de pesquisa nacionais, similar ao COG-UK realizado no Reino Unido (COVID-19 GENOMICS UK CONSORTIUM, 2020; NICHOLLS et al., 2021), é essencial para fornecer conjuntos de dados espaço-temporalmente representativos que permitam caracterizar mais precisamente a evolução do SARS-CoV-2 e demais patógenos emergentes no Brasil, identificando prontamente novas variantes para melhor responder e controlar sua propagação.

Adicionalmente, uma representação espaço-temporal desproporcional influencia fortemente as análises filodinâmicas e filogeográficas, uma vez que ligações epidemiológicas entre regiões podem ser maximizadas ou minimizadas devido à super ou subamostragem. Os modelos contínuos de dispersão geográfica utilizados são mais adequados para o espalhamento por terra em distâncias curtas, em oposição às longas distâncias percorridas de avião, uma prática comum no território brasileiro devido a seu tamanho continental. A principal consequência é que as inferências da localização dos nós próximos à raiz das árvores devem ser analisadas cuidadosamente e apresentam maior incerteza associada, tanto devido às limitações do modelo relacionadas à captura de viagens de longa distância quanto aos dados amostrados de modo esparso e à incerteza das reconstruções filogenéticas no contexto da COVID-19 previamente mencionados.

Finalmente, obtivemos estimativas das taxas evolutivas para algumas árvores MCC clado-específicas significativamente menores em relação à descobertas anteriores (8 a 9×10^{-4} subst/site/ano) (RAMBAUT, 2020; SU et al., 2020). Essas diferenças contribuíram para datações mais antigas dos ancestrais comuns mais recentes, uma vez que nesse caso é esperado que o acúmulo de mais mutações ocorra em um período de tempo mais longo. Embora algumas datas

inferidas provavelmente não sejam realistas no contexto da pandemia de COVID-19, elas destacam problemas com o processo de coleta de dados. As possíveis explicações para este comportamento são: amostragem não aleatória, amostras proximamente relacionadas com a mesma idade (agrupamento filo-temporal), e a presença de heterogeneidade nas taxas evolutivas entre linhagens (TONG et al., 2018).

7. CONCLUSÕES

Neste trabalho, objetivamos caracterizar a distribuição de linhagens e padrões de espalhamento geográfico do vírus SARS-CoV-2, causador da pandemia de COVID-19. Para este fim, utilizamos amostras do município de Esteio na primeira fase da epidemia (maio a outubro de 2020), do Rio Grande do Sul em período de aumento de hospitalizações e mortes (março de 2021) devido ao surgimento da variante P.1 (Gama) e de todo o território brasileiro no primeiro ano da epidemia (fevereiro de 2020 a fevereiro de 2021). Ao utilizar dados epidemiológicos e genomas completos do SARS-CoV-2 dos pacientes locais e de um conjunto representativo da diversidade viral mundial depositado no banco de dados GISAID, conseguimos contextualizar os genomas sequenciados e realizar análises detalhadas para compreender o espalhamento geográfico e dinâmica viral em escalas locais, regionais e internacionais.

A pandemia de COVID-19 foi a primeira na qual a comunidade científica internacional foi capaz de aplicar sequenciamento de genomas completos praticamente em tempo real. O compartilhamento desses dados e as análises resultantes informaram decisões de saúde pública, permitiram a detecção de mutações e variantes que poderiam afetar a virulência, patogênese, transmissibilidade, alcance de novos hospedeiros, escape imunológico e eficácia das vacinas. Tanto para essa quanto para futuras pandemias, tais iniciativas e colaborações irão permitir uma resposta ainda mais rápida às emergências de saúde pública. Contudo, predições genótipo-fenótipo não podem ser realizadas no mesmo ritmo do sequenciamento genômico. Portanto, o próximo nível da vigilância genômica provavelmente irá abranger uma abordagem sistemática capaz de correlacionar a detecção e a caracterização fenotípica das novas variantes em

tempo oportuno, visando contribuir para o desenvolvimento e atualização de testes diagnósticos, vacinas, estratégias terapêuticas e adoção de intervenções não-farmacológicas.

REFERÊNCIAS BIBLIOGRÁFICAS

ABU-RADDAD, L. J.; CHEMAITELLY, H.; BUTT, A. A. Effectiveness of the BNT162b2 Covid-19 Vaccine against the B.1.1.7 and B.1.351 Variants. **New England Journal of Medicine**, v. 385, n. 2, p. 187–189, 8 jul. 2021.

ALI, F.; KASRY, A.; AMIN, M. The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. **Medicine in Drug Discovery**, v. 10, p. 100086, 1 jun. 2021.

ALLICOCK, O. M. et al. Phylogeography and Population Dynamics of Dengue Viruses in the Americas. **Molecular Biology and Evolution**, v. 29, n. 6, p. 1533–1543, 1 jun. 2012.

ALPERT, T. et al. Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. **Cell**, v. 184, n. 10, p. 2595- 2604.e13, 13 maio 2021.

ALTMANN, D. M.; BOYTON, R. J.; BEALE, R. Immunity to SARS-CoV-2 variants of concern. **Science**, v. 371, n. 6534, p. 1103–1104, 12 mar. 2021.

ANDERSEN, K. G. et al. The proximal origin of SARS-CoV-2. **Nature Medicine**, v. 26, n. 4, p. 450-452, 17 mar. 2020.

ANDERSON, R. M. et al. Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 359, n. 1447, p. 1091–1105, 29 jul. 2004.

ARMOUGOM, F. et al. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. **Nucleic Acids Research**, v. 34, p. W604–W608, 1 jul. 2006.

AVANZATO, V. A. et al. Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. **Cell**, v. 183, n. 7, p. 1901-1912.e9, 23 dez. 2020.

BAUM, A. et al. Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. **Science**, v. 369, n. 6506, p. 1014–1018, 21 ago. 2020.

BAZYKIN, G. et al. Emergence of Y453F and Δ69-70HV mutations in a lymphoma patient with long-term COVID-19. **Virological**. Disponível em: <https://virological.org/t/emergence-of-y453f-and-69-70hv-mutations-in-a-lymphoma-patient-with-long-term-covid-19/580>. Acesso em: 2 set. 2021.

BERMINGHAM, E.; MORITZ, C. Comparative phylogeography: concepts and applications. **Molecular Ecology**, v. 7, n. 4, p. 367–369, 1998.

BERTRAM, S. et al. TMPRSS2 activates the human coronavirus 229E for cathepsin-independent host cell entry and is expressed in viral target cells in the respiratory epithelium. **Journal of Virology**, v. 87, n. 11, p. 6150–6160, jun. 2013.

BITTAR, C. et al. The Emergence of the New P.4 Lineage of SARS-CoV-2 With Spike L452R Mutation in Brazil. **Frontiers in Public Health**, v. 9, p. 1465, 2021.

BLOOM, J. D. et al. Investigate the origins of COVID-19. **Science**, v. 372, n. 6543, p. 694–694, 14 maio 2021.

BLOOM, J. D. Recovery of deleted deep sequencing data sheds more light on the early Wuhan SARS-CoV-2 epidemic. **Molecular Biology and Evolution**, v. 38, n. 12, p. 5211-5224, 16 ago. 2021.

BOLYEN, E. et al. Reproducibly sampling SARS-CoV-2 genomes across time, geography, and viral diversity. **F1000Research**, v. 9, p. 657, 29 jun. 2020.

BONI, M. 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, nov. 2020.

BRITO, A. F. et al. Global disparities in SARS-CoV-2 genomic surveillance. **MedRxiv**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.08.21.21262393v1>. Acesso em: 21 set. 2021.

BUSS, L. F. et al. Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic. **Science**, v. 371, n. 6526, p. 288-292, 8 dez. 2020.

CAI, Y. et al. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. Science, v. 373, n. 6555, p. 642–648, 6 ago. 2021.

CANDIDO, D. et al. Routes for COVID-19 importation in Brazil. **Journal of Travel Medicine**, v. 27, n. 3, 18 maio 2020a.

CANDIDO, D. et al. Evolution and epidemic spread of SARS-CoV-2 in Brazil. **Science**, v. 369, n. 6508, p. 1255–1260, 4 set. 2020b.

CARREÑO, J. M. et al. Activity of convalescent and vaccine serum against SARS-CoV-2 Omicron. **Nature**, 31 dez. 2021.

CASTRO, M. C. et al. Spatiotemporal pattern of COVID-19 spread in Brazil. **Science**, v. 372, n. 6544, p. 821–826, 21 maio 2021.

CELE, S. et al. SARS-CoV-2 Omicron has extensive but incomplete escape of Pfizer BNT162b2 elicited neutralization and requires ACE2 for infection. **MedRxiv**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.12.08.21267417v2>. Acesso em: 14 dez. 2021.

CENTRO ESTADUAL DE VIGILÂNCIA EM SAÚDE. Boletim Genômico 5 (16/04/2021). Disponível em: <https://coronavirus.rs.gov.br/upload/arquivos/202104/16173629-vigilancia-genomica-rs-boletim05-compactado.pdf>. Acesso em: 20 abr. 2021.

CERUTTI, G. et al. Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. **Cell Host & Microbe**, v. 29, n. 5, p. 819- 833.e7, 12 maio 2021.

CHAIN, P. et al. An applications-focused review of comparative genomics tools: Capabilities, limitations and future challenges. **Briefings in Bioinformatics**, v. 4, n. 2, p. 105–123, 1 jun. 2003.

CHANG, C. et al. The SARS coronavirus nucleocapsid protein--forms and functions. **Antiviral Research**, v. 103, p. 39–50, mar. 2014.

CHEN, X. et al. Neutralizing antibodies against SARS-CoV-2 variants induced by natural infection or vaccination: a systematic review and pooled meta-analysis. **Clinical Infectious Diseases**, p. ciab646, 24 jul. 2021.

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.

CHI, X. et al. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. **Science**, v. 369, n. 6504, p. 650–655, 7 ago. 2020.

CHOI, B. et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. **New England Journal of Medicine**, v. 383, n. 23, p. 2291–2293, 3 dez. 2020.

CLARO, I. M. 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.

COBEY, S.; KOELLE, K. Capturing escape in infectious disease dynamics. **Trends in Ecology & Evolution**, v. 23, n. 10, p. 572–577, out. 2008.

COREY, L. et al. SARS-CoV-2 Variants in Patients with Immunosuppression. **New England Journal of Medicine**, v. 385, n. 6, p. 562–566, 5 ago. 2021.

COTTEN, M. et al. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. **The Lancet**, v. 382, n. 9909, p. 1993–2002, 14 dez. 2013.

COVID-19 GENOMICS UK CONSORTIUM. An integrated national scale SARS-CoV-2 genomic surveillance network. **The Lancet Microbe**, v. 1, n. 3, p. e99–e100, 1 jul. 2020.

CZELUSNIAK, J. et al. Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. **Methods in Enzymology**, v. 183, p. 601–615, 1990.

DARWIN, C. On the origin of species by means of natural selection. 1. ed. **London: John Murray**, 1859.

DAVIES, N. G. et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. **Science**, v. 372, n. 6538, p. eabg3055, 3 mar. 2021a.

DAVIES, N. G. et al. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. **Nature**, v. 593, n. 7858, p. 270–274, maio 2021b.

DE ALMEIDA, L. G. et al. Genomic Surveillance of SARS-CoV-2 in the State of Rio de Janeiro, Brazil: technical briefing. **Virological**. Disponível em: <https://virological.org/t/genomic-surveillance-of-sars-cov-2-in-the-state-of-rio-de-janeiro-brazil-technical-briefing/683>. Acesso em: 4 maio. 2021.

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

DE WILDE, A. H. et al. Host Factors in Coronavirus Replication. **Current Topics in Microbiology and Immunology**, v. 419, p. 1–42, 2018.

DE WIT, E. et al. SARS and MERS: recent insights into emerging coronaviruses. **Nature Reviews Microbiology**, v. 14, n. 8, p. 523–534, 27 jun. 2016.

DEARLOVE, B. et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. **Proceedings of the National Academy of Sciences of the United States of America**, v. 117, n. 38, p. 23652–23662, 22 2020.

DEDIEGO, M. L. et al. A severe acute respiratory syndrome coronavirus that lacks the E gene is attenuated in vitro and in vivo. **Journal of Virology**, v. 81, n. 4, p. 1701–1713, fev. 2007.

DEJNIRATTISAI, W. et al. Antibody evasion by the P.1 strain of SARS-CoV-2. **Cell**, v. 184, n. 11, p. 2939- 2954.e9, 27 maio 2021.

DELAUNE, D. et al. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. **Nature Communications**, v. 12, n. 1, p. 6563, 9 nov. 2021.

DEMOLINER, M. 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, 9 nov. 2021.

DENG, X. et al. Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. **Cell**, v. 184, n. 13, p. 3426-3437.e8, 20 abr. 2021.

DENG, X.; DEN BAKKER, H. C.; HENDRIKSEN, R. S. Genomic Epidemiology: Whole-Genome-Sequencing–Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. **Annual Review of Food Science and Technology**, v. 7, n. 1, p. 353–374, 2016.

DIEKMANN, O.; HEESTERBEEK, J. A. P.; METZ, J. A. J. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. **Journal of Mathematical Biology**, v. 28, n. 4, p. 365–382, 1 jun. 1990.

DOS SANTOS, M. C. et al. First reported cases of SARS-CoV-2 sub-lineage B.1.617.2 in Brazil: an outbreak in a ship and alert for spread. **Virological**. Disponível em: <https://virological.org/t/first-reported-cases-of-sars-cov-2-sub-lineage-b-1-617-2-in-brazil-an-outbreak-in-a-ship-and-alert-for-spread/706>. Acesso em: 1 jun. 2021.

DOWNING, T. Tackling Drug Resistant Infection Outbreaks of Global Pandemic Escherichia coli ST131 Using Evolutionary and Epidemiological Genomics. **Microorganisms**, v. 3, n. 2, p. 236–267, 20 maio 2015.

DRUMMOND, A. J. et al. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. **Genetics**, v. 161, n. 3, p. 1307–1320, 1 jul. 2002.

DRUMMOND, A. J. et al. Relaxed Phylogenetics and Dating with Confidence. **PLOS Biology**, v. 4, n. 5, p. e88, 14 mar. 2006.

DRUMMOND, A. J.; RAMBAUT, A. BEAST: Bayesian evolutionary analysis by sampling trees. **BMC Evolutionary Biology**, v. 7, n. 1, p. 214, 8 nov. 2007.

DU PLESSIS, L. et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. **Science**, 8 jan. 2021.

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

DUDAS, G. et al. MERS-CoV spillover at the camel-human interface. eLife, v. 7, p. e31257, 16 jan. 2018.

DUDAS, G.; RAMBAUT, A. MERS-CoV recombination: implications about the reservoir and potential for adaptation. **Virus Evolution**, v. 2, n. 1, p. vev023, 20 jan. 2016.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, 2004.

EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL. Distribution of confirmed cases of MERS-CoV by place of infection and month of onset, from March 2012 to 2 December 2019. Disponível em: <https://www.ecdc.europa.eu/en/publications-data/distribution-confirmed-cases-mers-cov-place-infection-and-month-onset-march-2012>. Acesso em: 27 maio. 2021.

EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL. Rapid Risk Assessment: Detection of new SARS-CoV-2 variants related to mink. Disponível em: <https://www.ecdc.europa.eu/en/publications-data/detection-new-sars-cov-2-variants-mink>. Acesso em: 2 set. 2021.

FARIA, N. et al. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. **Science**, v. 372, n. 6544, p. 815–821, 14 abr. 2021.

FEHR, A. R.; PERLMAN, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. **Methods in molecular biology (Clifton, N.J.)**, v. 1282, p. 1–23, 2015.

FELSENSTEIN, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. **Journal of Molecular Evolution**, v. 17, n. 6, p. 368–376, 1 nov. 1981.

FELSENSTEIN, J. Distance Methods for Inferring Phylogenies: A Justification. **Evolution**, v. 38, n. 1, p. 16–24, 1984.

FELSENSTEIN, J. CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP. **Evolution; International Journal of Organic Evolution**, v. 39, n. 4, p. 783–791, jul. 1985.

FENG, D.-F.; DOOLITTLE, R. F. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. **Journal of Molecular Evolution**, v. 25, n. 4, p. 351–360, 1 ago. 1987.

FERGUSON, N. et al. Report 49 - Growth, population distribution and immune escape of Omicron in England. Disponível em: <http://www.imperial.ac.uk/medicine/departments/school-public-health/infectious-disease-epidemiology/mrc-global-infectious-disease-analysis/covid-19/report-49-omicron/>. Acesso em: 2 jan. 2022a.

FERGUSON, N. et al. Report 50 - Hospitalisation risk for Omicron cases in England. Disponível em: <http://www.imperial.ac.uk/medicine/departments/school-public-health/infectious-disease-epidemiology/mrc-global-infectious-disease-analysis/covid-19/report-50-severity-omicron/>. Acesso em: 3 jan. 2022b.

FERRAREZE, P. A. G. et al. E484K as an innovative phylogenetic event for viral evolution: Genomic analysis of the E484K spike mutation in SARS-CoV-2 lineages from Brazil. **Infection, Genetics and Evolution**, v. 93, p. 104941, 1 set. 2021.

FRANCESCHI, V. B. et al. Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil. **BMC Genomics**, v. 22, n. 1, p. 371, 20 maio 2021a.

FRANCESCHI, V. B. et al. Predominance of the SARS-CoV-2 Lineage P.1 and Its Sublineage P.1.2 in Patients from the Metropolitan Region of Porto Alegre, Southern Brazil in March 2021. **Pathogens**, v. 10, n. 8, p. 988, ago. 2021b.

FRANCESCHI, V. B. et al. Mutation hotspots and spatiotemporal distribution of SARS-CoV-2 lineages in Brazil, February 2020-2021. **Virus Research**, v. 304, p. 198532, 15 out. 2021c.

FRANCISCO JUNIOR, R. DA S. et al. Pervasive transmission of E484K and emergence of VUI-NP13L with evidence of SARS-CoV-2 co-infection events by two different lineages in Rio Grande do Sul, Brazil. **Virus Research**, v. 296, p. 198345, 15 abr. 2021a.

FRANCISCO JUNIOR, R. DA S. et al. Turnover of SARS-CoV-2 Lineages Shaped the Pandemic and Enabled the Emergence of New Variants in the State of Rio de Janeiro, Brazil. **Viruses**, v. 13, n. 10, p. 2013, out. 2021b.

GHOSH, S. et al. β-Coronaviruses Use Lysosomes for Egress Instead of the Biosynthetic Secretory Pathway. **Cell**, v. 183, n. 6, p. 1520- 1535.e14, 10 dez. 2020.

GIOVANETTI, M. et al. Genomic epidemiology reveals how restriction measures shaped the SARS-CoV-2 epidemic in Brazil. **MedRxiv**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.10.07.21264644v1>. Acesso em: 8 nov. 2021.

GOBEIL, S. M.C. et al. Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. **Science**, v. 373, n. 6555, p. eabi6226, 2021a.

GOBEIL, S. M.-C. et al. D614G Mutation Alters SARS-CoV-2 Spike Conformation and Enhances Protease Cleavage at the S1/S2 Junction. **Cell Reports**, v. 34, n. 2, p. 108630,

12 jan. 2021b.

GOLDMAN, N.; YANG, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. **Molecular Biology and Evolution**, v. 11, n. 5, p. 725–736, set. 1994.

GRÄF, T. et al. Identification of SARS-CoV-2 P.1-related lineages in Brazil provides new insights about the mechanisms of emergence of Variants of Concern. **Virological**. Disponível em: <https://virological.org/t/identification-of-sars-cov-2-p-1-related-lineages-in-brazil-provides-new-insights-about-the-mechanisms-of-emergence-of-variants-of-concern/694/1>. Acesso em: 17 maio. 2021a.

GRÄF, T. et al. Phylogenetic-based inference reveals distinct transmission dynamics of SARS-CoV-2 variant of concern Gamma and lineage P.2 in Brazil. **MedRxiv**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.10.24.21265116v1>. Acesso em: 8 nov. 2021b.

GRAHAM, M. S. et al. Changes in symptomatology, reinfection, and transmissibility associated with the SARS-CoV-2 variant B.1.1.7: an ecological study. **The Lancet Public Health**, v. 6, n. 5, p. e335–e345, 1 maio 2021.

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.

GRANT, R. et al. Impact of SARS-CoV-2 Delta variant on incubation, transmission settings and vaccine effectiveness: Results from a nationwide case-control study in France. **The Lancet Regional Health – Europe**, 25 nov. 2021.

GREANEY, A. J. et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. **Cell Host & Microbe**, v. 29, n. 1, p. 44-57.e9, 19 nov. 2020.

GREANEY, A. J. et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. **Cell Host & Microbe**, v. 29, n. 3, p. 463-476.e6, 8 fev. 2021a.

GREANEY, A. J. et al. Antibodies elicited by mRNA-1273 vaccination bind more broadly to the receptor binding domain than do those from SARS-CoV-2 infection. **Science Translational Medicine**, v. 13, n. 600, p. eabi9915, 30 jun. 2021b.

GRENFELL, B. T. et al. Unifying the epidemiological and evolutionary dynamics of pathogens. **Science (New York, N.Y.)**, v. 303, n. 5656, p. 327–332, 16 jan. 2004.

GRONVALL, G. K. The Contested Origin of SARS-CoV-2. **Survival**, v. 63, n. 6, p. 7–36, 2 nov. 2021.

GRUBAUGH, N. D.; PETRONE, M. E.; HOLMES, E. C. We shouldn't worry when a virus mutates during disease outbreaks. **Nature Microbiology**, v. 5, n. 4, p. 529–530, abr. 2020.

GU, H. et al. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. **Science**, v. 369, n. 6511, p. 1603–1607, 25 set. 2020.

GUAN, Y. et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. **Science (New York, N.Y.)**, v. 302, n. 5643, p. 276–278, 10 out. 2003.

GUTIERREZ, B. et al. Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America. **MedRxiv**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.11.19.21266601v1>. Acesso em: 24 nov. 2021.

HAAS, E. J. et al. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. **The Lancet**, v. 397, n. 10287, p. 1819–1829, 15 maio 2021.

HADFIELD, J. et al. Nextstrain: real-time tracking of pathogen evolution. **Bioinformatics**, v. 34, n. 23, p. 4121–4123, 1 dez. 2018.

HARVEY, W. T. et al. SARS-CoV-2 variants, spike mutations and immune escape. **Nature Reviews Microbiology**, v. 19, n. 7, p. 409-424, 1 jun. 2021.

HASEGAWA, M.; KISHINO, H.; YANO, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. **Journal of Molecular Evolution**, v. 22, n. 2, p. 160–174, 1985.

HE, W.-T. et al. Total virome characterizations of game animals in China reveals a spectrum of emerging viral pathogens. **BioRxiv**. Disponível em: <https://www.biorxiv.org/content/10.1101/2021.11.10.467646v1>. Acesso em: 24 nov. 2021.

HEMELAAR, J. et al. Global trends in molecular epidemiology of HIV-1 during 2000–2007. **AIDS (London, England)**, v. 25, n. 5, p. 679–689, 13 mar. 2011.

HILLIS, D. M. Approaches for Assessing Phylogenetic Accuracy. **Systematic Biology**, v. 44, n. 1, p. 3–16, 1995.

HODCROFT, E. B. et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. **Nature**, v. 595, n. 7869, p. 707-71, 7 jun. 2021.

HOFFMANN, M. et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. **Cell**, v. 181, n. 2, p. 271- 280.e8, 16 abr. 2020.

HOFFMANN, M. et al. SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. **Cell**, v. 184, n. 9, p. 2384- 2393.e12, 29 abr. 2021.

HOFFMANN, M.; KLEINE-WEBER, H.; PÖHLMANN, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. **Molecular Cell**, v. 78, n. 4, p. 779- 784.e5, 21 maio 2020.

HOLMES, E. C. Evolutionary History and Phylogeography of Human Viruses. **Annual Review of Microbiology**, v. 62, n. 1, p. 307–328, 2008.

HOLMES, E. C. et al. The Origins of SARS-CoV-2: A Critical Review. **Cell**, v. 184, n. 19, p. 4848–4856, 2021.

HON, C.C. et al. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. **Journal of Virology**, v. 82, n. 4, p. 1819–1826, fev. 2008.

HUANG, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. **The Lancet**, v. 395, n. 10223, p. 497–506, 15 fev. 2020a.

HUANG, Y. et al. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. **Acta Pharmacologica Sinica**, v. 41, n. 9, p. 1141–1149, set. 2020b.

HUELSENBECK, J. P. et al. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. **Science**, v. 294, n. 5550, p. 2310–2314, 14 dez. 2001.

HUSSAIN, S. et al. Identification of novel subgenomic RNAs and noncanonical transcription initiation signals of severe acute respiratory syndrome coronavirus. **Journal of Virology**, v. 79, n. 9, p. 5288–5295, maio 2005.

JACKSON, B. et al. Generation and transmission of inter-lineage recombinants in the SARS-CoV-2 pandemic. **Cell**, v. 184, n. 20, p. 5179-5188.e8, 16 ago. 2021.

JANSEN, L. et al. Investigation of a SARS-CoV-2 B.1.1.529 (Omicron) Variant Cluster — Nebraska, November–December 2021. **MMWR. Morbidity and Mortality Weekly Report**, v. 70, n. 5152, p. 1782–1784, 31 dez. 2021.

JOHNS HOPKINS CORONAVIRUS RESOURCE CENTER. COVID-19 Map. Disponível em: <https://coronavirus.jhu.edu/map.html>. Acesso em: 27 maio. 2021.

JOHNSON, B. A. et al. Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. **Nature**, v. 591, n. 7849, p. 293–299, mar. 2021a.

JOHNSON, B. A. et al. Nucleocapsid mutations in SARS-CoV-2 augment replication and pathogenesis. **BioRxiv**. Disponível em: <https://www.biorxiv.org/content/10.1101/2021.10.14.464390v1>. Acesso em: 16 out. 2021b.

JUKES, T. H.; CANTOR, C. R. CHAPTER 24 - Evolution of Protein Molecules. In: MUNRO, H. N. (Ed.). M**ammalian Protein Metabolism. Academic Press**, 1969. p. 21–132.

KARIM, S. S. A.; KARIM, Q. A. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. **The Lancet**, v. 398, n. 10317, p. 2126–2128, 11 dez. 2021.

KEETON, R. et al. SARS-CoV-2 spike T cell responses induced upon vaccination or infection remain robust against Omicron. **MedRxiv**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.12.26.21268380v1>. Acesso em: 2 jan. 2022.

KEMP, S. A. et al. SARS-CoV-2 evolution during treatment of chronic infection. **Nature**, v. 592, n. 7853, p. 277-282, 5 fev. 2021.

KENNEDY, D. A.; READ, A. F. Why does drug resistance readily evolve but vaccine resistance does not? **Proceedings of the Royal Society B: Biological Sciences**, v. 284, n. 1851, p. 20162562, 29 mar. 2017.

KEOGH-BROWN, M. R.; SMITH, R. D. The economic impact of SARS: How does the reality match the predictions? **Health Policy**, v. 88, n. 1, p. 110–120, 1 out. 2008.

KILPATRICK, A. M. et al. Predicting pathogen introduction: West Nile virus spread to Galapagos. **Conservation Biology: The Journal of the Society for Conservation Biology**, v. 20, n. 4, p. 1224–1231, ago. 2006.

KIM, Y.-I. et al. Infection and Rapid Transmission of SARS-CoV-2 in Ferrets. **Cell Host & Microbe**, v. 27, n. 5, p. 704- 709.e2, 13 maio 2020.

KIMURA, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. **Journal of Molecular Evolution**, v. 16, n. 2, p. 111–120, 1 jun. 1980.

KIMURA, M. The Neutral Theory of Molecular Evolution. **Cambridge: Cambridge University Press**, 1983.

KINGMAN, J. F. C. On the Genealogy of Large Populations. **Journal of Applied Probability**, v. 19, p. 27–43, 1982.

KISTLER, K. E.; HUDDLESTON, J.; BEDFORD, T. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. **BioRxiv**. Disponível em: <https://www.biorxiv.org/content/10.1101/2021.09.11.459844v1>. Acesso em: 28 set. 2021.

KNOWLES, L. L. Statistical Phylogeography. **Annual Review of Ecology, Evolution, and**

Systematics, v. 40, n. 1, p. 593–612, 2009.

KOELLE, K. et al. Epochal Evolution Shapes the Phylodynamics of Interpandemic Influenza A (H3N2) in Humans. **Science**, v. 314, n. 5807, p. 1898–1903, 22 dez. 2006.

KOOPMANS, M. SARS-CoV-2 and the human-animal interface: outbreaks on mink farms. **The Lancet Infectious Diseases**, v. 21, n. 1, p. 18–19, 1 jan. 2021.

KORBER, B. 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, 20 ago. 2020.

KUMAR, S. et al. An Evolutionary Portrait of the Progenitor SARS-CoV-2 and Its Dominant Offshoots in COVID-19 Pandemic. **Molecular Biology and Evolution**, v. 38, n. 8, p. 3046-3059, 4 maio 2021.

LAI, M. M. et al. Recombination between nonsegmented RNA genomes of murine coronaviruses. **Journal of Virology**, v. 56, n. 2, p. 449–456, nov. 1985.

LAM, T. T.-Y.; HON, C.-C.; TANG, J. W. Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. **Critical Reviews in Clinical Laboratory Sciences**, v. 47, n. 1, p. 5–49, 1 fev. 2010.

LAMARCA, A. P. et al. Genomic Surveillance Tracks the First Community Outbreak of the SARS-CoV-2 Delta (B.1.617.2) Variant in Brazil. **Journal of Virology**, v. JVI.01228-21, 2021a.

LAMARCA, A. P. et al. Genomic surveillance of SARS-CoV-2 tracks early interstate transmission of P.1 lineage and diversification within P.2 clade in Brazil. **PLOS Neglected Tropical Diseases**, v. 15, n. 10, p. e0009835, 13 out. 2021b.

LAN, J. et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. **Nature**, v. 581, n. 7807, p. 215–220, maio 2020.

LASEK-NESSELQUIST, E. et al. The localized rise of a B.1.526 SARS-CoV-2 variant containing an E484K mutation in New York State. **MedRxiv**, Disponível em: <https://www.medrxiv.org/content/10.1101/2021.02.26.21251868v1>. Acesso em: 1 mar. 2021a.

LASEK-NESSELQUIST, E. et al. A tale of three SARS-CoV-2 variants with independently acquired P681H mutations in New York State. **MedRxiv**, Disponível em: <https://www.medrxiv.org/content/10.1101/2021.03.10.21253285v1.full>. Acesso em: 12 mar. 2021b.

LEMEY, P. et al. Bayesian Phylogeography Finds Its Roots. **PLOS Computational Biology**, v. 5, n. 9, p. e1000520, 25 set. 2009.

LEMEY, P. et al. Untangling introductions and persistence in COVID-19 resurgence in Europe. **Nature**, v. 595, n. 7869, p. 713-717, 30 jun. 2021.

LEUNG, K. et al. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. **Eurosurveillance**, v. 26, n. 1, p. 2002106, 7 jan. 2021.

LEWIS, F. et al. Episodic Sexual Transmission of HIV Revealed by Molecular Phylodynamics. **PLOS Medicine**, v. 5, n. 3, p. e50, 18 mar. 2008.

LI, Q. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. **New England Journal of Medicine**, 13. v. 382, p. 1199–1207, 29 jan. 2020.

LI, W. et al. Bats Are Natural Reservoirs of SARS-Like Coronaviruses. **Science**, v. 310, n. 5748, p. 676–679, 28 out. 2005.

LIU, C. et al. Reduced neutralization of SARS-CoV-2 B.1.617 by vaccine and convalescent serum. **Cell**, v. 184, n. 16, p. 4220- 4236.e13, 5 ago. 2021a.

LIU, C. et al. Structural basis of mismatch recognition by a SARS-CoV-2 proofreading enzyme. **Science**, v. 373, n. 6559, p. 1142–1146, 3 set. 2021b.

LIU, J. et al. BNT162b2-elicited neutralization of B.1.617 and other SARS-CoV-2 variants. **Nature**, v. 596, n. 7871, p. 273–275, ago. 2021c.

LIU, Y. et al. The N501Y spike substitution enhances SARS-CoV-2 transmission. **BioRxiv**, Disponível em: <https://www.biorxiv.org/content/10.1101/2021.03.08.434499v1>. Acesso em: 9 mar. 2021d.

LOPEZ BERNAL, J. et al. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. **New England Journal of Medicine**, v. 385, n. 7, p. 585–594, 12 ago. 2021.

LU, H.; STRATTON, C. W.; TANG, Y.-W. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. **Journal of Medical Virology**, v. 92, n. 4, p. 401–402, abr. 2020.

LYTRAS, S. et al. The animal origin of SARS-CoV-2. **Science**, v. 373, n. 6558, p. 968–970, 27 ago. 2021.

LYTRAS, S.; MACLEAN, O.; ROBERTSON, D. L. The Sarbecovirus origin of SARS-CoV-2's furin cleavage site. **Virological**. Disponível em: <https://virological.org/t/the-sarbecovirus-origin-of-sars-cov-2-s-furin-cleavage-site/536>. Acesso em: 10 ago. 2021.

MA, Y. et al. Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. **Proceedings of the National Academy of Sciences**, v. 112, n. 30, p. 9436–9441, 28 jul. 2015.

MADHI, S. A. et al. Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. **New England Journal of Medicin**e, v. 384, n. 20, p. 1885–1898, 20 maio 2021.

MARGOLIASH, E. Primary structure and evolution of cytochrome C. **Proceedings of the National Academy of Sciences of the United States of America**, v. 50, n. 4, p. 672–679, out. 1963.

MARTIN, D. P. et al. The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. **Cell**, v. 184, n. 20, p. 5189-5200.e7, 6 set. 2021.

MARTINS, A. F. et al. Detection of SARS-CoV-2 lineage P.1 in patients from a region with exponentially increasing hospitalisation rate, February 2021, Rio Grande do Sul, Southern Brazil. **Eurosurveillance**, v. 26, n. 12, p. 2100276, 25 mar. 2021.

MASTERS, P. S. The Molecular Biology of Coronaviruses. In: **Advances in Virus Research. Academic Press**, 2006. v. 66p. 193–292.

MCCORMICK, K. D.; JACOBS, J. L.; MELLORS, J. W. The emerging plasticity of SARS-CoV-2. **Science**, v. 371, n. 6536, p. 1306–1308, 26 mar. 2021.

MEMISH, Z. A. et al. Respiratory tract samples, viral load, and genome fraction yield in patients with Middle East respiratory syndrome. **The Journal of Infectious Diseases**, v. 210, n. 10, p. 1590–1594, 15 nov. 2014.

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

MENG, B. et al. SARS-CoV-2 Omicron spike mediated immune escape, infectivity and cell-cell fusion. **BioRxiv**. Disponível em: <https://www.biorxiv.org/content/10.1101/2021.12.17.473248v2>. Acesso em: 2 jan. 2022b.

MEYER, B. et al. Antibodies against MERS coronavirus in dromedary camels, United Arab Emirates, 2003 and 2013. **Emerging Infectious Diseases**, v. 20, n. 4, p. 552–559, abr. 2014.

MILLER, W. et al. Comparative Genomics. **Annual Review of Genomics and Human Genetics**, v. 5, n. 1, p. 15–56, 2004.

MIR, D. et al. Recurrent Dissemination of SARS-CoV-2 Through the Uruguayan–Brazilian Border. **Frontiers in Microbiology**, v. 12, p. 1018, 2021.

MISHRA, S. et al. Changing composition of SARS-CoV-2 lineages and rise of Delta variant in England. **EClinicalMedicine**, v. 39, p. 101064, 1 set. 2021.

MOREIRA, F. R. R. et al. Epidemiological dynamics of SARS-CoV-2 VOC Gamma in Rio de Janeiro, Brazil. **Virus Evolution**, v. 7, n. 2, p. veab087, 1 out. 2021.

MOREL, B. et al. Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. **Molecular Biology and Evolution**, v. 38, n. 5, p. 1777–1791, 1 maio 2021.

MULLEN, J. L. et al. outbreak.info. Disponível em: <https://outbreak.info>. Acesso em: 19 jul. 2021a.

MULLEN, J. L. et al. B.1.617.2 Lineage Report. Disponível em: <https://outbreak.info/situation-reports?pango=B.1.617.2&loc=IND&loc=GBR&loc=USA&selected=IND>. Acesso em: 27 maio. 2021b.

MULLEN, J. L. et al. Lineage Comparison. Disponível em: <https://outbreak.info/compare-lineages>. Acesso em: 27 maio. 2021c.

MUNDO EDUCAÇÃO. Regiões do Brasil. Disponível em: < https://mundoeducacao.uol.com.br/geografia/as-regioes-brasil.htm>. Acesso em: 8 fev. 2022.

MUNNINK, B. B. O. et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. **Science**, v. 371, n. 6525, p. 172–177, 8 jan. 2021.

NAVECA, F. G. et al. COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. **Nature Medicine**, v. 27, n. 7, p. 1230-1238, 25 maio 2021a.

NAVECA, F. G. et al. Spread of Gamma (P.1) sub-lineages carrying Spike mutations close to the furin cleavage site and deletions in the N-terminal domain drives ongoing transmission of SARS-CoV-2 in Amazonas, Brazil. **Virological**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.09.12.21263453v1>. Acesso em: 22 set. 2021b.

NAVECA, F. G. et al. The SARS-CoV-2 variant Delta displaced the variants Gamma and Gamma plus in Amazonas, Brazil. **Virological**. Disponível em: <https://virological.org/t/the-sars-cov-2-variant-delta-displaced-the-variants-gamma-and-gamma-plus-in-amazonas-brazil/765>. Acesso em: 24 nov. 2021c.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology**, v. 48, n. 3, p. 443–453, 28 mar. 1970.

NEI, M.; GOJOBORI, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. **Molecular Biology and Evolution**, v. 3, n. 5, p. 418–426, 1 set. 1986.

NELSON, G. et al. Molecular dynamic simulation reveals E484K mutation enhances spike

RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. **BioRxiv**. Disponível em: <https://www.biorxiv.org/content/10.1101/2021.01.13.426558v1>. Acesso em: 13 jan. 2021.

NICHOLLS, S. M. et al. CLIMB-COVID: continuous integration supporting decentralised sequencing for SARS-CoV-2 genomic surveillance. **Genome Biology**, v. 22, n. 1, p. 196, 1 jul. 2021.

NICOLELIS, M. A. L. et al. The impact of super-spreader cities, highways, and intensive care availability in the early stages of the COVID-19 epidemic in Brazil. **Scientific Reports**, v. 11, n. 1, p. 13001, 21 jun. 2021.

NOTREDAME, C.; HIGGINS, D. G.; HERINGA, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. **Journal of Molecular Biology**, v. 302, n. 1, p. 205–217, 8 set. 2000.

OGANDO, N. S. et al. The Enzymatic Activity of the nsp14 Exoribonuclease Is Critical for Replication of MERS-CoV and SARS-CoV-2. **Journal of Virology**, v. 94, n. 23, p. e01246-20, 9 nov. 2020.

ORESHKOVA, N. et al. SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. **Eurosurveillance**, v. 25, n. 23, 11 jun. 2020.

ORTEGO, J. et al. Absence of E protein arrests transmissible gastroenteritis coronavirus maturation in the secretory pathway. **Virology**, v. 368, n. 2, p. 296–308, 25 nov. 2007.

O'TOOLE, Á. et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. **Virus Evolution**, v. 7, n. 2, p. veab064, 30 jul. 2021.

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

PAIVA, M. H. S. et al. 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, dez. 2020.

PAULES, C. I.; MARSTON, H. D.; FAUCI, A. S. Coronavirus Infections—More Than Just the Common Cold. **JAMA**, v. 323, n. 8, p. 707–708, 25 fev. 2020.

PEACOCK, T. P. et al. SARS-CoV-2 one year on: evidence for ongoing viral adaptation. **Journal of General Virology**, v. 102, n. 4, p. 001584, 2021a.

PEACOCK, T. P. et al. The SARS-CoV-2 variants associated with infections in India, B.1.617, show enhanced spike cleavage by furin. **BioRxiv**. Disponível em: <https://www.biorxiv.org/content/10.1101/2021.05.28.446163v1>. Acesso em: 28 maio 2021b.

PEIRIS, J. S. M. et al. Clinical progression and viral load in a community outbreak of coronavirus-associated SARS pneumonia: a prospective study. **Lancet (London, England)**, v. 361, n. 9371, p. 1767–1772, 24 maio 2003.

PEKAR, J. et al. Timing the SARS-CoV-2 index case in Hubei province. **Science**, v. 372, n. 6540, p. 412-417, 18 mar. 2021a.

PEKAR, J. et al. Evidence Against the Veracity of SARS-CoV-2 Genomes Intermediate between Lineages A and B. **Virological**. Disponível em: <https://virological.org/t/evidence-against-the-veracity-of-sars-cov-2-genomes-intermediate-between-lineages-a-and-b/754>. Acesso em: 21 set. 2021b.

PENG, J. et al. Estimation of secondary household attack rates for emergent spike L452R SARS-CoV-2 variants detected by genomic surveillance at a community-based testing site in San Francisco. **Clinical Infectious Diseases**, n. ciab283, 31 mar. 2021.

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

PEVSNER, J. Bioinformatics and functional genomics. 3rd ed. Hoboken, N.J: **Wiley-Blackwell**, 2015.

PIPES, L. et al. Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny. **Molecular Biology and Evolution**, v. 38, n. 4, p. 1537–1543, 1 abr. 2021.

PLANAS, D. et al. Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. **Nature Medicine**, v. 27, n. 5, p. 917–924, maio 2021a.

PLANAS, D. et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. **Nature**, v. 596, n. 7871, p. 276–280, ago. 2021b.

PLANTE, J. A. et al. Spike mutation D614G alters SARS-CoV-2 fitness. **Nature**, v. 592, n. 7852, p. 116-121, 26 out. 2020.

POND, S. L. K.; FROST, S. D. W.; MUSE, S. V. HyPhy: hypothesis testing using phylogenies. **Bioinformatics**, v. 21, n. 5, p. 676–679, 1 mar. 2005.

PULLIAM, J. R. C. et al. Increased risk of SARS-CoV-2 reinfection associated with emergence of the Omicron variant in South Africa. **MedRxiv**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.11.11.21266068v2>. Acesso em: 14 dez. 2021.

PYBUS, O. G. et al. The epidemic behavior of the hepatitis C virus. **Science (New York, N.Y.)**, v. 292, n. 5525, p. 2323–2325, 22 jun. 2001.

PYBUS, O. G.; RAMBAUT, A. Evolutionary analysis of the dynamics of viral infectious disease. **Nature Reviews Genetics**, v. 10, n. 8, p. 540–550, ago. 2009.

PYRC, K.; BERKHOUT, B.; HOEK, L. VAN DER. The Novel Human Coronaviruses NL63 and HKU1. **Journal of Virology**, v. 81, n. 7, p. 3051–3057, 1 abr. 2007.

RAMBAUT, A. Phylodynamic Analysis | 176 genomes | 6 Mar 2020 - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology. **Virological**. Disponível em: <https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356>. Acesso em: 11 fev. 2021.

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

RAMBAUT, A. et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. **Virological**. Disponível em: <https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563>. Acesso em: 4 jan. 2021b.

REDE GENÔMICA FIOCRUZ. Vigilância Genômica do SARS-CoV-2 no Brasil - Dashboard. Disponível em: <http://www.genomahcov.fiocruz.br/dashboard/>. Acesso em: 13 ago. 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, p. 615280, 2021a.

RESENDE, P. C. et al. Identification of a new B.1.1.33 SARS-CoV-2 Variant of Interest

(VOI) circulating in Brazil with mutation E484K and multiple deletions in the amino (N)-terminal domain of the Spike protein. **Virological**. Disponível em: <https://virological.org/t/identification-of-a-new-b-1-1-33-sars-cov-2-variant-of-interest-voi-circulating-in-brazil-with-mutation-e484k-and-multiple-deletions-in-the-amino-n-terminal-domain-of-the-spike-protein/675>. Acesso em: 1 jun. 2021b.

RESENDE, P. C. et al. A Potential SARS-CoV-2 Variant of Interest (VOI) Harboring Mutation E484K in the Spike Protein Was Identified within Lineage B.1.1.33 Circulating in Brazil. **Viruses**, v. 13, n. 5, p. 724, maio 2021c.

ROTA, P. A. et al. Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. Science, v. 300, n. 5624, p. 1394–1399, 30 maio 2003.

RUAN, Y. et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. **The Lancet**, v. 361, n. 9371, p. 1779–1785, 24 maio 2003.

SABINO, E. C. et al. Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. **The Lancet**, v. 397, n. 10273, p. 452–455, 6 fev. 2021.

SAITO, A. et al. SARS-CoV-2 spike P681R mutation enhances and accelerates viral fusion. **BioRxiv**. Disponível em: <https://www.biorxiv.org/content/10.1101/2021.06.17.448820v1>. Acesso em: 17 jun. 2021.

SANT'ANNA, F. H. et al. Emergence of the novel SARS-CoV-2 lineage VUI-NP13L and massive spread of P.2 in South Brazil. **Emerging Microbes & Infections**, v. 10, n. 1, p. 1431–1440, 1 jan. 2021.

SHEIKH, A. et al. SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. **The Lancet**, v. 397, n. 10293, p. 2461–2462, 26 jun. 2021.

SHI, J. et al. Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2. **Science**, v. 368, n. 6494, p. 1016–1020, 29 maio 2020.

SHINDE, V. et al. Efficacy of NVX-CoV2373 Covid-19 Vaccine against the B.1.351 Variant. **New England Journal of Medicine**, v. 384, n. 20, p. 1899–1909, 20 maio 2021.

SHU, Y.; MCCAULEY, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. **Eurosurveillance**, v. 22, n. 13, 30 mar. 2017.

SIMPSON, G. G.; SIMPSON, L. The meaning of evolution: a study of the history of life and of its significance for man. **Yale University Press**, v. 25, 1949.

SINGER, J. et al. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. **Preprints**. Disponível em: <https://www.preprints.org/manuscript/202006.0225/v1>. Acesso em: 18 jun. 2020.

SLAVOV, S. N. et al. Genomic monitoring unveil the early detection of the SARS-CoV-2 B.1.351 (beta) variant (20H/501Y.V2) in Brazil. **Journal of Medical Virology,** v. 93, n. 12, p. 6782–6787, 2021.

SMITH, G. J. D. et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. **Nature**, v. 459, n. 7250, p. 1122–1125, jun. 2009.

SOUZA, W. M. et al. Neutralisation of SARS-CoV-2 lineage P.1 by antibodies elicited through natural SARS-CoV-2 infection or vaccination with an inactivated SARS-CoV-2 vaccine: an immunological study. **The Lancet Microbe**, v. 2, n. 10, p. e527–e535, 1 out. 2021.

STARR, T. N. et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain

Reveals Constraints on Folding and ACE2 Binding. **Cell**, v. 182, n. 5, p. 1295- 1310.e20, 3 set. 2020.

SU, Y. C. F. et al. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. **Nature Communications**, v. 6, n. 1, p. 7952, 6 ago. 2015.

SU, Y. C. F. et al. Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. **mBio**, v. 11, n. 4, 25 ago. 2020.

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

SUN, J. et al. COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. **Trends in Molecular Medicine**, 5. v. 26, p. 483–495, 21 mar. 2020.

TAMURA, K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. **Molecular Biology and Evolution**, v. 9, n. 4, p. 678–687, jul. 1992.

TAO, K. et al. The biological and clinical significance of emerging SARS-CoV-2 variants. **Nature Reviews Genetics**, v. 22, n. 12, p. 757-773, 17 set. 2021.

TAVARE, S. Some probabilistic and statistical problems in the analysis of DNA sequences. **Lectures on mathematics in the life sciences**, v. 17, n. 2, p. 57-86, 1986.

TAY, J. H. et al. The emergence of SARS-CoV-2 variants of concern is driven by acceleration of the evolutionary rate. **MedRxiv**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.08.29.21262799v1>. Acesso em: 15 nov. 2021.

TAYLOR, L. H.; LATHAM, S. M.; WOOLHOUSE, M. E. Risk factors for human disease emergence. **Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences**, v. 356, n. 1411, p. 983–989, 29 jul. 2001.

TEGALLY, H. et al. Detection of a SARS-CoV-2 variant of concern in South Africa. **Nature**, v. 592, n. 7854, p. 438–443, abr. 2021.

TEMMAM, S. et al. Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula. **Research Square**. Disponível em: <https://www.researchsquare.com/article/rs-871965/v1>. Acesso em: 21 set. 2021

THOMSON, E. C. et al. Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. **Cell**, v. 184, n. 5, p. 1171- 1187.e20, 4 mar. 2021.

TIMMERS, L. F. S. M. et al. SARS-CoV-2 mutations in Brazil: from genomics to putative clinical conditions. **Scientific Reports**, v. 11, n. 1, p. 11998, 7 jun. 2021.

TONG, K. J. et al. A comparison of methods for estimating substitution rates from ancient DNA sequence data. **BMC Evolutionary Biology**, v. 18, n. 1, p. 70, 16 maio 2018.

TRUONG, T. T. et al. Increased viral variants in children and young adults with impaired humoral immunity and persistent SARS-CoV-2 infection: A consecutive case series. **EBioMedicine**, v. 67, p. 103355, maio 2021.

TURAKHIA, Y. et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. **Nature Genetics**, v. 53, n. 6, p. 809–816, jun. 2021.

VARELA, A. P. M. et al. SARS-CoV-2 introduction and lineage dynamics across three epidemic peaks in Southern Brazil: massive spread of P.1. **Infection, Genetics and Evolution**, v. 96, p. 105144, 17 nov. 2021.

VERHEIJE, M. H. et al. The Coronavirus Nucleocapsid Protein Is Dynamically Associated with the Replication-Transcription Complexes. **Journal of Virology**, v. 84, n. 21, p. 11575–11579, 1 nov. 2010.

VIJAYKRISHNA, D. et al. Evolutionary Insights into the Ecology of Coronaviruses. **Journal of Virology**, v. 81, n. 8, p. 4012–4020, 15 abr. 2007.

V'KOVSKI, P. et al. Coronavirus biology and replication: implications for SARS-CoV-2. **Nature Reviews Microbiology**, v. 19, n. 3, p. 155–170, mar. 2021.

VOLOCH, C. M. et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. **Journal of Virology**, 10. v. 95, p. e00119-21, 1 mar. 2021.

VOLZ, E. 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, 7 jan. 2021a.

VOLZ, E. et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. **Nature**, v. 593, n. 7858, p. 266–269, maio 2021b.

VOLZ, E. M.; KOELLE, K.; BEDFORD, T. Viral Phylodynamics. **PLoS Computational Biology**, v. 9, n. 3, 21 mar. 2013.

WALLS, A. C. et al. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. **Cell**, v. 181, n. 2, p. 281- 292.e6, 16 abr. 2020.

WALSH, D.; MOHR, I. Viral subversion of the host protein synthesis machinery. **Nature Reviews Microbiology**, v. 9, n. 12, p. 860–875, dez. 2011.

WAN, Y. et al. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. **Journal of Virology**, v. 94, n. 7, 17 mar. 2020.

WANG, P. et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. **Nature**, v. 593, n. 7857, p. 130–135, maio 2021a.

WANG, P. et al. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. **Cell Host & Microbe**, v. 29, n. 5, p. 747- 751.e4, 12 maio 2021b.

WANG, R. et al. Analysis of SARS-CoV-2 variant mutations reveals neutralization escape mechanisms and the ability to use ACE2 receptors from additional species. **Immunity**, v. 54, n. 7, p. 1611- 1621.e5, 13 jul. 2021c.

WEISBLUM, Y. et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. **eLife**, v. 9, p. e61312, 28 out. 2020.

WEISSMAN, D. et al. D614G Spike Mutation Increases SARS CoV-2 Susceptibility to Neutralization. **Cell Host & Microbe**, v. 29, n. 1, p. 23- 31.e4, 13 jan. 2021.

WHO MERS-COV RESEARCH GROUP. State of Knowledge and Data Gaps of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in Humans. **PLoS currents**, v. 5, 12 nov. 2013.

WILLETT, B. et al. The hyper-transmissible SARS-CoV-2 Omicron variant exhibits significant antigenic change, vaccine escape and a switch in cell entry mechanism. Disponível em: <https://www.gla.ac.uk/researchinstitutes/iii/cvr/engage/news/headline_829358_en.html>. Acesso em: 2 jan. 2022.

WILSON, L. et al. SARS coronavirus E protein forms cation-selective ion channels. **Virology**, v. 330, n. 1, p. 322–331, 5 dez. 2004.

WOLTER, N. et al. Early assessment of the clinical severity of the SARS-CoV-2 Omicron variant in South Africa. **MedRxiv**. Disponível em: <https://www.medrxiv.org/content/10.1101/2021.12.21.21268116v1>. Acesso em: 2 jan. 2022.

WORLD HEALTH ORGANIZATION. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. Disponível em: <https://www.who.int/publications/m/item/summary-of-probable-sars-cases-with-onset-of-illness-from-1-november-2002-to-31-july-2003>. Acesso em: 27 maio. 2021.

WORLD HEALTH ORGANIZATION. Middle East respiratory syndrome coronavirus (MERS-CoV) – Saudi Arabia. Disponível em: <http://www.who.int/csr/don/26-april-2016-mers-saudi-arabia/en/>. Acesso em: 27 maio. 2021.

WORLD HEALTH ORGANIZATION. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. Disponível em: <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>. Acesso em: 27 maio. 2021a.

WORLD HEALTH ORGANIZATION. Statement on the second meeting of the International Health Regulations (2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV). Disponível em: <https://www.who.int/news/item/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov)>. Acesso em: 28 maio. 2021b.

WORLD HEALTH ORGANIZATION. WHO Coronavirus Disease (COVID-19) Dashboard. Disponível em: <https://covid19.who.int>. Acesso em: 27 maio. 2021a.

WORLD HEALTH ORGANIZATION. Universal Health Coverage. Disponível em: <https://www.who.int/westernpacific/health-topics/universal-health-coverage>. Acesso em: 27 maio. 2021b.

WORLD HEALTH ORGANIZATION. WHO-convened global study of origins of SARS-CoV-2: China Part. Disponível em: <https://www.who.int/health-topics/coronavirus/origins-of-the-virus>. Acesso em: 27 maio. 2021c.

WORLD HEALTH ORGANIZATION. Tracking SARS-CoV-2 variants. Disponível em: <https://www.who.int/emergencies/emergency-health-kits/trauma-emergency-surgery-kit-who-tesk-2019/tracking-SARS-CoV-2-variants>. Acesso em: 30 set. 2021d.

WOROBEY, M. et al. The emergence of SARS-CoV-2 in Europe and North America. **Science (New York, N.Y.)**, v. 370, n. 6516, p. 564–570, 30 out. 2020.

WOROBEY, M. Dissecting the early COVID-19 cases in Wuhan. **Science**, v. 374, n. 6572, p. 1202-1204, 2021.

WRAPP, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. **Science**, v. 367, n. 6483, p. 1260–1263, 13 mar. 2020.

WU, Y.; ZHAO, S. Furin cleavage sites naturally occur in coronaviruses. **Stem Cell Research**, v. 50, p. 102115, 1 jan. 2021.

XAVIER, J. et al. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. **Emerging Microbes & Infections**, v. 9, n. 1, p. 1824–1834, 1 jan. 2020.

XIA, S. et al. The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. **Signal Transduction and Targeted Therapy**, v. 5, n. 1, p. 1–3, 12 jun. 2020.

XIAO, X. et al. Animal sales from Wuhan wet markets immediately prior to the COVID-19 pandemic. **Scientific Reports**, v. 11, n. 1, p. 11898, 7 jun. 2021.

YANG, Q. et al. Just 2% of SARS-CoV-2−positive individuals carry 90% of the virus circulating in communities. **Proceedings of the National Academy of Sciences**, v. 118, n. 21, 25 maio 2021.

YANG, Z. PAML 4: phylogenetic analysis by maximum likelihood. **Molecular Biology and Evolution**, v. 24, n. 8, p. 1586–1591, ago. 2007.

YANG, Z.; RANNALA, B. Molecular phylogenetics: principles and practice. **Nature Reviews Genetics**, v. 13, n. 5, p. 303–314, maio 2012.

YURKOVETSKIY, L. et al. Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. **Cell**, v. 183, n. 3, p. 739- 751.e8, 29 out. 2020.

ZAKI, A. M. et al. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. **The New England Journal of Medicine**, v. 367, n. 19, p. 1814–1820, 8 nov. 2012.

ZHANG, L. et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. **Nature Communications**, v. 11, n. 1, p. 6013, 26 nov. 2020.

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

ZHONG, N. et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. **The Lancet**, v. 362, n. 9393, p. 1353–1358, 25 out. 2003.

ZHOU, H. et al. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. **Current Biology**, 11 maio 2020a.

ZHOU, H. et al. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. **Cell**, v. 184, n. 17, p. 4380-4391.e14, 9 jun. 2021.

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

ZHU, N. et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. **New England Journal of Medicine**, 8. v. 382, p. 727–733, 24 jan. 2020.

ZUCKERKANDL, E.; PAULING, L. Molecular Disease, Evolution, and Genic Heterogeneity. **Horizons in Biochemistry**, p. 189–225, 1962.

ZUMLA, A.; HUI, D. S.; PERLMAN, S. Middle East respiratory syndrome. **The Lancet**, v. 386, n. 9997, p. 995–1007, 5 set. 2015.

CURRICULUM VITAE RESUMIDO

FRANCESCHI, VINICIUS BONETTI.

1. DADOS PESSOAIS

Nome:

Vinicius Bonetti Franceschi

Local e Data de Nascimento:

Porto Alegre, Rio Grande do Sul, Brasil, 18/06/1998.

Endereço Profissional:

Universidade Federal do Rio Grande do Sul, Centro de Biotecnologia

Avenida Bento Gonçalves, 9500 Prédio 43421, sala 221

91501-970 Porto Alegre, RS, Brasil

E-mail:

vinicius.franceschi@ufrgs.br

vinibfranc@gmail.com

2. FORMAÇÃO

2020 - Atual

Mestrado em Biologia Celular e Molecular

Universidade Federal do Rio Grande do Sul, UFRGS, Porto Alegre, Brasil

Orientadora: Claudia Elizabeth Thompson

Coorientadora: Gabriela Bettella Cybis

Bolsista: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

2016 - 2019

Graduação em Informática Biomédica

Universidade Federal de Ciências da Saúde de Porto Alegre, UFCSPA, Porto Alegre, Brasil

Orientadora: Claudia Elizabeth Thompson

Título do Trabalho de Conclusão: Análise metagenômica de dados de pacientes com infecções do Sistema Nervoso Central

3. ESTÁGIOS

2019 - 2020

Vínculo: Bolsista

Enquadramento Funcional: Bolsista de Apoio Técnico

Carga horária: 30h

Local: Universidade Federal de Ciências da Saúde de Porto Alegre

Bolsista PROPLAN/UFCSPA do Programa de acompanhamento de indicadores de produção científica e tecnológica, auxiliando no levantamento de dados para o Catálogo da Produção Científica da UFCSPA e desenvolvendo um sistema para registro centralizado e recuperação de indicadores da Pró-Reitoria de Pesquisa e Pós Graduação (ProPPG-UFCSPA).

Supervisoras: Dinara Jaqueline Moura e Marcia Giovenardi

2019 - 2019

Vínculo: Estagiário

Enquadramento Funcional: Estágio Curricular Obrigatório

Carga horária: 20h

Local: Universidade Federal do Rio Grande do Sul

Desenvolvimento de uma plataforma para consulta e integração de dados de expressão diferencial, anotação funcional e ortologia de organismos proximamente relacionados.

Supervisor: Charley Christian Staats

2017 - 2018

Vínculo: Bolsista

Enquadramento Funcional: Iniciação Científica (PIC/UFCSPA)

Carga horária: 20

Desenvolvimento do aplicativo móvel chamado MySleep, o qual monitora a noite de sono do usuário, auxiliando, assim, o diagnóstico e tratamento de distúrbios do sono. Além disso, foi desenvolvido um portal web para acesso

às métricas coletadas pelo aplicativo em dashboards e gráficos que facilitam a interpretação e tomada de decisão dos especialistas.
Orientador: Cristiano Bonato Both

4. PRÊMIOS E DISTINÇÕES

2020 5º tweet mais curtido na Reunião Anual do Programa de Pós-Graduação em Biologia Celular e Molecular, UFRGS.
2019 Destaque da sessão 002 (Iniciação Científica - Ciência da Computação - Informática Biomédica) no Congresso UFCSPA: conectando saúde e sociedade, UFCSPA.
2019 Menção Honrosa: Finalista do Prêmio Divulgação Acadêmica na categoria Pesquisa no Congresso UFCSPA: conectando saúde e sociedade, UFCSPA.

5. PROJETOS DE PESQUISA

2020 - Atual
Percurso Epidemiológico, Genômico e Clínico do vírus SARS-CoV-2 causador de COVID-19
Natureza: Pesquisa.
Alunos envolvidos: Graduação: (7) / Mestrado acadêmico: (7) / Doutorado: (3) .
Integrantes: Vinicius Bonetti Franceschi - Integrante / Claudia Elizabeth Thompson - Coordenador / Charley Christian Staats - Integrante / Nêmora Tregnano Barcellos - Integrante / Lívia Kmetzsch Rosa e Silva - Integrante / Ana Trindade Winck - Integrante / Carla Diniz Lopes Becker - Integrante / Cecília Dias Flores - Integrante / Cristine Souza Goebel - Integrante / Filipe Santana da Silva - Integrante / Liane Nanci Rotta - Integrante / Luciano Costa Blomberg - Integrante / Luiz Carlos Rodrigues Junior - Integrante / Paulo Ricardo Gazzola Zen - Integrante / Pedro Roosevelt Torres Romão - Integrante / Thatiane Alves Pianoschi Alva - Integrante / Viviane Rodrigues Botelho - Integrante / Alice da Rocha Nascimento - Integrante / Carem Luana Machado Lessa - Integrante / Júlia Geremias Martins - Integrante / Júlia

Gonçalves Küchle - Integrante / Sofia Faber Silveira - Integrante / Andressa Schneiders Santos - Integrante / Andressa Barreto Glaeser - Integrante / Francyelle Diones da Silva Oliveira - Integrante / Gabriel Dickin Caldana - Integrante / Gustavo Henrique Cervi - Integrante / Igor Martins da Silva - Integrante / Fábio Brambilla - Integrante / Andréa Aparecida Konzen - Integrante / Alvaro Vigo - Integrante / Edison Pignaton de Freitas - Integrante / Esequia Sauter - Integrante / Fabio Souto de Azevedo - Integrante / Gabriela Bettella Cybis - Integrante / Marilene Henning Vainstein - Integrante / Gabriela Prado Paludo - Integrante / Meiski Mariá Vedovatto - Integrante / Amanda de Menezes Mayer - Integrante / Elson Romeu Farias - Integrante / José Antônio Tesser Poloni - Integrante / Marcos Pascoal Pattussi - Integrante / Maria Leticia Rodrigues Ikeda - Integrante / Viviane Schmitt Jahnke - Integrante / Fernando Rosado Spilki - Integrante / Juliane Deise Fleck - Integrante.

2020 - Atual

Inovação em Diagnóstico Molecular de Infecções de difícil tratamento do Sistema Nervoso Central

Descrição: Projeto aprovado no Edital CNPq/AWS Nº 032/2019 - Acesso às Plataformas de Computação em Nuvem da Empresa AMAZON Web Service (Cloud Credits for Research)

Natureza: Pesquisa.

Alunos envolvidos: Graduação: (4) / Mestrado acadêmico: (3) / Doutorado: (1) .

Integrantes: **Vinicius Bonetti Franceschi - Integrante** / Claudia Elizabeth Thompson - Coordenador / Ana Trindade Winck - Integrante / Carla Diniz Lopes Becker - Integrante / Cecília Dias Flores - Integrante / Luciano Costa Blomberg - Integrante / Paulo Ricardo Gazzola Zen - Integrante / Viviane Rodrigues Botelho - Integrante / Alice da Rocha Nascimento - Integrante / Júlia Geremias Martins - Integrante / Júlia Gonçalves Küchle - Integrante / Sofia Faber Silveira - Integrante / Francyelle Diones da Silva Oliveira - Integrante / Gustavo Henrique Cervi - Integrante / Meiski Mariá Vedovatto - Integrante / Guilherme Loss de Moraes - Integrante / Maria Ismenia Zulian Lionzo - Integrante / Paulo Valdeci Worm - Integrante / Ricardo Gurgel Rebouças -

Integrante / Felipe Martins de Lima Cecchini - Integrante / Luiz Felipe Valter de Oliveira - Integrante / Daniela Carla Lunelli - Integrante / Marcelo Martins dos Reis - Integrante / Ana Paula Christoff - Integrante.

2017 - 2018

angELO: Uma Plataforma baseada em computação em nuvem para Mobile Health

Natureza: Pesquisa.

Alunos envolvidos: Graduação: (2) / Mestrado acadêmico: (1)

Integrantes: **Vinicius Bonetti Franceschi - Integrante** / Cristiano Bonato Both - Coordenador / Carlos Fabiel Bublitz - Integrante / Ana Carolina Ribeiro Teixeira - Integrante / Ícaro Maia Santos de Castro - Integrante.

6. ARTIGOS COMPLETOS PUBLICADOS EM PERIÓDICOS

FRANCESCHI, VINÍCIUS BONETTI; FERRAREZE, PATRÍCIA ALINE GRÖHS ; ZIMERMAN, RICARDO ARIEL ; CYBIS, GABRIELA BETTELLA ; THOMPSON, CLAUDIA ELIZABETH . Mutation hotspots and spatiotemporal distribution of SARS-CoV-2 lineages in Brazil, February 2020-2021. VIRUS RESEARCH, v. 304, p. 198532, 2021. Citações: 10.

FRANCESCHI, VINÍCIUS BONETTI; CALDANA, GABRIEL DICKIN; PERIN, CHRISTIANO; HORN, ALEXANDRE; PETER, CAMILA; CYBIS, GABRIELA BETTELLA; FERRAREZE, PATRÍCIA ALINE GRÖHS; ROTTA, LIANE NANCI; CADEGIANI, FLÁVIO ADSUARA; ZIMERMAN, RICARDO ARIEL; THOMPSON, CLAUDIA ELIZABETH. Predominance of the SARS-CoV-2 Lineage P.1 and Its Sublineage P.1.2 in Patients from the Metropolitan Region of Porto Alegre, Southern Brazil in March 2021. PATHOGENS, v. 10, p. 988, 2021. Citações: 4.

FERRAREZE, PATRÍCIA ALINE GRÖHS; **FRANCESCHI, VINÍCIUS BONETTI**; MAYER, AMANDA DE MENEZES; CALDANA, GABRIEL DICKIN; ZIMERMAN, RICARDO ARIEL; THOMPSON, CLAUDIA ELIZABETH. E484K as an innovative phylogenetic event for viral evolution: Genomic analysis of

the E484K spike mutation in SARS-CoV-2 lineages from Brazil. INFECTION GENETICS AND EVOLUTION, v. 93, p. 104941, 2021. Citações: 43.

FRANCESCHI, VINÍCIUS BONETTI; CALDANA, GABRIEL DICKIN; DE MENEZES MAYER, AMANDA; CYBIS, GABRIELA BETTELLA; NEVES, CARLA ANDRETTA MOREIRA; FERRAREZE, PATRÍCIA ALINE GRÖHS; DEMOLINER, MERIANE; DE ALMEIDA, PAULA RODRIGUES; GULARTE, JULIANA SCHONS; HANSEN, ALANA WITT; WEBER, MATHEUS NUNES; FLECK, JULIANE DEISE; ZIMERMAN, RICARDO ARIEL; KMETZSCH, LÍVIA; SPILKI, FERNANDO ROSADO; THOMPSON, CLAUDIA ELIZABETH. Genomic epidemiology of SARS-CoV-2 in Esteio, Rio Grande do Sul, Brazil. BMC GENOMICS, v. 22, p. 371, 2021. Citações: 12.

FRANCESCHI, Vinicius Bonetti; SANTOS, Andressa Schneiders; GLAESER, Andressa Barreto; PAIZ, Janini Cristina; CALDANA, Gabriel Dickin; MACHADO LESSA, Carem Luana; MAYER, Amanda de Menezes; KÜCHLE, Júlia Gonçalves; ZEN, Paulo Ricardo Gazzola; VIGO, Alvaro; WINCK, Ana Trindade; ROTTA, Liane Nanci; THOMPSON, Claudia Elizabeth. Population- based prevalence surveys during the Covid- 19 pandemic: A systematic review. Reviews in Medical Virology, v. 31, p. e2200, 2020. Citações: 17.

DECONTE, Desirée; DOS SANTOS, Catarine Benta Lopes; DE MORAIS, Camila Ohomoto; YONAMINE, Tatiane Mayumi; NOGUEIRA, Letícia Thaís; FERREIRA, Maria Angélica Tosi; **FRANCESCHI, Vinicius Bonetti**; LONGHI, André Luís Soares; VILLACIS, Rolando André Rios; ROGATTO, Silvia Regina; LIGABUE-BRAUN, Rodrigo; ZEN, Paulo Ricardo Gazzola; MACHADO ROSA, Rafael Fabiano; FIEGENBAUM, Marilu. Unusual features in a child with Marshall-Smith syndrome due to a novel NFIX variant: Evidence for an abnormal protein function. Gene Reports, v. 22, p. 100991, 2021.

7. RESUMOS E TRABALHOS APRESENTADOS EM CONGRESSOS

FRANCESCHI, VINÍCIUS BONETTI; FERRAREZE, PATRÍCIA ALINE GRÖHS ; ZIMERMAN, RICARDO ARIEL ; CYBIS, GABRIELA BETTELLA ; THOMPSON, CLAUDIA ELIZABETH . Mutation hotspots and spatiotemporal distribution of SARS-CoV-2 lineages in Brazil, February 2020-2021. In: 5th Workshop on Virus Dynamics, Seattle (Estados Unidos). 2021. (Apresentação de trabalho).

FRANCESCHI, VINÍCIUS BONETTI; CALDANA, GABRIEL DICKIN; PERIN, CHRISTIANO; HORN, ALEXANDRE; PETER, CAMILA; CYBIS, GABRIELA BETTELLA; FERRAREZE, PATRÍCIA ALINE GRÖHS; ROTTA, LIANE NANCI; CADEGIANI, FLÁVIO ADSUARA; ZIMERMAN, RICARDO ARIEL; THOMPSON, CLAUDIA ELIZABETH. Broad distribution of the P.1 SARS-CoV-2 lineage and its sublineage P.1.2 in patients from Southern Brazil in March 2021: a phylogenomic analysis. In: Virus Genomics and Evolution Conference, Londres (Inglaterra). 2021. (Apresentação de trabalho).

FRANCESCHI, Vinicius Bonetti; THOMPSON, Claudia Elizabeth. Pipeline metagenômico como ferramenta para diagnóstico de neuroinfecções. In: Congresso UFCSPA: conectando saúde e sociedade, 2019, Porto Alegre. Anais do Congresso UFCSPA: conectando saúde e sociedade, 2019.

FRANCESCHI, Vinicius Bonetti; THOMPSON, Claudia Elizabeth. Metagenomic pipeline for pathogen detection in neuroinfections. In: Escola Gaúcha de Bioinformática (EGB) 2019, Porto Alegre. 2019. (Apresentação de Trabalho).

FRANCESCHI, Vinicius Bonetti; BUBLITZ, Carlos Fabiel; TEIXEIRA, Ana Carolina Ribeiro; CASTRO, Ícaro Maia Santos; BOTH, Cristiano Bonato. MySleep: A Sleep Monitoring Application. In: IV Mostra de Ensino, Pesquisa e Extensão da UFCSPA, Porto Alegre. 2018. (Apresentação de Trabalho).

APÊNDICES

Apresento como apêndices alguns artigos publicados no período do Mestrado relacionados à COVID-19, como coautor ou em outros aspectos da doença.

APÊNDICE 1

O manuscrito que constitui o Apêndice 1, intitulado “E484K as an innovative phylogenetic event for viral evolution: Genomic analysis of the E484K spike mutation in SARS-CoV-2 lineages from Brazil” objetivou caracterizar o aumento na proporção de genomas brasileiros com a mutação E484K na proteína *Spike* e avaliar sítios positivamente selecionados nesta proteína. Encontra-se publicado na revista Infection, Genetics and Evolution (https://www.sciencedirect.com/journal/infection-genetics-and-evolution), com fator de impacto JCR 2021 = 3,342 e Qualis/CAPES = A2. O manuscrito e os materiais suplementares estão disponíveis mediante assinatura no seguinte *link*: https://www.sciencedirect.com/science/article/abs/pii/S1567134821002380.

APÊNDICE 2

O manuscrito que constitui o Apêndice 2, intitulado “Population-based prevalence surveys during the Covid-19 pandemic: A systematic review” objetivou avaliar aspectos qualitativos de estudos populacionais que estimassem a prevalência da COVID-19, identificando seu grau de confiabilidade e práticas ou variáveis relacionadas à qualidade metodológica. Encontra-se publicado na revista Reviews in Medical Virology (https://onlinelibrary.wiley.com/journal/10991654), com fator de impacto JCR 2021 = 6,989 e Qualis/CAPES = A1. O manuscrito e os materiais suplementares estão disponíveis na íntegra (*Open access*) no seguinte *link*: https://onlinelibrary.wiley.com/doi/full/10.1002/rmv.220