UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS BIOLÓGICAS

DEPARTAMENTO DE GENÉTICA, ECOLOGIA E EVOLUÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

DISSERTAÇÃO DE MESTRADO

VIGILÂNCIA GENÔMICA DO SARS-CoV-2 EM BETIM, MINAS GERAIS

ORIENTADO: Diego Menezes Bonfim

ORIENTADOR: Dr. Renato Santana De Aguiar

COORIENTADOR: Dr. Renan Pedra De Souza

BELO HORIZONTE

Dezembro de 2021

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE GENÉTICA, ECOLOGIA E EVOLUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

Vigilância genômica do SARS-CoV-2 em BETIM, Minas Gerais

Diego Menezes Bonfim

Dissertação submetida ao programa de Pós-graduação em Genética da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do título de Mestre em Genética.

Orientador: Dr. Renato Santana de Aguiar
Coorientador: Dr. Renan Pedra de Souza

Área de concentração: genética molecular, de microrganismos e biotecnologia

043 Bonfim, Diego Menezes.
Vigilância Genômica do SARS-CoV-2 em Betim-MG. [manuscrito] / Diego Menezes Bonfim. – 2021.
70 f. : il. ; 29,5 cm.

Orientador: Dr. Renato Santana de Aguiar. Coorientador: Dr. Renan Pedra de Souza.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genética. 2. Genômica. 3. COVID-19. 4. Filogenia. I. Aguiar, Renato Santana de. II. Souza, Renan Pedra de. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 575

Ficha catalográfica elaborada por Cleide Vieira de Faria – CRB: 6/2314

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Genética

ATA DE DEFESA DE DISSERTAÇÃO

ATA DA DEFESA DE DISSERTAÇÃO	**318/2021** **Entrada 1º/2020**
Diego Menezes Bonfim	**CPF: 123.219.026-80**

Às quatorze horas do dia **06 de dezembro de 2021**, reuniu-se remotamente a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Vigilância Genômica do SARS-CoV-2 em Betim-MG"**, requisito para obtenção do grau de Mestre em **Genética.** Abrindo a sessão, o Presidente da Comissão, **Renato Santana de Aguiar**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	**Instituição**	**CPF**	**Indicação**
Renato Santana de Aguiar	UFMG	000.086.336-06	Aprovado
Fabrício Rodrigues dos Santos	UFMG	567.487.446-87	Aprovado
Felipe Campos de Melo Iani	Fundação Ezequiel Dias	015.130.086-09	Aprovado

Pelas indicações, o candidato foi considerado: APROVADO

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

Belo Horizonte, 06 de dezembro de 2021.

Renato Santana de Aguiar (UFMG)

Fabrício Rodrigues dos Santos (UFMG)

Felipe Campos de Melo Iani (Fundação Ezequiel Dias)

Assinatura dos membros da banca examinadora:

Documento assinado eletronicamente por **Renato Santana de Aguiar**, **Professor do Magistério Superior**, em 06/12/2021, às 16:33, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Fabricio Rodrigues dos Santos**, **Membro de comissão**, em 06/12/2021, às 17:36, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Felipe Campos de Melo Iani**, **Usuário Externo**, em 09/12/2021, às 14:33, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1129921** e o código CRC **B7DCA0A8**.

Referência: Processo nº 23072.262674/2021-91 SEI nº 1129921

AGRADECIMENTOS

Agradeço aos meus pais, Ernane e Petrânia, por me proporcionarem um ambiente de criatividade e fertilidade mental em meio às dificuldades da vida, atributo necessário ao alcance dos sonhos.

Ao meu irmão, Tiago, por sempre me ajudar a manter minha mente presa à realidade, atributo igualmente necessário.

À todos os professores do ensino básico, que de alguma forma me inspiraram ou acreditaram em mim.

Ao meu melhor amigo, Thales, que plantou em mim a semente da ciência e lutamos juntos desde os 9 anos de idade. Vencemos!

Aos camaradas da graduação, em especial Alexandre, Matheus e Rafael por estarem presentes em todos os momentos e por tudo o que vivemos juntos.

Aos camaradas do laboratório, em especial ao João, pela valiosíssima colaboração e fraternidade.

Ao meu orientador, Renato, e ao meu coorientador, Renan, por serem ambos exemplo de ética e método científico e pela mão sempre estendida para me ajudar, possibilitando que eu conciliasse trabalho e carreira acadêmica e chegasse até aqui. Obrigado!

Ao Álvaro e ao Braulino, fontes eternas de luz e amor.

E especialmente à Cíntia, amor da minha vida e minha esposa, por criar e manter comigo um universo de realidades que seriam impossíveis de serem conquistadas de maneira isolada, provando que a força de 1 + 1 é maior que 2, ou como diria Aristóteles, que "*o todo é maior que a simples soma de suas partes*."

EPÍGRAFE

“The other side of this coin is that viruses may be genes who have broken loose from 'colonies' such as ourselves. (…) If this is true, we might just as well regard ourselves as colonies of viruses!”

(Richard Dawkins)

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIAÇÕES

2019-nCoV - 2019 novel Coronavirus

ACE-2 - Angiotensin-converting enzyme - 2

B814 - espécime de número B814

BR-040 - Rodovia interestadual 040

BR-262 - Rodovia interestadual 262

BR-381 - Rodovia interestadual 381

CDC - Centers for Disease Control and Prevention

cDNA - complementary DNA

COVID-19 - Coronavirus disease 2019

CoVs - Coronavírus

CT - Cycle Threshold

DNA - Deoxyribonucleic Acid

E - envelope

HCoV-229E - Human Coronavirus - 229E

HCoV-HKU1 - Human Coronavirus - HKU1

HCoV-NL63 - Human Coronavirus - NL63

HCoV-OC43 - Human Coronavirus - OC43

MERS-CoV Middle east respiratory syndrome - Coronavirus

MERS - Middle east respiratory syndrome

MG-050 - Rodovia intermunicipal 050

MG-060 - Rodovia intermunicipal 060

MG - Estado de Minas Gerais

M - membrane

N - nucleocapsid

NGS - Next-Generation Sequencing

NPIs - Nonpharmaceutical Interventions

nsp2 - non structural protein 2

OMS - Organização Mundial de Saúde

ORF - Open Reading Frame

RaTG13 - *Rhinolophus affinis* Tongguan 2013

RBD - Receptor Binding Domain

RdRp - RNA dependent RNA polymerase

RmYN02 - *Rhinolophus malayanus* Yunnan - consensus sequence 02

RNA - Ribonucleic acid

RNAseP - Ribonuclease P

RNPs - Ribonucleoproteins

RT-qPCR - Reverse Transcriptase - quantitative Polymerase Chain Reaction

S_1 - Spike protein subunit 1

S_2 - Spike protein subunit 2

SARS-CoV-2 - Severe acute respiratory syndrome - Coronavirus - 2

SARS-CoV - Severe acute respiratory syndrome - Coronavirus

SARS - Severe acute respiratory syndrome

SNP - Single Nucleotide Polymorphism

SP - Estado de São Paulo

S - spike

TCLE - Termo de Consentimento Livre e Esclarecido

TMPRSS2 - Transmembrane serine protease 2

TRSs - Transcriptional Regulatory Sequences

VOC - Variant of Concern

VOI - Variant of Interest

RESUMO

Coronavírus (CoVs) são vírus de RNA fita positiva que constituem a subfamília Orthocoronavirinae. Sete diferentes tipos de CoV são capazes de infectar humanos, dentre elas o Betacoronavirus SARS-CoV-2, proximamente relacionado a linhagens que infectam morcegos, sua provável origem zoonótica. O novo coronavírus possui polimorfismos em seu genoma especialmente na sua proteína de superfície spike, utilizada para infectar células humanas através do receptor ACE-2 sendo associado ao desenvolvimento da nova doença COVID-19, de caráter pandêmico. Tendo origem na China no fim de dezembro de 2019, o SARS-CoV-2 teve seu primeiro caso de infecção confirmada no Brasil no dia 26 de fevereiro de 2020 e no estado de Minas Gerais no dia 04 de março de 2020. Estratégias constantes de vigilância genômica são importantes no controle da transmissão e dispersão do vírus na população humana. Com o surgimento de variantes de interesse (VOI) e de preocupação (VOC) do SARS-CoV-2, o cenário local e global da pandemia se tornou mais complexo, o que reforça a necessidade dessa estratégia. O objetivo dessa dissertação foi realizar a vigilância genômica do SARS-CoV-2 na cidade de Betim-MG, quinta mais populosa do estado, com importância industrial e atravessada por estradas estaduais e federais recebendo um fluxo migratório humano considerável. Para isso, um total de 3239 amostras de swab naso-faríngeo coletadas de junho a julho de 2020 foram testadas por RT-qPCR para o SARS-CoV-2, triadas para sequenciamento na plataforma MiSeq (Illumina) através de uma metodologia baseada em amplificação com iniciadores cobrindo todo o genoma de SARS-CoV-2. No final, foram obtidos 35 novos genomas de SARS-CoV-2 (cobertura > 79%; profundidade > 200x). Estes genomas foram incluídos em reconstruções filogenéticas para identificação de linhagens virais circulantes, modelos filogeográficos datados, interpolação para análise de padrões de dispersão e mapas genéticos para visualização de polimorfismos. Dos 35 genomas obtidos, 18 (51,4%) foram classificados como linhagem B.1.1.28 e 17 (48,6%) como linhagem B.1.1.33, circulantes na primeira fase da pandemia do estado de MG. O modelo filogenético não-enraizado confirmou as linhagens e sugeriu múltiplas introduções para ambos os clados. Os modelos filogeográficos sugeriram pelo menos 7 e 12 introduções distintas para as linhagens B.1.1.28 e B.1.1.33, respectivamente. Os modelos de dispersão geográfica mostram a propagação da pandemia ao longo do tempo, mas não sugerem padrões geográficos de cada linhagem. Destacamos um genoma B.1.1.33 que possui 12 mutações específicas e que está posicionado com considerável distância em comparação aos demais no modelo filogeográfico desta linhagem. Os resultados sugerem um fluxo interestadual de migração da população humana na dispersão do SARS-CoV-2, fato que deve ser levado em conta na adoção de medidas de controle da pandemia na cidade de Betim-MG.

Palavras-chave: 2019-nCoV; COVID-19; genômica; filogenia; variantes, epidemiologia

ABSTRACT

Coronaviruses (CoVs) are positive strand RNA viruses that constitute the subfamily Orthocoronavirinae. Seven different types of CoV are capable of infecting humans, including the Betacoronavirus SARS-CoV-2, closely related to strains that infect bats, its probable zoonotic origin. The new coronavirus has polymorphisms in its genome, specially in its gene encoding the spike surface protein, used to infect human cells through the ACE-2 receptor, and is associated with the development of the new pandemic disease named COVID-19. Originating in China at the end of December 2019, SARS-CoV-2 had its first confirmed case of infection in Brazil on February 26, 2020 and in the state of Minas Gerais on March 4, 2020. Constant genomic surveillance strategies are important to control the transmission and spread of the virus in the human population. With the emergence of SARS-CoV-2 variants of interest (VOI) and concern (VOC), the local and global scenario of the pandemic has become more complex, which reinforces the need for this type of study. The goal of this dissertation was to carry out the genomic surveillance of SARS-CoV-2 in the city of Betim-MG, the fifth most populous city in the state, with industrial importance and crossed by state and federal roads receiving a considerable human migratory flow. For this, a total of 3239 nasopharyngeal swab samples collected from June to July 2020 were tested by RT-qPCR for SARS-CoV-2, sorted for sequencing on the MiSeq platform (Illumina) using an amplification-based methodology with primers covering the whole SARS-CoV-2 genome. In the end, 35 new SARS-CoV-2 genomes were obtained (coverage > 79%; depth > 200x). These genomes were included in phylogenetic reconstructions to identify circulating viral lineages, dated phylogeographic models, interpolation to analyze dispersion patterns and genetic maps to visualize polymorphisms. Of the 35 genomes obtained, 18 (51.4%) were classified as lineage B.1.1.28 and 17 (48.6%) as lineage B.1.1.33, circulating in the first phase of the pandemic in the state of MG. The unrooted phylogenetic model confirmed the lineages and suggested multiple introductions for both clades. The phylogeographic models suggested at least 7 and 12 distinct introductions for the B.1.1.28 and B.1.1.33 lineages, respectively. Geographic dispersion models show the spread of the pandemic over time, but do not suggest geographic patterns for each lineage. We highlight a B.1.1.33 genome that presents 12 specific mutations and that is positioned at a considerable distance compared to the others in the phylogeographic model of this strain. The results suggest an interstate migration flow of the human population in the dispersion of SARS-CoV-2, fact that must be taken into account when adopting measures to control the pandemic in the city of Betim-MG.

Keywords: 2019-nCoV; genetics; virology; phylogeny; epidemiology.

1. INTRODUÇÃO

1.1. CORONAVÍRUS

Coronavírus (CoVs) são os membros da subfamília *Orthocoronavirinae*, que juntamente com a subfamília *Letovirinae* compõem a família *Coronaviridae* (ordem Nidovirales), descrita pela primeira vez no final dos anos 60[1]. Os representantes deste grupo são vírus envelopados que possuem fita única de RNA com polaridade positiva[2] capazes de infectar hospedeiros de diferentes ordens: de invertebrados a aves e mamíferos[3]. Até o presente momento, levando em conta o novo coronavírus, sete tipos de CoVs mostraram-se capazes de infectar seres humanos[4].

A estrutura geral da partícula viral dos coronavírus consiste em um envelope esférico composto de membrana bilipídica, proteínas envelope (E) e de membrana (M) e atravessado por proteínas do tipo *spike* (S), que promovem a entrada celular destes vírus. Internamente há o nucleocapsídeo (N): uma fita de RNA genômico ligada a proteínas do tipo N organizada de maneira espiralar. O formato destas estruturas, em especial o envelope e as proteínas *spike* faz com que a partícula viral microscopicamente se assemelhe ao halo ou corona solar, característica que deu nome ao grupo. O tamanho médio do genoma dos coronavírus é o maior dentre os vírus de RNA, indo de 26 a 32kb. O gene que codifica a proteína replicase (RNA polimerase dependente de RNA) é o mais extenso, ocupando 2/3 do genoma. A organização genômica é conservada na subfamília: a sequência que codifica a replicase (ORF1ab) aparece seguida da sequência da proteína *spike* (S), da proteína do envelope (E), da proteína de membrana (M) e finalmente da proteína do nucleocapsídeo (N) no sentido 5' – 3'[3]. A identidade nucleotídica destas regiões também é altamente conservada entre diferentes linhagens de CoVs, isto se deve principalmente ao fato da RNA polimerase dependente de RNA (RdRp) dos vírus da ordem Nidovirales possuir capacidade revisora através da atividade de exonuclease 3'-5' promovida por esta proteína[5].

Atualmente a subfamília *Orthocoronavirinae* divide-se em quatro gêneros: os *Alpha-, Beta-, Gamma- e Deltacoronavírus*[2]. A classificação mais recente ocorreu em 2009, sendo os três primeiros clados derivados de uma nomenclatura anterior: *Alpha-, Beta-,* e *Gammacoronavirus* oriundos dos Coronavírus grupo 1, 2 e 3, respectivamente. Tal classificação agrupava os patógenos pelo tipo de hospedeiro: os grupos 1 e 2 compreendiam coronavírus que infectam mamíferos, ao passo que o grupo 3 era composto de patógenos que infectam aves. Apesar desta correlação entre gênero e tipo de hospedeiro permanecer significativa, estudos posteriores mostraram que alguns dos gêneros atuais, como os *Alphacoronavirus* são bastante parafiléticos e que estamos longe de entender completamente

a sua origem. O gênero *Deltacoronavirus* é o mais recente, tendo emergido após a mudança na nomenclatura[6,7].

1.2. INFECÇÕES POR CORONAVÍRUS HUMANOS

Os coronavírus isolados em humanos com reconhecida importância médica totalizam sete diferentes tipos, das quais duas estão classificadas no gênero *Alphacoronavirus*: HCoV-NL63 e HCoV-229E. Todas as demais: HCoV-OC43, HCoV-HKU1, SARS-CoV, MERS-CoV e mais recentemente o SARS-CoV-2 estão incluídas no gênero *Betacoronavirus*[2,8]. As linhagens HCoV-229E e HCoV-OC43 são conhecidas desde o final dos anos 60. Outra possível linhagem (B814) chegou a ser isolada e cultivada previamente. Desde a época, estas infecções foram caracterizadas pela transmissão por vias aéreas e manifestação de sintomas comuns de resfriado[9-11].

O próximo coronavírus humano a ser conhecido foi descrito apenas em 2003 na China, tendo sua provável origem associada ao consumo de carne de civetas. SARS-CoV, também transmitido por vias aéreas, foi o responsável por uma epidemia de consequências catastróficas. Assim como o SARS-CoV-2, este vírus utiliza o receptor ACE-2 para infectar as células do epitélio pulmonar humano[2,12]. Progressivamente, pacientes infectados por SARS evoluíam de um quadro clínico de resfriado comum com febre, tosse seca, dispneia e dor de cabeça para sintomas mais graves como hipoxemia e falência respiratória causados pelo dano alveolar, causando os piores danos em indivíduos idosos e/ou com comorbidades. 10% dos 8096 casos confirmados vieram a óbito[13,14]. As consequências impactantes da SARS fizeram com que a comunidade médica e científica atribuísse maior importância aos CoVs. Com o aumento da pesquisa na área, dois novos coronavírus que infectam o trato respiratório humano foram descritos nos anos seguintes: a linhagem HCoV-NL63, em 2004 e a HCoV-HKU1 em 2005. Ambos são transmitidos por vias aéreas e normalmente desencadeiam apenas casos de resfriado comum, embora a linhagem HCoV-HKU1 tenha sido descrita pela primeira vez em pacientes com quadro de pneumonia[2,15,16].

Uma década após a epidemia de SARS, o sexto CoV humano foi capaz de desencadear outra epidemia devastadora. MERS-CoV foi descrito pela primeira vez em um paciente com pneumonia na Arábia Saudita em 2012. Tendo se originado em camelídeos, esta segue sendo sua principal fonte de infecção uma vez que a transmissão humana por vias aéreas ocorre com transmissão limitada. O quadro clínico da MERS é o mais amplo dentre as infecções causadas por coronavírus, indo desde a ausência de sintomas à falha respiratória aguda seguida de morte. Tipicamente, indivíduos infectados apresentam febre igual ou superior à 38°C acompanhada de calafrios. Outros sintomas comuns incluem tosse,

dispneia, falta de ar, náusea, vômito, diarreia, mialgia generalizada, mal estar, sonolência, confusão mental, quadros clássicos de pneumonia, falha renal, inflamação do pericárdio e coagulopatia. Até o presente momento, 34% dos 2519 casos registrados vieram a óbito, a maior taxa de mortalidade de uma infecção causada por coronavírus[14,17,18].

1.3. SARS-CoV-2

1.3.1. ORIGEM

Ainda que atualmente não se possa afirmar com exatidão a origem do novo coronavírus, desde sua descrição o patógeno 2019-nCoV foi incluído no subgênero *Sarbecovirus* (Subfamília Orthocoronavirinae; Gênero *Betacoronavirus*), tendo sido posteriormente nomeado SARS-CoV-2[4,8]. O sétimo coronavírus humano, descrito no início de 2020, é proximamente relacionado com linhagens virais que infectam pangolins, morcegos e outros animais silvestres. A carne e outros materiais biológicos destes animais são encontrados em mercados de produtos tradicionais, para consumo alimentício ou fabricação de medicamentos alternativos em comércios de algumas cidades da China, o que facilita o contato e estabelecimento destes vírus em humanos[19]. Duas linhagens virais mostraram-se especialmente relacionadas ao causador da COVID-19: RaTG13[20], com tempo de divergência entre esta espécie e o SARS-CoV-2 de aproximadamente 52 anos e RmYN02[21], com tempo de divergência entre esta espécie e o SARS-CoV-2 de aproximadamente 37 anos, de acordo com estudos filogenéticos. Ambas foram isoladas de morcegos do gênero *Rhinolophus*. O enquadramento filogenético do SARS-CoV-2 com base nestas linhagens, no entanto, tem sido alvo de questionamentos. Fatores como as fortes pressões seletivas na passagem de um tipo de hospedeiro a outro, a possibilidade de recombinação intra-hospedeiro com outros vírus e a alta taxa de mutação inerente aos vírus de genoma de RNA, fazem com que a modelagem da evolução destas linhagens ajustada pelo relógio molecular seja um processo difícil. Tais fatores são potenciais fontes de viés que devem ser consideradas no processo de inferência filogenética[22].

1.3.2. MORFOLOGIA

SARS-CoV-2 possui o "*Bauplan"* tradicional dos Orthocoronavirinae. Isto implica dizer que as estruturas canônicas dos demais CoVs estão presentes. Isto inclui o envelope esférico composto de camada bilipídica, proteínas envelope (E) e membrana (M), que atinge 64,8-96,6nM de diâmetro nesta linhagem viral, atravessado por proteínas do tipo *spike* (S) em formato de corona solar. No caso do SARS-CoV-2, chama atenção o núcleocapsídeo (N), que

é composto de 30 a 35 Ribonucleoproteínas (RNPs) em formato de G reverso de 15nm de diâmetro por vírion[23]. O novo coronavírus, assim como os demais membros de sua subfamília, promove a entrada celular a partir do reconhecimento de proteínas receptoras de membrana do hospedeiro pelas suas proteínas *spike* (S). A glicoproteína *spike* (S) dos coronavírus é composta de 2 subunidades: S_1, que faz contato direto com o receptor celular do hospedeiro e S_2, que promove a fusão das membranas viral e celular. Estas proteínas formam um homotrímero em formato de Y que atravessa o envelope viral[24]. Os receptores celulares são exclusivos para cada hospedeiro, portanto, a estrutura destas proteínas varia de maneira inter- e intra-específica. Em geral, as regiões codificadoras do gene S possuem a maior diversidade entre as diferentes linhagens virais e sofrem a maior pressão seletiva por serem reconhecidas pela resposta imunológica de seus hospedeiros. De fato, um estudo evolutivo comparando sequências de aminoácidos encontrou uma alta divergência entre a proteína *spike* (S) do SARS-CoV-2 em comparação com outros CoVs proximamente relacionados[20] (Fig. 1).

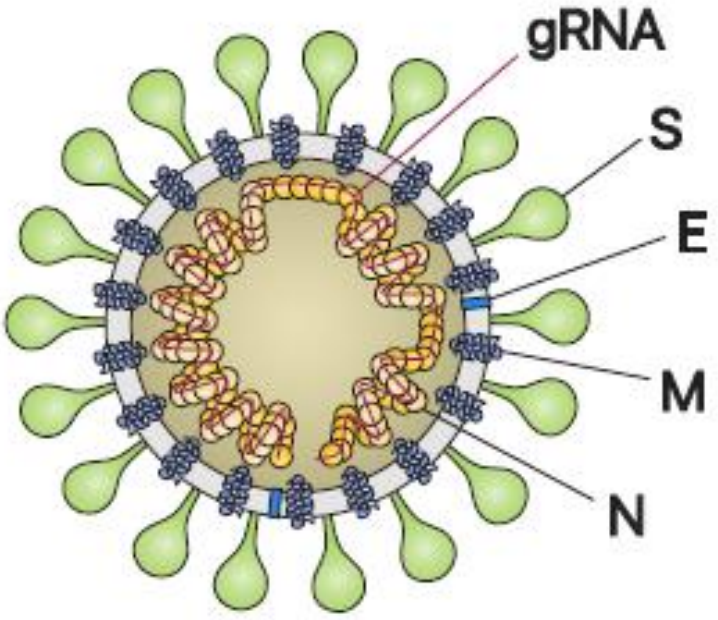

Figura 1. Representação gráfica da estrutura morfológica do SARS-CoV-2. Estão presentes: o RNA genômico (gRNA), as proteínas *Spike* (S), do envelope (E), membrana (M) e as ribonucleoproteínas que compõem o núcleocapsídeo (N) (adaptado de Kim et al, 2020)[41].

1.3.3. GENOMA

O genoma do SARS-CoV-2 é altamente conservado em relação aos demais CoVs. A ordem que as principais regiões codificadoras se apresentam é mantida: o primeiro gene é o

que codifica a replicase (ORF1ab), seguido dos genes que codificam as proteínas *spike* (S), envelope (E), membrana (M) e núcleocapsídeo (N). O sentido da tradução bem como a identidade nucleotídica das sequências também é conservado em relação aos outros CoVs. Com um tamanho de cerca de 30 kilobases, o genoma do SARS-CoV-2 possui uma taxa de mutação de 2 nucleotídeos ao mês[25]. As principais regiões codificadoras identificadas até o

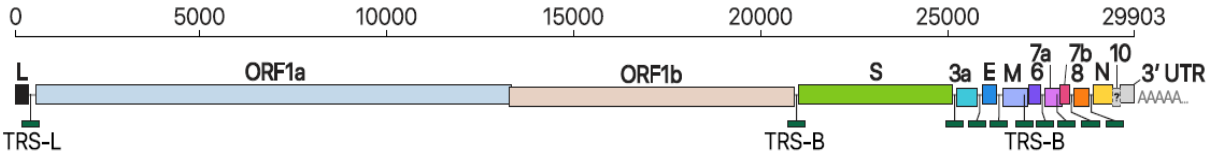

Figura 2. Representação esquemática do genoma do SARS-CoV-2. As principais regiões codificadoras: ORF1ab, S, 3a, E, M, 6, 7a, 7b, 8, N, 10 além das sequências reguladoras de transcrição (TRSs) estão representadas por diferentes cores na ordem em que aparecem no genoma. (adaptado de Kim et al, 2020)[26].

Destas, a sequência mais extensa é a que codifica a poliproteína ORF1ab composta de 16 proteínas não-estruturais em sequência (nsp1 a nsp16) que compõem a replicase dependente de RNA (RdRP), ocupando mais de 72% do genoma. Uma análise transcriptômica mostrou que a região mais abundantemente expressa é a da proteína do nucleocapsídeo (N), seguida das regiões S, 7a, 3, 8, M, E, 6, e 7b25,26. Em comparação com a linhagem mais proximamente relacionada (RmYN02), sequências de regiões como a ORF1ab, 3a, E, 6, 7a, N e 10 chegam a apresentar mais de 96% de similaridade. Como esperado, esta taxa cai significativamente ao comparar as sequências da região S entre as duas linhagens virais (71,8%), especialmente se o trecho comparado for a sequência que codifica o RBD (62,4%)[21] (Tabela 1).

Tabela 1. Regiões codificadoras do SARS-CoV-2. Lista dos genes do SARS-CoV-2 incluindo sua posição no genoma e suas funções (adaptado de Bai et al, 2021)[27]

Gene / ORF		**Posição (nt)**		**Função**
		Início	**Fim**	
1a	nsp1	266	805	Acelera a degradação do mRNA, bloqueia respostas imunes inatas.
	nsp2	806	2719	Sofre pressão seletiva positiva.
	nsp3	2720	8554	Liga às proteínas do hospedeiro, media a replicação viral.

	nsp4	8555	10054	Media a interação com membranas celulares.
	nsp5	10055	10972	A principal protease com atividade de processamento da poliproteína dos CoVs.
	nsp6	10973	11842	Restringe a expressão do autofagossomo.
	nsp7	11843	12091	Forma complexo com nsp8, estabilizando-a.
	nsp8	12092	12685	Forma complexo com nsp7, catalisa a síntese de iniciadores de RNA.
	nsp9	12686	13024	Proteína de ligação à ácidos nucleicos.
	nsp10	13025	13441	Interage com nsp14 e nsp16, regula a função da replicase viral.
	nsp11	13442	13480	Função desconhecida.
1b	nsp12	13442	16236	Catalisa a síntese do RNA viral.
	nsp13	16237	18039	Interage com nsp8 e nsp12, regula a atividade da helicase.
	nsp14	18040	19620	Exonuclease. Diminui a incidência de nucleotídeos mal pareados.
	nsp15	19621	20658	Atividade de RNAse. Degrada RNA viral.
	nsp16	20659	21552	Atividade de 2′-O-Mtase. Escapa da imunidade inata.
S		21563	25384	Liga-se aos receptores celulares do hospedeiro, media a ligação e fusão do envelope e membrana celular do hospedeiro.
3a		25393	26220	Induz apoptose, patogenicidade e liberação dos vírions, media a ativação do inflamassomo.
3b		25814	25882	Está relacionado à ativação de AP-1 pelas vias ERK e JNK.
E		26245	26472	Interage com a proteína de membrana para constituir o envelope.
M		26523	27191	Determina a forma do vírion. É central na organização da constituição da estrutura dos CoVs.
6		27202	27387	Inibe a tradução das proteínas celulares.
7a		27394	27759	Relacionado à interação vírus-hospedeiro.
7b		27756	27887	Relacionado à interação vírus-hospedeiro.

8b	27894	28259	Media a supressão imunológica e evasão viral.
N	28274	29533	Liga-se ao DNA genômico viral, forma o nucleocapsídeo.
9b	28284	28942	Media respostas imunes.
9c	28733	29577	Modifica a atividade mitocondrial do hospedeiro.
10	29558	29674	Modifica a atividade mitocondrial do hospedeiro.

1.3.4. CICLO REPLICATIVO

A penetração do SARS-CoV-2 nas células do epitélio nasal e bronquial se dá principalmente pelo contato do domínio de ligação ao receptor (RBD) localizado na subunidade 1 do homotrímero spike (S1) com o domínio extracelular do receptor de membrana da enzima de conversão da Angiotensina-2 (ACE-2)[28]. À interação entre a região RDB e ACE-2, segue-se a clivagem de um sítio específico da subunidade 2 do spike (S2) por proteases celulares, principalmente a Serinoprotease Transmembrana-2 Humana (TMPRSS2), desencadeando o processo de endocitose e entrada da partícula viral na célula com posterior liberação do RNA genômico viral no citoplasma celular[30-32]. Apesar de o receptor (ACE-2) utilizado pelo novo coronavírus ser o mesmo utilizado pelo SARS-CoV, a proteína spike (S) do SARS-CoV-2 possui um sítio adicional de clivagem que é reconhecida pela proteína furina localizada no complexo de golgi celular envolvida na maturação das proteínas de superfície viral[33]. Posteriormente, o vírus migra para os pneumócitos, especialmente para o tecido epitelial alveolar do tipo II, cujas células apresentam os maiores níveis de maior expressão e disponibilidade deste receptor no corpo humano[29] (Fig. 3).

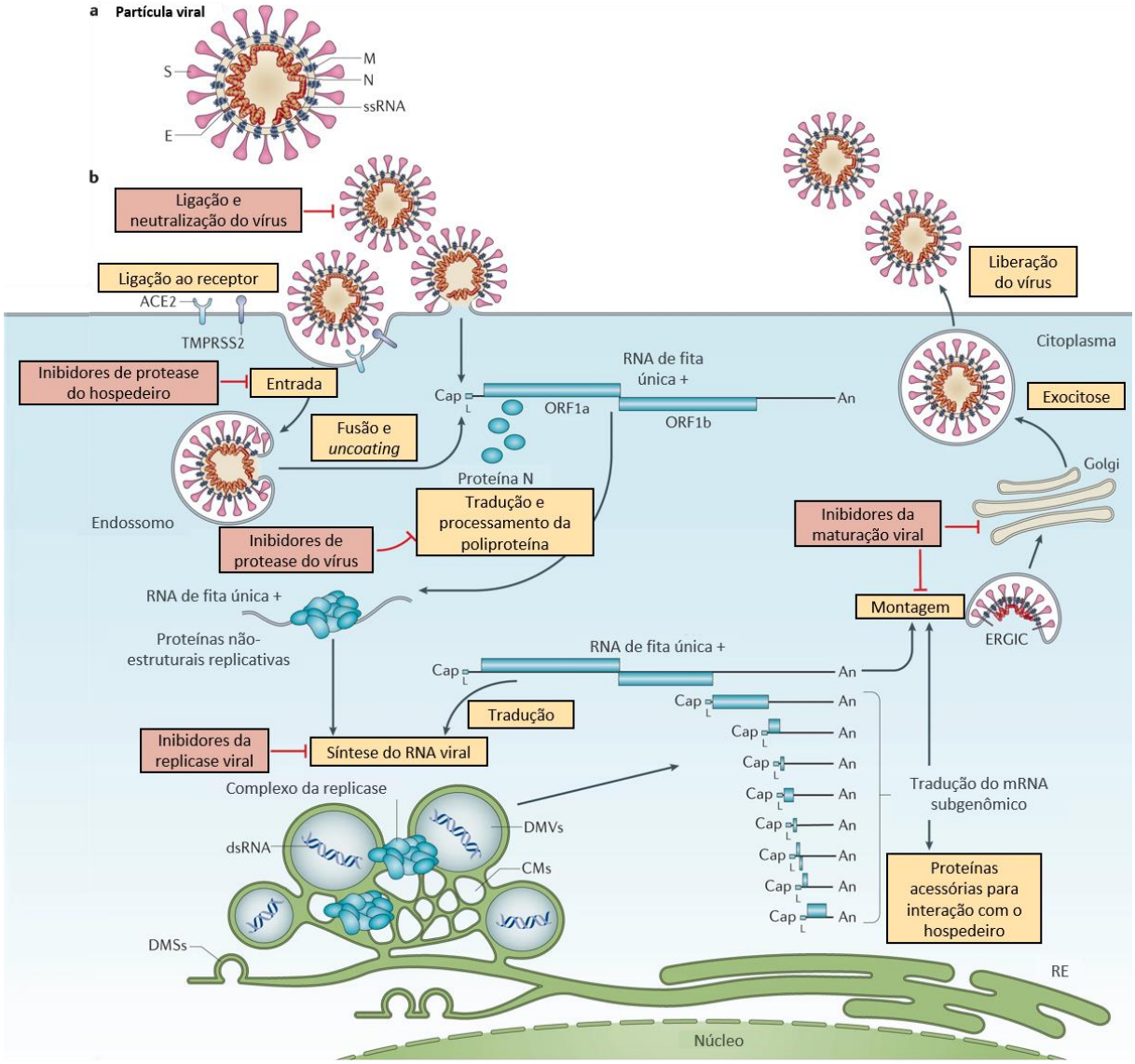

Figura 3. Representação gráfica do ciclo replicativo do SARS-CoV-2. As principais etapas incluindo o reconhecimento dos receptores ACE-2 e TMPRSS2 pela proteína *Spike*, fusão e *uncoating*, tradução e maturação das proteínas virais, exocitose e liberação dos vírions estão representadas. (adaptado de V'kovski et al, 2021)[30].

1.4. COVID-19

1.4.1. A INFECÇÃO POR SARS-CoV-2

COVID-19, a infecção causada por SARS-CoV-2 é transmitida de pessoa a pessoa através das vias áreas superiores. Tendo infectado as células pulmonares, o SARS-CoV-2 induz forte resposta imunológica que pode culminar na inflamação dos alvéolos. A linfopenia em decorrência da infecção dos linfócitos-T, somada à apoptose linfocítica que decorre da resposta inflamatória, acabam por obstruir os alvéolos afetados, o que pode levar a um quadro

de pneumonia. Esta pneumonia pode progredir para um quadro de Síndrome Aguda Respiratória Severa (SARS) a depender de uma série de fatores da infecção, como a proporção de alvéolos obstruídos, o grau de obstrução de cada alvéolo e o sistema imune e fisiologia do hospedeiro[14,29].

Em se tratando de aspectos clínicos, a COVID-19 começa a apresentar sintomas em média cinco dias após o contato com o patógeno. Sintomas mais severos que culminam na procura por atendimento hospitalar são observados no décimo primeiro dia da infecção. O sintoma mais relatado é a febre, observado em uma média de 80% dos pacientes. Outros sintomas comuns em pacientes hospitalizados incluem tosse seca, falta de ar, fatiga, mialgia, náusea, diarreia, dor de cabeça, fraqueza, rinorreia, anosmia e ageusia[29]. Com a progressão do quadro clínico, segue-se a pneumonia que pode desencadear a SARS, requerendo suporte respiratório que pode ser invasiva ou não dependendo da severidade. Frequentemente os casos mais severos são acompanhados por danos no fígado, rins e coração[29]. Danos neurológicos ocasionalmente são observados: uma média de 8% destas complicações são transitórias e cerca de 3%, permanentes. Fatores como sexo, idade e a presença de comorbidades são determinantes no curso da COVID-19. Dados de pacientes hospitalizados de diversas coortes apontam uma média de 60% como sendo do sexo masculino, 80% como possuindo mais de 50 anos e 52,5% como portadores de hipertensão. Outras comorbidades comumente observadas incluem diabetes, doenças crônicas cardiovasculares, pulmonares, hepáticas e renais e doenças oncológicas[29].

1.4.2 A PANDEMIA DA COVID-19

O primeiro caso de pneumonia de agente etiológico desconhecido na China foi reportado localmente em 08 de dezembro de 2019 e globalmente após 22 dias através de nota à OMS na noite de 30 de dezembro de 2019. A nota fazia associação entre o paciente zero e um mercado de frutos do mar e outros produtos tradicionais da cidade de Wuhan, na província de Hubei. A descrição do patógeno se deu em 24 de janeiro de 2020[4,35,36]. Após quase 3 meses do primeiro caso, em 29 de fevereiro de 2020 a província de Hubei contava com 66.337 infectados enquanto as demais províncias chinesas contavam com 13.057 casos. Em 30 de janeiro do mesmo ano, o resto do mundo contava com 82 casos confirmados e nenhuma morte. Um mês depois a pandemia progrediu para 6009 casos reportados distribuídos em 53 países. Em março de 2020, após o vírus ter sido reportado em 197 países a OMS oficializou a situação como uma pandemia[35]. Até o presente momento, a COVID-19 conta com um total de cerca de 258 milhões de casos confirmados. Com taxa de mortalidade de 2%, o sétimo coronavírus foi responsável por uma perda humana muito maior que as

epidemias causadas pelos demais coronavírus, levando a um total de 5,15 milhões de óbitos[37] (Fig. 4).

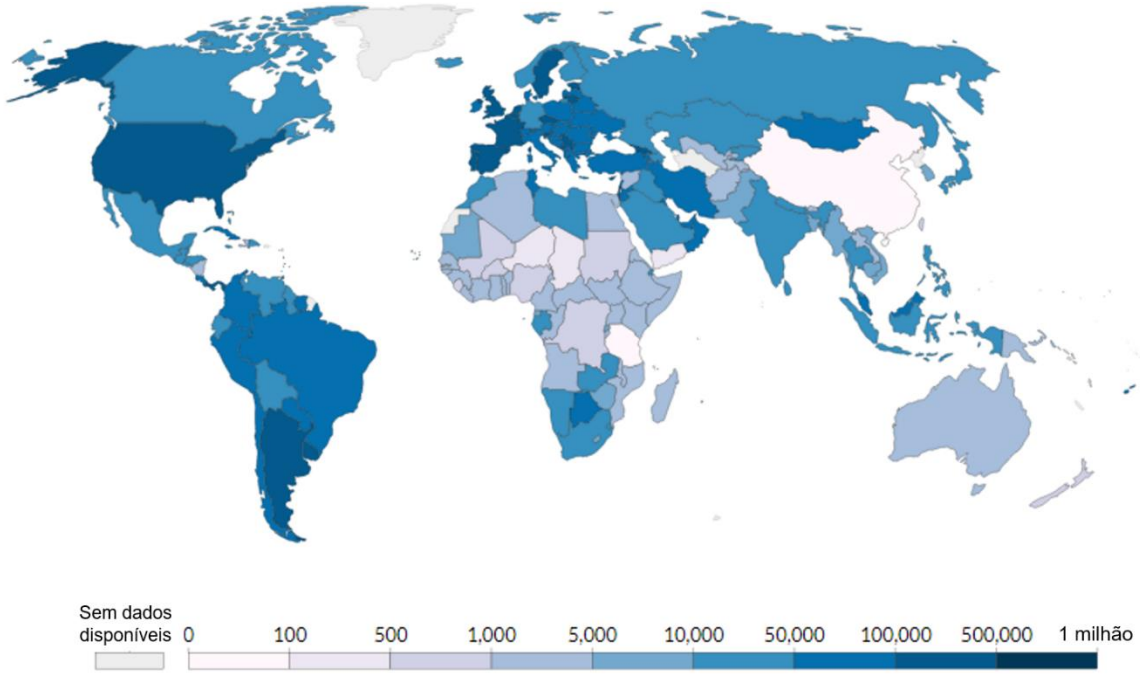

Figura 4. Total de casos confirmados por milhão de habitantes. O mapa inclui dados cumulativos ao redor do mundo até 08 de setembro de 2021. Os casos estão representados em escala de azul com a intensificação da cor associada ao aumento do número de casos. (adaptado de Ritchie et al, 2020)[38].

O primeiro caso de infecção por SARS-CoV-2 no Brasil - também o primeiro da América do Sul - foi reportado em 26 de fevereiro de 2020 na cidade de São Paulo (SP). O indivíduo do sexo masculino com 61 anos de idade tinha histórico de viagem à Itália, na região da Lombardia, epicentro epidemiológico do país na época. Na data, o mundo contava 81.109 casos distribuídos em 38 países. O nosso grupo de pesquisa participou do sequenciamento genomico deste primeiro caso no Brasil e a comparação com genomas de referência de Wuhan e da Itália revelaram duas mutações adicionais: uma na sequência da proteína não estrutural 2 (nsp2) e outra no gene que codifica a proteína *spike* (S)[39,40]. A rápida progressão da COVID-19 ao redor do mundo obrigou a adoção de medidas de intervenção não-farmacológicas (NPIs), como fechamento de escolas e outros serviços não-essenciais além da interrupção de voos internacionais para tentar mitigar o impacto da COVID-19. Mesmo tendo demonstrado ser eficazes no retardamento da transmissão do SARS-CoV-2, tais medidas não foram adequadamente observadas por muito tempo no Brasil. Fatores econômicos, sociais e políticos desnudados com o avanço da pandemia, o alto grau de subnotificação e a adesão de *Fake News* por parte significativa da população podem explicar tais inadequações[41]. A falta de assertividade na implementação destas medidas fez com que

a pandemia de COVID-19 se prolongasse em muitos países até o presente momento. No caso do Brasil, observam-se dois picos no que diz respeito à casos e mortes: o primeiro, no fim de julho de 2020 seguido de queda até novembro do mesmo ano, e o segundo no fim de março de 2021 (Fig 5).

Figura 5. Curva da pandemia da COVID-19 no Brasil. O gráfico mostra os novos casos de infecção por SARS-CoV-2 reportados desde o início das introduções em território nacional até 05/12/2021.

Apesar das estatísticas mostrarem que ainda estamos longe de chegarmos a um momento confortável, a situação no país começa a melhorar com o avanço da vacinação, com as curvas de infecções e principalmente óbitos demonstrando considerável queda. Até o presente momento, o SARS-CoV-2 infectou um total de 22 milhões de brasileiros, levando a um total de mais de 613 mil mortes. Sendo o sexto país em termos populacionais, o Brasil é responsável por 8,53% dos casos globais e 11,9% das mortes[37].

Minas Gerais é o segundo estado mais populoso do Brasil sendo fortemente conectado aos estados de São Paulo e Rio de Janeiro com os quais faz fronteira, ambos polos de importante atividade econômica do país. A soma destes fatores fazem com que o estado seja vulnerável à COVID-19, do ponto de vista epidemiológico. O primeiro caso no estado foi reportado em 4 de março de 2020, uma semana depois do primeiro caso do país em São Paulo. O indivíduo, do sexo feminino, com 38 anos de idade tinha acabado de desembarcar de uma viagem à Israel. O quadro da pandemia em Minas Gerais seguiu a tendência nacional, com um pico ligeiramente mais achatado em julho de 2020 e um segundo pico muito mais acentuado no fim de março de 2021. Assim como no país, o avanço da campanha de vacinação tem contribuído para a tendência de queda dos números de transmissão e óbito no

estado de Minas Gerais. Até o presente momento, 2,2 milhões de casos e mais de 56 mil óbitos foram reportados no território estadual[42,43].

1.4.3. DIVERSIFICAÇÃO E EVOLUÇÃO DO SARS-CoV-2

A classificação de organismos vivos e outros entes replicativos em grupos com base na similaridade é um problema clássico das ciências naturais. No caso dos vírus a questão da taxonomia é ainda mais complexa, tendo em vista a alta taxa de mutação e diversidade viral. De acordo com o Comitê Internacional de Taxonomia de Vírus (ICTV), até o momento não há consenso em relação aos critérios de classificação de diversidade viral abaixo do nível de espécies[44]. Tipicamente, os clados virais são nomeados com uma grande variedade de termos como 'subtipos', 'genótipos', 'grupos' e 'tipos virais'. Atualmente, mesmo a nomenclatura de espécie para vírus tem sido questionada e novas propostas, como o conceito de quasispecies tem ganhado cada vez mais força[45].

No caso do SARS-CoV-2, duas linhagens (A e B) do vírus que remontam à Wuhan rapidamente se espalharam por todo o território chinês. Os genomas pertencentes à linhagem A conservam 2 nucleotídeos (posição 8782 na nsp4 da ORF1a e a posição 28144 na ORF8) em relação às linhagens de CoV mais próximas que infectam morcegos (RaTG13 e RmYN02). Apesar dos genomas de vírus pertencentes ao clado B terem sido sequenciados antes dos genomas da linhagem A, estes não apresentam os nucleotídeos em ambas as posições, o que sugere que os genomas da linhagem A componham o ramo mais antigo do SARS-CoV-2[46]. Desde então, Rambaut e colaboradores, 2020 propuseram um conjunto de regras para a nomenclatura das diferentes linhagens do SARS-CoV-2, que tem sido observadas até o presente momento: *"em primeiro lugar, um suposto clado emergente deve exibir evidência filogenética de ser descendente de um clado ancestral e deve apresentar evidências de ganho evolutivo em termos de aumento da frequência em uma região geograficamente distinta da região de origem do clado ancestral*".

Posteriormente, ainda segundo os autores, para o registro da nova linhagem devem ser obedecidos os seguintes critérios: *"(a) os novos genomas devem exibir um ou mais nucleotídeos de diferença dos ramos ancestrais; (b) novos clados devem possuir 5 genomas independentes com cobertura genômica > 95%; (c) os genomas dentro de um novo clado devem exibir pelo menos uma troca nucleotídica conservada; e (d) um valor de bootstrap > 70% para o nó definidor da linhagem"* (Fig. 6)[46].

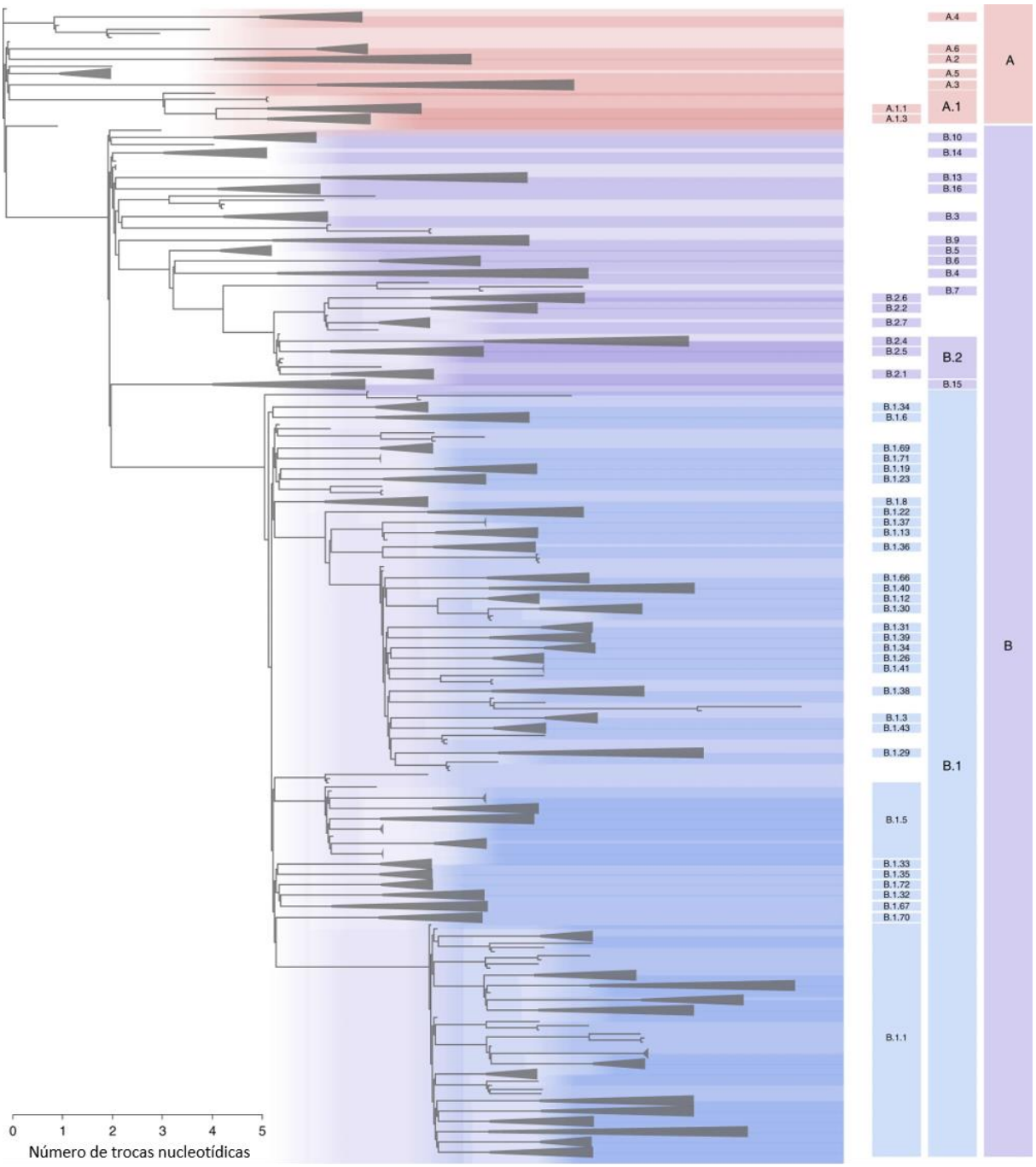

Figura 6. Reconstrução filogenética das primeiras linhagens de SARS-CoV-2. A árvore inclui genomas até 18 de maio de 2020. Cinco genomas representativos estão incluídos para cada linhagem definida. As linhagens mais representadas estão destacadas por áreas coloridas e numeradas à direita (adaptado de Rambaut et al, 2020)[46].

A nível global, os estudos de vigilância genômica conduzidos nos primeiros meses mostravam que a transmissão das linhagens descendentes do clado B predominavam em relação as linhagens derivadas de A, especialmente na Europa, fato que está associado com a incorporação de mutações que se traduzem em um ganho na transmissibilidade. Na Itália, primeiro epicentro da pandemia fora da China, a linhagem B.1 e sua derivada B.1.1 já haviam

sido reportadas como dominantes em um estudo de fevereiro de 2020[35,47]. A linhagem B.1, assim como as linhagens descendentes deste clado, é marcada pelo surgimento da mutação no gene S:D614G, uma mutação não-sinônima na posição de aminoácido 614 que resulta na substituição de um ácido aspártico por uma glicina. Desde o seu aparecimento, estudos de vigilância genômica apontaram um aumento significativo da frequência de 614G em relação aos genomas que apresentavam 614D no gene que codifica o *Spike*. A mutação S:D614G permitiu a diversificação dos clados descendentes de B.1 em diferentes regiões geográficas e se reflete na dominância absoluta destes clados em todo o globo[47].

O primeiro trabalho que buscou acessar o cenário inicial da COVID-19 no Brasil por vigilância genômica foi um estudo que sequenciou amostras dos 6 primeiros indivíduos atestados como positivos no país, por meio de RT-qPCR. Os pacientes eram indivíduos que estavam voltando de viagem à Europa ou que tiveram contato próximo com viajantes que desembarcaram do continente. Os genomas obtidos não mostraram diferenças em comparação com as linhagens que predominavam na Itália no momento do estudo[49]. Nos próximos meses, estudos de vigilância genômica de diferentes regiões do Brasil começaram a emergir na tentativa de estabelecer a frequência das diferentes linhagens circulando no país.

Estudos que sequenciaram amostras coletadas até o fim de abril de 2020 de diferentes regiões do país mostraram a dominância das linhagens derivadas do clado B em comparação com o clado A: Um estudo publicado em julho de 2020 comparou 490 sequências brasileiras de SARS-CoV-2 isoladas entre 5 de março e 30 de abril de 2020. A comparação com genomas de referência de Wuhan mostrou que a maior parte (485) das sequências isoladas no Brasil pertenciam à linhagem B do vírus, sendo apenas 5 pertencentes à linhagem A. A análise filogenética das sequências permitiu a observação de 3 clados até então: Clado 1, concentrado em São Paulo; Clado 2, espalhado em 16 estados, evidenciando a transmissão local e Clado 3, concentrado no Ceará, composto majoritariamente de sequências importadas de outros países. O estudo também avaliou o impacto das diferentes intervenções não-farmacológicas nos primeiros meses da pandemia no Brasil. Dados de março de 2020 mostraram que as NPIs levaram a uma queda na taxa de transmissibilidade (R), definida como a média de infecções comunitárias promovidas por indivíduos infectados. Em São Paulo e Rio de Janeiro esta taxa brevemente passou de $R>3$ a $R<1$, indicando um decréscimo na transmissibilidade logo após a adoção destas medidas[50].

Publicado também em julho de 2020, um estudo simultâneo compilou dados de 40 sequências de SARS-CoV-2 isoladas em Minas Gerais. Os resultados demonstraram a mesma tendência dos estudos epidemiológicos no restante do país. O estudo revelou a conservação de alguns padrões nacionais no estado, como o alto grau de subnotificação e

particularidades locais como o fato dos números de casos e óbitos estarem geograficamente mais distribuídos, ao contrário dos outros estados onde os indicadores sobressaíam-se nas capitais. Com relação às linhagens encontradas, 37 dos 40 genomas (92,5%) agrupavam-se no clado B.1., 2 (5%) no clado B e apenas 1 (2,5%) no clado A. A análise filogenética revelou ainda que o início da pandemia no estado havia sido marcado por múltiplas introduções e que este tipo de entrada predominava em relação à transmissão local no momento do estudo[51]. Na mesma época, um estudo que sequenciou amostras coletadas em Rondônia em março de 2020 encontrou apenas linhagens derivadas de B compondo o cenário local na proporção 75% B.1.1 / 25% B.1[52].

Com o avanço rápido da pandemia no Brasil, diferentes linhagens derivadas do clado B começaram a emergir e aumentar de frequência no território nacional, evoluindo na direção de uma maior transmissibilidade. Duas cepas particularmente bem sucedidas na colonização do ambiente brasileiro foram as linhagens B.1.1.28 e B.1.1.33. Visando acessar a prevalência da linhagem B.1.1.33 em diferentes regiões do país, um estudo coletou amostras do Rio de Janeiro, Pernambuco e Pará até 30 de abril de 2020. Cerca de 33% das sequências agruparam-se neste clado mediante modelos de máxima-verossimilhança[53]. O aumento da circulação destas duas linhagens foi observado também em dois estudos independentes que sequenciaram amostras coletadas no estado de Pernambuco e Sergipe[54,55]. Caminhando para o final do ano, um estudo que sequenciou amostras do Rio de Janeiro coletadas entre abril e novembro de 2020 mostrou a prevalência (94,44%) dos clados B.1.1.28 e B.1.1.33 e linhagens derivadas[56].

No fim do ano de 2020 com o maior afrouxamento das medidas sanitárias, o país vivenciava seu pior momento na pandemia com o crescimento de uma segunda onda cujo ápice seria atingido em março de 2021. Dentre outras coisas, este momento se destacou pela emergência de variantes de interesse e de preocupação nacionais além da introdução e estabelecimento de variantes internacionais no país. Variantes de interesse (VOI) são linhagens que possuem mutações estruturais com potencial de alterarem a dinâmica da infecção, modificando atributos como as taxas de transmissibilidade, mortalidade e resistência às vacinas disponíveis. Variantes de preocupação (VOC) são linhagens que possuem mutações estruturais que estão associadas com a alteração de um ou mais destes atributos, com evidências suportadas pela literatura. Conforme esperado, a pressão seletiva da passagem de um hospedeiro para o outro ocorre principalmente na proteína *Spike*, o que faz com que a sequência do gene que codifica esta proteína apresente a maior variabilidade em variantes bem sucedidas na colonização de novos ambientes. Deste modo, em geral os polimorfismos no gene S tem sido majoritariamente considerados para classificar linhagens

como VOC e VOI. Esta classificação tem sido importante para guiar os estudos de vigilância genômica do SARS-CoV-2 ao redor do globo[57] (Tabela 2).

Tabela 2. Variantes de interesse e preocupação do SARS-CoV-2. Lista de todas as variantes classificadas como VOC ou VOI pela OMS até o presente momento incluindo os sítios de mutações nos aminoácidos e país de origem. (adaptado de Centers for Disease Control and Prevention, 2021)[58].

Kappa (B.1.617.1)	T95I, G142D, E154K, L452R, E484Q, D614G, P681R, Q1071H	VOI	Índia
B.1.617.3	T19R, G142D, L452R, E484Q, D614G, P681R, D950N	VOI	Índia
Alpha (B.1.1.7 e linhagens derivadas)	69del, 70del, 144del, E484K, S494P, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H K1191N	VOC	Reino Unido
Beta (B.1.351, B.1.351.2 e B.1.351.3)	D80A, D215G, 241del, 242del, 243del, K417N, E484K, N501Y, D614G, A701V	VOC	África do Sul
Delta (B.1.617.2 e linhagens derivadas)	T19R, V70F, T95I, G142D, E156-, F157-, R158G, A222V, W258L, K417N, L452R, T478K, D614G, P681R, D950N	VOC	Índia
Gamma (P.1 e linhagens derivadas)	L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I	VOC	Brasil/Japão

Dentre as variantes de emergência nacional destacam-se a P.1 e a P.2, por sua prevalência no cenário atual. Ambas foram descritas em março de 2021 e descendem da linhagem B.1.1.28. A primeira, descrita em Manaus, espalhou-se de maneira muito bem sucedida por todo o país. A segunda, descrita no Rio de Janeiro, concentrou-se principalmente nas regiões sudeste e sul do país. Nota-se ainda a emergência de outras VOIs, como a P.4 e a N.9, mas a dinâmica epidemiológica e o espaço destas linhagens no que tange à prevalência no panorama nacional são restritos [58,59-64]. Além das linhagens nacionais, a introdução e/ou transmissão local de linhagens como a B.1.1.7, B.1.525, B.1.1.251 além da VOC indiana Delta (B.1.617.2), mais transmissível e virulenta tem sido reportada de maneira recorrente[65-68]. O monitoramento destas linhagens por meio de estudos de vigilância genômica é de suma importância para o entendimento e controle da COVID-19 no país.

1.5. VIGILÂNCIA GENÔMICA

Em meados de 2007, Sintchenko e colaboradores publicaram um artigo propondo a criação e compartilhamento de perfis genômicos de patógenos de interesse médico expondo a necessidade de preencher lacunas do conhecimento em epidemiologia da época. Em suas palavras, a *"combinação de marcadores genômicos ou de outras ciências ômicas de modo a disponibilizar dados para serem integrados e compartilhados é essencial para uma vigilância bem sucedida e controle de doenças*"[69]. Foi somente na década seguinte com a explosão de geração e publicação de dados de sequenciamento, entretanto, que o termo "vigilância genômica" passou a ser visto nos trabalhos relacionados a monitoramento de patógenos sem, porém, uma preocupação em definir a ferramenta. Chan e Rabadan, em 2013 propuseram o coeficiente q^2 como uma medida da proporção da distância genética 'R' do ancestral mais proximamente relacionado para a análise de sequências de influenza. Os autores declararam que "*sistemas de vigilância genômica servem como ferramentas de valor incalculável para detecção antecipada de surtos epidêmicos, determinação da variação genética de uma população, melhoramento do design de vacinas e superação do problema da resistência à antibióticos"*[70].

A última década foi marcada pelo acúmulo de dados genômicos e desenvolvimento de ferramentas para lidar com esses dados, além de práticas coordenadas que vão desde a predição de mecanismos biológicos com base em sequências nucleotídicas até a alimentação de bancos de dados genômicos e a execução de pesquisas robustas tendo como base estes bancos a fim de obter respostas mais precisas para questões biológicas recorrentes. No contexto da epidemiologia, a vigilância genômica tem emergido como uma ferramenta que integra dados genômicos com dados clínicos, de mobilidade e transmissibilidade para modelar padrões de evolução e dispersão geográfica de agentes etiológicos de interesse.

Os anos de 2020 e 2021 tem sido extremamente prolíficos em se tratando de trabalhos de vigilância genômica como seu tema principal ou secundário até o momento. A urgência em obter respostas concretas que pudessem guiar as medidas durante o curso da pandemia de COVID-19 impulsionou a publicação de um número nunca antes alcançado deste tipo de trabalho além da disponibilização de bancos de dados de sequenciamento, ferramentas de análise e outros recursos voltados para a vigilância genômica do SARS-CoV-2[71,72].

A pandemia de COVID-19 revelou o papel da mobilidade humana na transmissão de um novo vírus em escala global. Uma epidemia de alta transmissibilidade foi capaz de impactar o mundo em pouquíssimo tempo, principalmente devido à ausência de um sistema eficaz de monitoramento de patógenos infecciosos a nível global. Epidemias não conhecem

fronteiras, ao contrário, são situações dinâmicas que exigem de nós o investimento em estratégias de controle e mitigação de impactos que saibam lidar com esse dinamismo. A vigilância genômica do SARS-CoV-2 foi importante desde o início da pandemia por identificar regiões conservadas no genoma do vírus para o desenho de iniciadores para diagnóstico, sequências imunogênicas a serem utilizadas em projetos de vacina, ancestrais comuns e reservatórios silvestres do vírus e acompanhar a evolução das linhagens virais através da identificação de novas mutações e padrões de dispersão da epidemia.

A vigilância genômica do SARS-CoV-2 se faz especialmente necessária em um país vasto, populoso e diverso como Brasil onde as características de cada localidade exigem análises particulares para a adoção de medidas de controle específicas. O município de Betim, alvo deste estudo, faz importante conexão geográfica, sendo atravessado por 2 vias intermunicipais: MG-060, MG-050 e 3 interestaduais: BR-381, BR-040, BR-262. Betim é a quinta cidade mais populosa do estado de Minas Gerais, reunindo 444.784 habitantes, segundo estimativas recentes[73-75]. Sendo parte da região metropolitana de Belo Horizonte, a cidade hospeda trabalhadores da capital do estado que compõem considerável fatia da população. Estas características fazem com que a cidade seja bastante propensa à introdução externa do SARS-CoV-2. O primeiro caso de infecção por SARS-CoV-2 no município foi em 21 de março de 2020 e até o presente momento, o município conta com 32602 casos e 1373 óbitos notificados[43].

Um estudo de vigilância genômica no município será de extrema importância para avaliar a importância das fronteiras estaduais na transmissão do vírus, além do surgimento e propagação de novas variantes nos estados brasileiros. Os resultados de um estudo como este podem guiar autoridades locais na adoção de medidas de controle da dispersão do SARS-CoV-2 aumentando significativamente a eficácia deste gerenciamento, o que traria benefícios a todos os cidadãos.

2. OBJETIVOS

2.1. Geral

Avaliar a introdução e circulação de linhagens de SARS-CoV-2 na cidade de Betim-MG durante a primeira fase da pandemia através de estudo de vigilância genômica viral.

2.2. Específicos

a) Obter genomas completos do SARS-CoV-2 a partir de indivíduos positivos para COVID-19 em Betim, MG
b) Identificar as linhagens de SARS-CoV-2 circulantes em Betim através de modelos filogenéticos.
c) Descrever e avaliar os polimorfismos encontrados nas sequências obtidas.
d) Avaliar a dispersão geográfica do SARS-CoV-2 na cidade.
e) Estimar as datas e origens de introdução do SARS-CoV-2 na cidade de Betim através de modelos filogeográficos datados.

3. METODOLOGIA

3.1. OBTENÇÃO DAS AMOSTRAS

Visando avaliar a soro-prevalência do SARS-CoV-2 em Betim, um estudo anterior coletou material biológico por meio de swab naso-faríngeo de 3.240 indivíduos domiciliados em Betim-MG em 3 rodadas entre junho e julho de 2020: A primeira rodada de coleta ocorreu nos dias 3 a 5 de junho de 2020, a segunda rodada nos dias 23 a 25 de junho de 2020 e a

terceira rodada nos dias 13 a 15 de julho de 2020. O estudo foi realizado de forma randomizada com busca ativa domiciliar promovido pela prefeitura de Betim em parceria com o nosso grupo de pesquisa. O estudo coletou ainda dados geoespaciais e epidemiológicos de todos os participantes. Esse estudo foi aprovado pelo comitê de ética (CAAE 31459220.2.0000.5651) com assinatura do termo de consentimento livre e esclarecido (TCLE) de todos os participantes.

3.2. DIAGNÓSTICO MOLECULAR DE COVID-19

O procedimento de extração de RNA das amostras coletadas se deu a partir do protocolo especificado pelo kit de extração utilizado, BioGene Extração de DNA/RNA Viral - Bioclin. O protocolo consiste em um fluxo de trabalho de 5 passos: 1. lise celular; 2. precipitação; 3. ligação; 4. lavagem e 5. Eluição. Conforme descrito pelo fabricante, na etapa da lise foram adicionados 5µL de Proteinase K e posteriormente 200 µL de tampão de lise (reagente incluso) à 200µL de cada amostra em um tubo de 1,5mL. Em ambos os momentos, as soluções foram devidamente homogeneizadas. Durante a etapa de precipitação, 5,6µL de RNA carreador foram adicionados às soluções, que foram homogeneizadas e então incubadas por 3 minutos a temperatura ambiente (18-25ºC). Após a incubação, 200µL de Etanol absoluto (96-100%) foram adicionados às soluções, que foram novamente homogeneizadas e então incubadas por 5 minutos a temperatura ambiente (18-25ºC). Na etapa de ligação, o conteúdo total das soluções foi centrifugado em coluna contendo tubo coletor (inclusos) por 3 minutos a 4000 x g, descartando por fim o tubo coletor com o filtrado. A etapa de lavagem consistiu em 3 centrifugações de reagentes de lavagem, seguidas do descarte dos filtrados para cada amostra: uma adicionando 400µL do reagente de lavagem 1 (incluso) por 1 minuto a 11000 x g, uma adicionando 400µL do reagente de lavagem 2 (incluso) por 1 minuto a 11000 x g e uma adicionando 200µL do reagente de lavagem 2 por 1 minuto a 15000 x g. Por fim, durante a etapa de eluição as colunas foram transferidas para tubos coletores finais de 1,5mL. Os tubos contendo as colunas foram incubados por 5 minutos a 56ºC com as tampas abertas. Posteriormente, 50µL de água livre de nuclease preaquecida a 70ºC foram adicionados diretamente no centro da membrana de cada coluna. As colunas e tubos foram então incubadas por 3 minutos a temperatura ambiente (18-25ºC), e por fim centrifugadas por 3 minutos a 15000 x g. Cada lote de extração incluiu 1 controle negativo contendo água livre de Nuclease no lugar da amostra.

As amostras de RNA extraído foram testadas para a presença de 2 alvos do gene do núcleocapsídeo (N1 e N2) do SARS-CoV-2 além de 1 alvo humano como controle endógeno (RNAseP) por meio de ensaio RT-qPCR. A sequência dos iniciadores e sondas fluorescentes

utilizados no diagnóstico foram sugeridas pelo CDC dos EUA (EUA 200001)[76]. Para o alvo N1, as sequências dos iniciadores direto, reverso e da sonda utilizadas foram, respectivamente: 5'-GACCCCAAAATCAGCGAAAT-3', 5'-TCTGGTTACTGCCAGTTGAATCTG-3', e 5'-FAM-ACCCCGCAT/ZEN/TACGTTTGGTGGACC-3IABkFQ-3'. Para o alvo N2, as sequências dos iniciadores direto, reverso e da sonda utilizadas foram, respectivamente: 5'-TTACAAACATTGGCCGCAAA-3', 5'- GCGCGACATTCCGAAGAA-3' e 5'-FAM-ACAATTTGC/ZEN/CCCCAGCGCTTCAG-3IABkF-3'. Para o alvo RNAseP, as sequências dos iniciadores direto, reverso e da sonda utilizadas foram, respectivamente: 5'-AGATTTGGACCTGCGAGCG-3', 5'- GAGCGGCTGTCTCCACAAGT-3' e 5'- FAM-TTCTGACCT/ZEN/GAAGGCTCTGCGCG-3IABkFQ-3'. Para o diagnóstico via RT-qPCR foi utilizado o kit GoTaq® Probe 1-Step RT-qPCR System - Promega.

As amostras foram diagnosticadas em placas com capacidade de 96 reações. Cada placa incluiu um controle negativo (água livre de nuclease) e um controle positivo (plasmídeos com sequências dos três alvos). O mix da reação de PCR foi preparado de acordo com as instruções do fabricante em um cálculo levando em conta o volume de cada reagente por reação multiplicado pelo número de reações executadas em cada lote, adicionando 5% do volume de cada reagente em função da perda de pipetagem. De acordo com o protocolo, para cada alvo foram adicionados a temperatura ambiente: 10µL do reagente GoTaq® Probe qPCR Master Mix with dUTP, 0,4µL da enzima GoScript™ RT Mix for 1-Step RT-qPCR, 1,6µL de água livre de nuclease, 1µL do iniciador direto a 5nM, 1µL do iniciador reverso a 5nM e 1µL do mix de sondas a 5nM além de 5µL do RNA a ser testado, resultando num volume total de 20µL por reação. As placas de diagnóstico correram em instrumento de PCR em tempo real, de acordo com a seguinte ciclagem: transcrição reversa a 45ºC por 15 minutos (ciclo único), inativação da transcriptase reversa e ativação da DNA polimerase a 95ºC por 2 minutos (ciclo único) e 40 ciclos de desnaturação (95ºC por 15 segundos) e anelamento e extensão (60ºC por 1 minuto). A amplificação dos 3 alvos para cada amostra foi analisada no *software* do instrumento de PCR em tempo real utilizado. Alvos com *Cycle Threshold* (CT) de amplificação ≤ 40 foram considerados detectáveis. Os resultados de cada amostra foram interpretados da seguinte maneira: detecção dos 2 alvos virais e do controle endógeno: amostra positiva para o SARS-CoV-2; detecção de um dos alvos virais e do controle endógeno: amostra indeterminada; detecção apenas do controle endógeno: amostra negativa para o SARS-CoV-2; ausência de detecção do controle endógeno: amostra inválida. Após o diagnóstico, as amostras foram armazenadas em ultra freezer a -80ºC até o momento do uso para melhor preservação.

3.3. AMPLIFICAÇÃO DOS GENOMAS VIRAIS E PREPARAÇÃO DAS BIBLIOTECAS DE DNA

As amostras que foram consideradas positivas para o SARS-CoV-2 foram triadas com base no CT da amplificação de RT-qPCR. Amostras com CT ≤ 30 foram consideradas elegíveis para os procedimentos de amplificação do genoma viral, montagem de bibliotecas e sequenciamento. O kit utilizado foi o QIAseq® SARS-CoV-2 Primer Panel - QIAGEN, uma abordagem de sequenciamento por amplificação via PCR. De acordo com as instruções do fabricante, primeiramente as amostras de RNA juntamente com 1 controle negativo (água livre de nuclease) para cada lote foram adicionadas a uma solução de água livre de nuclease iniciadores aleatórios em uma reação preparada a 4ºC na seguinte proporção: 5µL de RNA, 1µL de iniciadores aleatórios a 0,35µM diluídos previamente com água livre de nuclease na proporção 1:11 e 6µL de água livre de nuclease, num volume final de 12µL. A solução foi incubada em termociclador direto com tampa aquecida a 105ºC de acordo com o seguinte programa: 65ºC por 5 minutos e logo depois a 4ºC por 1 minuto. Em seguida, as amostras foram convertidas em cDNA por meio de transcrição reversa. A reação preparada a 4ºC incluiu: os 12µL da solução anterior de RNA e iniciadores aleatórios, 4µL do reagente *Multimodal RT Buffer* - 5x, 2µL de água livre de nuclease, 1µL de inibidor de RNAse e 1µL da transcriptase reversa EZ, num volume final de 20µL. A reação foi incubada em termociclador direto com tampa aquecida a 105ºC de acordo com o seguinte programa: 42ºC por 50 minutos e em sequência a 70ºC por 15 minutos e então resfriada até chegar a 4ºC.

Os cDNAs gerados foram amplificados por meio de reação de PCR utilizando pares de iniciadores adjacentes e intercalados entre dois *pools* (*pool* 1 e 2) desenhados para cobrir todo o genoma viral, com uma distância de 400nt entre 2 iniciadores adjacentes (Anexo 1), segundo protocolo Artic[77]. Foram preparadas 2 reações de PCR independentes a 4ºC para cada amostra, sendo uma utilizando o *pool* 1 de iniciadores e outra utilizando o *pool* 2 de iniciadores Cada reação incluiu: 2,5µL do cDNA recém-preparado da amostra, 3µL do *pool* de iniciadores 12µM, 12,5µL do reagente *QIAseq 2X HiFi MM*, e 7µL de água livre de nuclease, em uma reação de volume final de 25µL. Todas as reações foram incubadas com tampa aquecida a 105ºC de acordo com o seguinte programa: ativação da DNA polimerase por 98ºC a 2 minutos (ciclo único) e 35 ciclos de desnaturação por 98ºC a 20 segundos e anelamento e extensão por 65ºC a 5 minutos, seguido de um resfriamento a 4ºC.

Posteriormente, o conteúdo dos tubos amplificados por cada *pool* de iniciadores foi transferido para tubos de 1,5mL e então passou por uma purificação de DNA por meio da adição e homogeneização de 25µL de *beads* de ferro, gerando uma solução de *beads* de ferro e material amplificado numa proporção de 1:1 previamente equilibradas à temperatura

ambiente (18-25ºC). A solução de 50µL foi incubada a temperatura ambiente (18-25ºC) por 5 minutos e então disposta em racks magnéticas. Após 3 minutos de incubação nas racks os 50µL de sobrenadante foram descartados e o *pellet* das *beads* com DNA passou por 2 lavagens por meio de adição e descarte 200µL de uma solução de álcool 80% recém-preparada nos tubos de 1,5mL. Após a lavagem, os tubos foram centrifugados para precipitação e descarte do álcool residual e colocados na rack com as tampas abertas para secagem. Após a secagem, o DNA foi eluido por meio da adição e homogeneização de 15µL de tampão de eluição (*Buffer EB*) aos *pellets* nos tubos de 1,5mL. Os tubos foram incubados por 3 minutos a temperatura ambiente (18-25ºC) e recolocados à rack magnética por 2 minutos para formação de *pellets*. Por fim, 14µL do DNA purificado foram coletados e transferidos para tubos finais.

O DNA das amostras amplificadas pelos 2 *pools* de iniciadores independentes foi quantificado por meio do ensaio Qubit dsDNA HS kit - ThermoFisher, através do preparo da reação de quantificação pela adição e homogeneização de 1µL de amostra para 199µL do reagente de quantificação (1:200). A solução foi incubada por 2 minutos à temperatura ambiente na ausência de luz e então quantificada por meio de aparelho de fluorimetria. Após a quantificação, 75ng de cada *pool* 1 e 2 das amostras foram adicionados a um volume variável de água livre de Nuclease em um tubo de 0,2mL, resultando em um volume de 35µL de solução a 150ng. O procedimento visa a normalização entre diferentes amostras e a equidade entre os trechos amplificados por diferentes *pools* na mesma amostra.

As amostras de DNA amplificado passaram por fragmentação enzimática por meio da adição de 5µL do reagente FX Buffer, 10x e de 10µL do *pool* de enzimas de restrição FX Enzyme Mix, em uma reação preparada a 4ºC com volume final de 50µL que foi incubada em termociclador direto com tampa aquecida a 105ºC de acordo com o seguinte programa: 4ºC por 1 minuto, 14ºC por 32 minutos, 65ºC por 30 minutos e por fim resfriada a 4ºC. Uma combinação de index com adaptadores i5 e i7 específica da placa tipo Illumina® QIAseq UDI Y-Adapter Plate A (96) layout (UDI 1–96) foi atribuída a cada amostra. Os 5µL de adaptadores foram ligados aos 50µL da reação anterior por meio da adição de 20µL de tampão da ligase, 10µL da enzima DNA ligase e 15µL da água livre de nuclease, em uma reação preparada a 4ºC com um volume final de 100µL. A reação foi incubada em termociclador direto sem aquecimento da tampa para a reação de ligação a 20ºC por 15 minutos e logo em seguida incubada em termociclador direto com tampa aquecida a 105ºC para inativação da ligase a 65ºC por 20 minutos. As bibliotecas de DNA foram transferidas para tubos de 1,5mL e então passaram por uma purificação por meio da adição e homogeneização de 80µL de *beads* de ferro previamente equilibradas à temperatura ambiente (18-25ºC). A solução de 180µL (1:0,8) foi incubada a temperatura ambiente (18-25ºC) por 5 minutos e então disposta em racks

magnéticas. Após 3 minutos de incubação nas racks os 180µL de sobrenadante foram descartados e o *pellet* das *beads* com DNA passou por 2 lavagens por meio de adição e descarte 200µL de uma solução de álcool 80% recém-preparada nos tubos de 1,5mL. Após a lavagem, os tubos foram centrifugados para precipitação e descarte do álcool residual e colocados na rack com as tampas abertas para secagem. Após a secagem, o DNA foi eluido por meio da adição e homogeneização de 52,5µL de tampão de eluição (*Buffer EB*) aos *pellets* nos tubos de 1,5mL. Os tubos foram incubados por 3 minutos a temperatura ambiente (18-25ºC) e recolocados à rack magnética por 2 minutos para formação de pellets. Por fim, 50µL das bibliotecas de DNA purificado foram coletados e transferidos para outro tubo de 1,5mL para passarem por uma segunda purificação visando a seleção de fragmentos com tamanho desejado (370pb) que se deu pela adição e homogeneização de 50µL de *beads* de ferro previamente equilibradas à temperatura ambiente (18-25ºC) ao volume eluido. A solução de 100µL (1:1) foi incubada a temperatura ambiente (18-25ºC) por 5 minutos e então disposta em racks magnéticas. Após 3 minutos de incubação nas racks os 100µL de sobrenadante foram descartados e o *pellet* das *beads* com DNA passou por 2 lavagens por meio de adição e descarte 200µL de uma solução de álcool 80% recém-preparada nos tubos de 1,5mL. Após a lavagem, os tubos foram centrifugados para precipitação e descarte do álcool residual e colocados na rack com as tampas abertas para secagem. Após a secagem, o DNA foi eluido por meio da adição e homogeneização de 26µL de tampão de eluição (Buffer EB) aos pellets nos tubos de 1,5mL. Os tubos foram incubados por 3 minutos a temperatura ambiente (18-25ºC) e recolocados à rack magnética por 2 minutos para formação de pellets. Por fim, 23,5µL das bibliotecas de DNA purificado foram coletados e armazenados em tubos finais de 1,5mL.

3.4. QUANTIFICAÇÃO E VERIFICAÇÃO DA INTEGRIDADE DAS BIBLIOTECAS DE DNA

O tamanho do fragmento das bibliotecas foi obtido por meio de corrida em equipamento Bioanalyzer 2100 - Agilent, através do kit Agilent DNA 1000 Kit. Tendo em vista a distância entre iniciadores adjacentes, a eficiência da fragmentação e a ligação dos adaptadores às extremidades, o tamanho médio esperado é de 370pb para cada fragmento. As bibliotecas foram quantificadas por reação de PCR através do ensaio QIAseq™ Library Quant Assay Kit - QIAGEN. Uma abordagem baseada na mensuração da fluorescência liberada pelo agente intercalante *SYBR green*. Os iniciadores utilizados neste kit estão desenhados para a hibridização com os indexes (parte comum) dos adaptadores tipo Illumina® utilizados. A sequência dos iniciadores direto e reverso utilizados é, respectivamente, 5'-AATGATACGGCGACCACCGA-3' e 5'-

CAAGCAGAAGACGGCATACGA-3'. De acordo com o que está especificado no kit, as bibliotecas de DNA foram diluídas por meio de tampão de diluição incluso para as concentrações 1:2000 e 1:20000. As 2 diluições para cada biblioteca de DNA foram adicionadas a uma reação que incluiu: 30,6µL de água livre de nuclease, 45µL do reagente *SYBR Green Mastermix*, 3,6µL do mix de iniciadores e 10,8 de biblioteca de DNA, num volume final de 90µL. Os mixes da reação, incluindo o controle negativo, foram distribuídos em placas de 96 poços em tréplicas de 25µL de volume. As placas incluíram ainda 5 tréplicas do controle positivo em diluições seriadas de 10 vezes, para construção de uma curva padrão além de um controle negativo de placa, também em tréplica. As placas correram em um instrumento *Real Time PCR*, de acordo com o seguinte programa: ativação da DNA polimerase a 95ºC por 10 minutos (ciclo único) e 30 ciclos de 95ºC a 15 segundos para desnaturação, 60ºC a 30 segundos para anelamento e 72ºC a 2 minutos para extensão. O resultado da quantificação de cada amostra por curva padrão foi exportado e incluído em tabela do fabricante que juntamente com o tamanho médio dos fragmentos calculado previamente foi utilizado para base de cálculo da molaridade das bibliotecas.

3.5. SEQUENCIAMENTO DO GENOMA VIRAL

As bibliotecas foram preparadas para sequenciamento em cartucho do tipo v3-600 ciclos em um instrumento MiSeq- Illumina® de acordo com o Protocolo A do kit MiSeq System Denature and Dilute Libraries Guide. Seguindo as orientações do protocolo, em um primeiro momento as bibliotecas foram todas normalizadas para 4nM com adição de água livre de nuclease em um volume final de 10µL tendo como base o valor de molaridade calculado anteriormente. Posteriormente foi preparado um *pool* composto por 5µL de cada biblioteca, que foi vigorosamente homogeneizado e armazenado a 4ºC. Para a desnaturação do *pool* de bibliotecas, foi preparada uma solução de 1mL de NAOH a 0,2N a partir da adição e homogeneização de 200µL de NAOH 1N a 800µL de água livre de nuclease. Em seguida, 5µL do *pool* de bibliotecas foram combinados com 5µL da solução de desnaturação recém-preparada em um tubo de 1,5mL. O tubo contendo os 10µL desta solução foi homogeneizado vigorosamente, centrifugado por 280 x g a 1 minuto e incubado a temperatura ambiente por 5 minutos. Posteriormente, 990µL do reagente de diluição HT1 previamente equilibrado à temperatura ambiente (18-25ºC) foram adicionados à solução de 10µL preparada anteriormente, resultando em uma solução de 20pM de bibliotecas de DNA desnaturadas. Para atingir a molaridade óptima para o sequenciamento, 300µL da solução anterior foram diluídos em outros 300µL do reagente de diluição HT1 previamente equilibrado à temperatura ambiente (18-25ºC), resultando em 10pM de bibliotecas em um volume final de 600µL em um

tubo de 1,5mL, que foi armazenado a temperatura ambiente (18-25ºC). O controle Phix da corrida, foi preparado de acordo com as instruções do protocolo: 2µL de controle Phix a 10nM foram adicionados em um tubo a 3µL de uma solução de Tris-Cl a 10 mM com, 0,1% de Tween 20, de pH final de 8,5, gerando uma solução de 5µL de controle Phix a 4nM, que foi combinada com 5µL da solução de desnaturação recém-preparada em um tubo de 1,5mL. O tubo contendo os 10µL desta solução foi homogeneizado vigorosamente, centrifugado por 280 x g a 1 minuto e incubado a temperatura ambiente por 5 minutos. Posteriormente, 990µL do reagente de diluição HT1 previamente equilibrado à temperatura ambiente (18-25ºC) foram adicionados à solução de 10µL preparada anteriormente, resultando em uma solução de 20pM de controle Phix desnaturado. Em seguida, 375µL desta solução foram diluídos em 225µL do reagente de diluição HT1 previamente equilibrado à temperatura ambiente (18-25ºC), resultando em uma solução de 600µL de controle Phix desnaturado a 12,5pM. Por fim, 6µL da solução de controle Phix a 12,5pM recém preparada foram adicionados a 594µL da solução de bibliotecas a 10pM. Os 600µL desta solução final foram inoculados no cartucho v3-600 ciclos, previamente equilibrado à temperatura ambiente (18-25ºC), que correu em instrumento MiSeq - Illumina, gerando arquivos de sequência FASTQ, que foram posteriormente descarregados do instrumento.

3.6. MONTAGEM DOS GENOMAS E DETERMINAÇÃO DAS LINHAGENS

Os arquivos de sequência FASTQ gerados no sequenciamento foram processados por meio de uma sequência de análises bioinformáticas com ferramentas estabelecidas conhecida como *workflow* ou *pipeline*. As sequências geradas passaram pela remoção de adaptadores, iniciadores, bases nucleotídicas com baixa qualidade de leitura (*phred score* < 30) e *reads* curtas (< 50 nt) através do *software* Trimmomatic v0.39[78] a fim de melhorar a qualidade das sequências, dando origem a um relatório de qualidade interpretado pela interface FASTQC. As *reads* foram mapeadas contra o genoma de referência de SARS-CoV-2 (código de acesso: NC_045512.2) através do *software* Bowtie2[79]. Os arquivos BAM gerados serviram de *input* para os softwares SAMtools, BCFtools[80] e BEDtools[81] gerando sequências-consenso. As sequências que atingiram o esperado no que diz respeito aos parâmetros de qualidade do sequenciamento (cobertura > 75%, profundidade > 200x), foram selecionadas para as análises posteriores.

As linhagens dos genomas obtidos foram acessadas *a priori* a partir do *software* online PANGOLIN 2.0 (versão de 2 de fevereiro de 2021), que classifica os genomas por meio da identificação de 'mutações classificadoras', que são mutações presentes em pelo menos 75% das sequências globais classificadas em um dado clado[46]. Para as reconstruções filogenéticas

estes genomas foram incluídos em uma árvore filogenética não-enraizada contendo 3814 sequências obtidas a partir do banco de dados GISAID[82] até 12 de janeiro de 2021, com data de coleta entre o início e o fim do estudo, sendo: todas as sequências nacionais disponibilizadas, além de uma sequência internacional por semana de cada país com sequências disponibilizadas no banco de dados. Os genomas selecionados foram alinhados a partir do *software* MAFFT v7.475[83] com inferência filogenética a partir do modelo de máxima verossimilhança inferida a partir do *software* IQ-Tree 2[84], com base no modelo GTR+F+I+G4[51,85]. O teste de verossimilhança aproximada de Shimoidara-Hasegawa (SH-aLRT) foi utilizado para acessar a confiabilidade estatística dos ramos da árvore[86].

Visando descrever as mutações em comparação com o genoma de referência do SARS-CoV-2 (NC_045512.2) presentes nas sequências obtidas, a posição e a frequência dos polimorfismos encontrados foi computada de maneira independente por linhagem. Os mapas genéticos foram construídos no *software* R v.4.0.5[87] Foi utilizado o pacote GenomeRanges[88] para construir a estrutura do SARS-CoV-2 e referenciar a localização dos polimorfismos e o pacote trackViewer[89] para representar graficamente os polimorfismos.

3.7. CONSTRUÇÃO DOS MODELOS DE DISPERSÃO VIRAL

Posteriormente, a dispersão geográfica do SARS-CoV-2 na cidade por rodada de coleta foi avaliada de modo generalista e por linhagem. Os dados geoespaciais juntamente com os resultados dos testes de RT-qPCR e da análise de linhagens foram interpolados através do pacote gstat[90] no *software* R v.4.0.5.[87] utilizando os seguintes parâmetros: idp = 2, maxdist = 2857. Os valores de escala da interpolação foram definidos da seguinte maneira, por rodada, para os modelos generalistas: 0 para pontos com resultados não-detectáveis e 1 para pontos com resultados detectáveis. De maneira semelhante, os valores da escala de interpolação por rodada, para cada linhagem foram definidos da seguinte maneira: 0 para pontos com resultados não-detectáveis ou para pontos detectáveis com linhagem desconhecida ou para pontos com linhagem diferente da linhagem do modelo e 1 para pontos com linhagem idêntica à linhagem do modelo. Os modelos generalistas e específicos para cada linhagem foram sobrepostos ao mapa das macrorregiões da cidade de Betim, gerando 9 mapas de calor indicativos da dispersão do SARS-CoV-2.

3.8. CONSTRUÇÃO DOS MODELOS FILOGEOGRÁFICOS

Após a confirmação das linhagens, os genomas obtidos foram incluídos em 2 modelos filogeográficos datados separados por linhagem de modo a avaliar a origem da introdução

das cepas circulantes através do *software* BEAST v.1.10.4[91], ambos utilizando: o modelo de substituição nucleotídica HKY+I+G4[92,93], o relógio molecular estrito, o *skygrid* coalescente não-paramétrico da árvore anterior[94] e um modelo filogeográfico simétrico discreto[95]. Uma distribuição normal anterior foi assumida para o relógio molecular, tendo a taxa de mutação estimada para SARS-CoV-2(média = 1.13 x 10^{-3}; desvio padrão = 5.1 x 10^{-4})[50]. Os valores de corte da árvore *skygrid* anterior foram baseados na estimativa de datas de emergência de cada linhagem[50].

A amostragem das sequências-consenso utilizadas nestes modelos foi realizada de maneira equitativa e estratificada de maneira temporal e espacial, incluindo sempre que possível o mesmo número de sequências para cada categoria e garantindo a inclusão de todas as sequências disponíveis para categorias pouco representadas. As 5 categorias discretas utilizadas baseiam-se nas localidades correlacionadas geográfica e economicamente ao município: Betim (as sequências obtidas no estudo), Minas Gerais, São Paulo, Rio de Janeiro, Brasileiras (outros estados) e Internacionais. A data de coleta dos genomas observou o período entre o surgimento das variantes encontradas e o fim do estudo. Todos os genomas de referência incluídos foram obtidos a partir do banco de dados GISAID[82] (Anexo 2). Para as bases de dados de cada linhagem foram utilizadas, respectivamente, três e duas cadeias independentes de 200 milhões de gerações de amostragem a cada 10 mil estados. O *software* Tracer v1.7.1[96] foi utilizado para verificar a interposição e convergência de cadeias (tamanho efetivo da amostra > 200 para todos os parâmetros) que foram então combinados com uma queima de 10% com logcombiner v1.10.4[91]. A credibilidade máxima dos clados da árvore foi determinada através do treeannotator v1.10.4.[91].

4. RESULTADOS

4.1. LINHAGENS DE SARS-CoV-2 IDENTIFICADAS EM BETIM-MG

O diagnóstico molecular do SARS-CoV-2 resultou em 84 amostras detectáveis para os 2 alvos virais dentre os 3240 indivíduos testados durante todo o período de duração do estudo, sendo 2 resultados detectáveis obtidos na primeira rodada de coleta (03-05/06/2020), 22 resultados detectáveis obtidos na segunda rodada de coleta (23-25/06/2020) e 60 resultados detectáveis obtidos na terceira rodada de coleta (13-15/07/2020) do estudo. Um dos testes positivos dentre os 3240 averiguados não foi levado em conta para o sequenciamento e análises posteriores visto que o indivíduo testado era menor de idade, o que fez jus à retirada desta amostra do estudo de vigilância genômica. A distribuição dos resultados detectáveis coincide com o grau de urbanização das diferentes áreas do município (Fig. 7; Fig. 8).

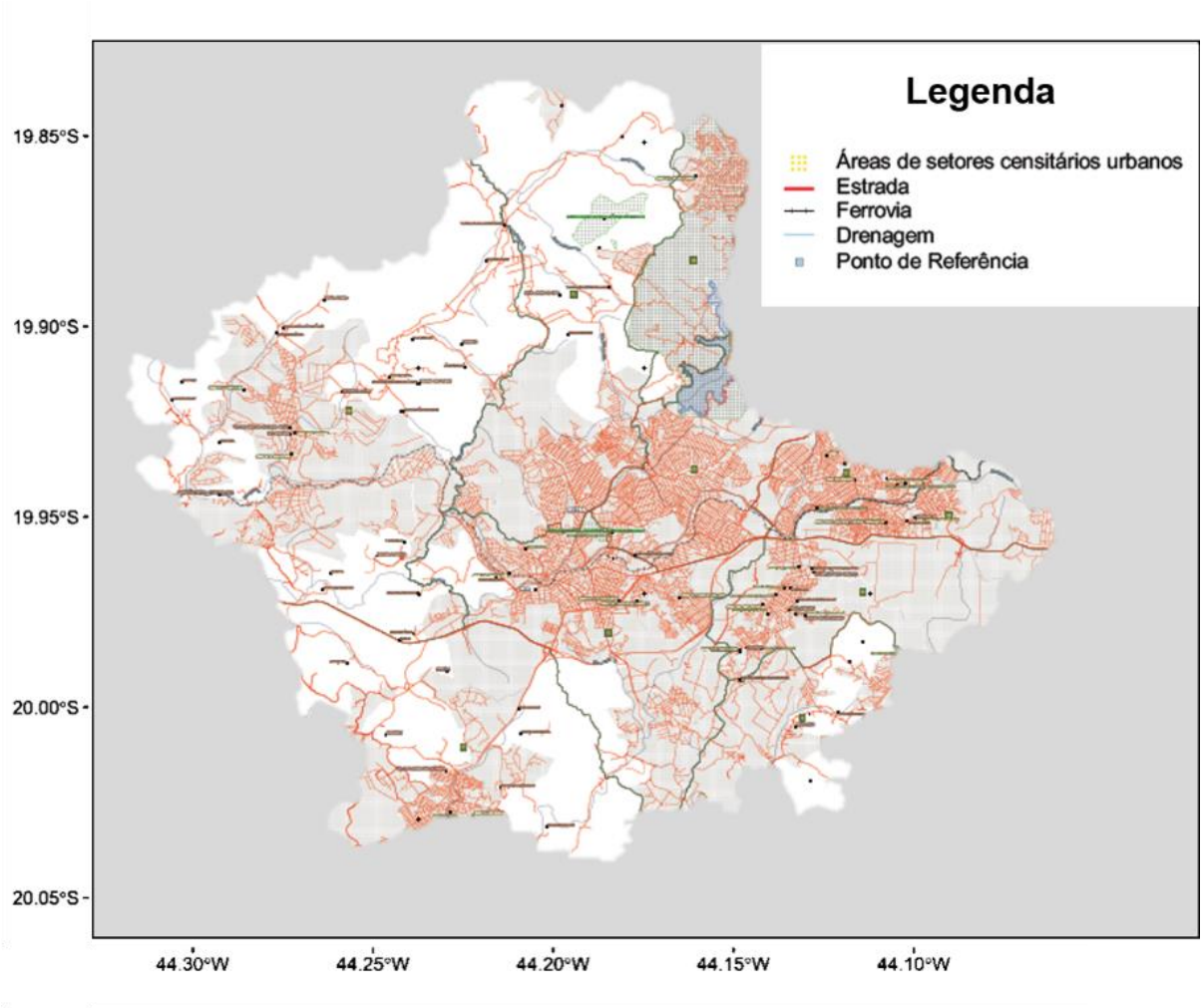

Figura 7. Grau de urbanização das microrregiões do município de Betim-MG. No mapa os elementos geográficos associados à urbanização: áreas de setores censitários urbanos, estradas, ferrovias, drenagens e pontos de referência estão representados por diferentes símbolos (adaptado de IBGE, 2020)[97].

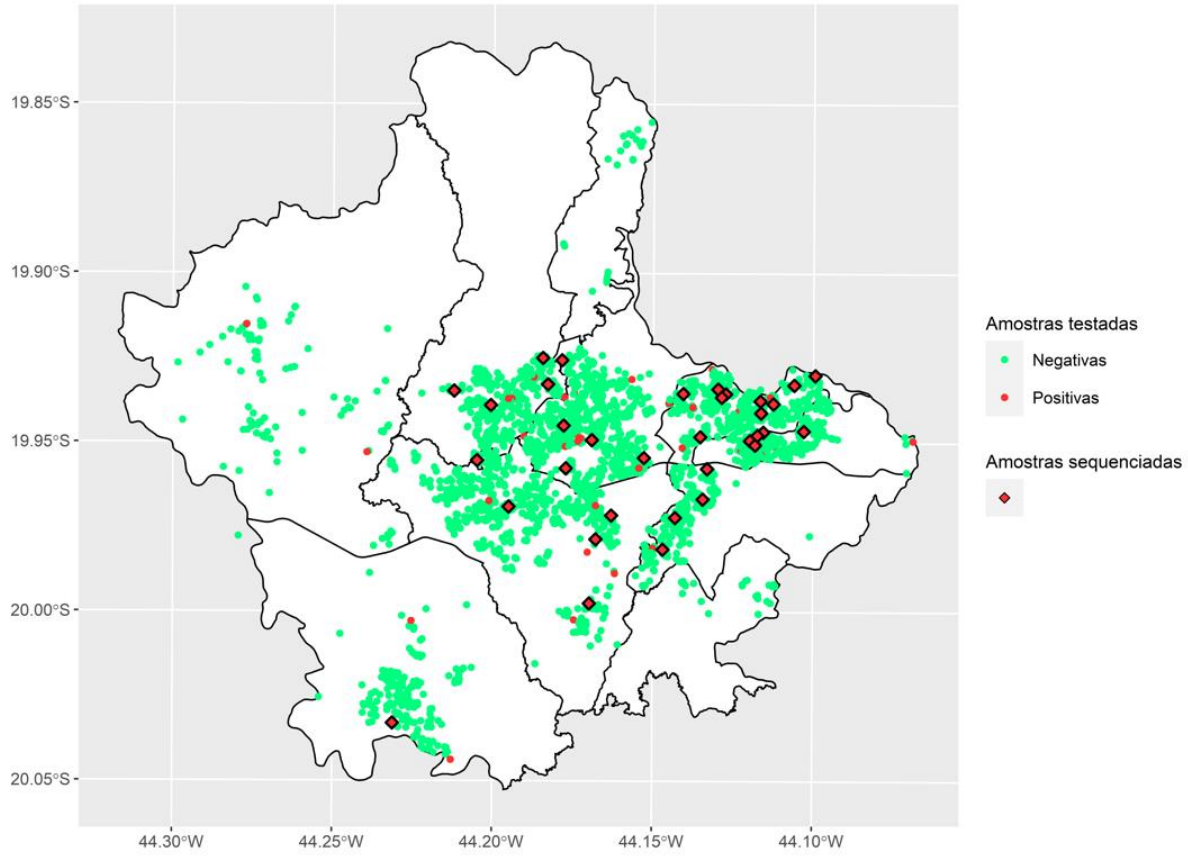

Figura 8. Distribuição espacial das amostras de Betim avaliadas nesse estudo. O mapa mostra a localização das 3239 amostras coletadas e testadas via RT-qPCR. As 3156 amostras que testaram negativo aparecem em verde. Em laranja, aparecem as 84 amostras que testaram positivo, com destaque (losango) para as 35 selecionadas para sequenciamento de genoma completo do SARS-CoV-2.

Após a triagem por meio dos valores de CT (≤ 30) da amplificação da reação, 39 das 84 amostras detectáveis foram consideradas elegíveis para a montagem de bibliotecas e sequenciamento. Após o processamento dos dados e análise qualitativa, 35 bibliotecas atingiram cobertura mínima superior a 79,4% e profundidade mínima de 239x o genoma total do vírus . Estas bibliotecas passaram pelo pipeline bioinformático, o que resultou em 35 novas sequências-consenso de SARS-CoV-2. Para determinar as linhagens circulantes, os genomas obtidos foram classificados por meio do *software* PANGOLIN 2.0[46] que identifica, por alinhamento, a presença das mutações definidoras de cada linhagem. A análise agrupou as 35 sequências em 2 clados: 18 (51,4%) genomas foram classificados na linhagem B.1.1.28 e 17 (48,6%) genomas foram classificados na linhagem B.1.1.33. (Tabela 3).

Tabela 3. Metadados dos 35 genomas de SARS-CoV-2 obtidos no estudo. A tabela inclui dados temporais e métricas referentes à amostra, tais como os CTs de amplificação da reação de RT-qPCR, o número médio de vezes que cada nucleotídeo foi lido no sequenciamento (profundidade média) e a porcentagem do genoma que foi coberta pelo sequenciamento (cobertura) além da classificação dos 35 genomas obtidos.

ID	Data de coleta	CT N1 (CDC)	CT N2 (CDC)	Rodada de coleta	Profundidade média	Cobertura	Linhagem
betim_0013	03/06/2020	8,99	9,21	1	1080,49	99,20%	B.1.1.33
betim_1128	24/06/2020	19,4	19,8	2	250	97,87%	B.1.1.33
betim_1155	23/06/2020	18,11	18,7	2	239	97,73%	B.1.1.28
betim_1338	24/06/2020	15,84	16,44	2	5622,51	99,70%	B.1.1.28
betim_1367	23/06/2020	12	12,32	2	1262,4	99,84%	B.1.1.33
betim_1520	23/06/2020	10,29	10,03	2	2670,83	99,86%	B.1.1.33
betim_1521	24/06/2020	11,9	12,44	2	3056	99,84%	B.1.1.28
betim_1706	23/06/2020	24,03	24,8	2	6275,13	97,63%	B.1.1.28
betim_1730	24/06/2020	13,89	14,09	2	4495,43	99,27%	B.1.1.33
betim_1756	24/06/2020	10,93	11,43	2	2972,99	99,86%	B.1.1.28
betim_1806	24/06/2020	22,28	22,41	2	10040,7	99,08%	B.1.1.33
betim_1834	24/06/2020	15,71	17,48	2	2027,59	99,84%	B.1.1.33
betim_1853	23/06/2020	11,21	11,22	2	996,48	95,28%	B.1.1.28
betim_1905	25/06/2020	22	23	2	6910,15	97,79%	B.1.1.28
betim_1957	23/06/2020	19,84	20,46	2	8795,96	89,12%	B.1.1.28
betim_2224	13/07/2020	9,48	10,04	3	13991,7	99,31%	B.1.1.28
betim_2256	14/07/2020	11,5	11,39	3	4972,17	98,02%	B.1.1.28
betim_2296	14/07/2020	21,99	22,46	3	7961,24	99,20%	B.1.1.33
betim_2405	13/07/2020	14,62	15,07	3	4070,16	99,32%	B.1.1.33
betim_2421	13/07/2020	19,44	20,87	3	3964,35	94,34%	B.1.1.28
betim_2427	13/07/2020	24,66	25,71	3	5320,31	79,40%	B.1.1.33
betim_2494	15/07/2020	24,32	25,08	3	10072	98,78%	B.1.1.28
betim_2621	14/07/2020	15,24	15,66	3	4050,13	97,95%	B.1.1.33
betim_2624	14/07/2020	9,58	9,97	3	2796	92,58%	B.1.1.33
betim_2626	13/07/2020	19,78	19,96	3	1129,37	94,82%	B.1.1.33
betim_2674	13/07/2020	19,35	20,09	3	6484,77	99,64%	B.1.1.33
betim_2769	13/07/2020	22,21	22,86	3	1210,8	81,73%	B.1.1.28
betim_2791	13/07/2020	11,9	12,25	3	2036,86	99,07%	B.1.1.28
betim_2808	14/07/2020	10,67	10,5	3	3741,48	94,73%	B.1.1.33
betim_2892	15/07/2020	22,45	23,63	3	11672,1	99,26%	B.1.1.28
betim_2905	14/07/2020	10,71	10,52	3	5093,82	95,97%	B.1.1.33
betim_2933	13/07/2020	16,26	16,61	3	3254,58	97,68%	B.1.1.33
betim_2964	15/07/2020	19,11	20,25	3	1762,05	94,30%	B.1.1.28
betim_3167	15/07/2020	11,83	12,18	3	5015,61	99,41%	B.1.1.28
betim_3231	14/07/2020	10,1	9,71	3	15183,4	99,73%	B.1.1.28

A cobertura média das sequências obtidas no nosso estudo foi de 96,78%, o que é considerado um ótimo parâmetro para tecnologias de sequenciamento de nova geração. As

linhagens encontradas nas 35 sequências obtidas (B.1.1.28 e B.1.1.33) estão de acordo com o esperado tendo em vista as linhagens dominantes na época da primeira onda da pandemia no estado de Minas Gerais.

A reconstrução filogenética dos genomas completos de SARS-CoV-2 oriundos da cidade de Betim-MG foi realizada através do modelo de máxima verossimilhança utilizando sequencias-referência brasileiras e internacionais. A árvore filogenética não-enraizada gerada demonstra que as 35 sequências de Betim se agruparam com sequências Brasileiras, o que aponta para a transmissão comunitária a nível nacional na cidade, sendo 18 (51,4%) incluídas no clado B.1.1.28 e 17 (48,6%) no clado B.1.1.33, confirmando a identidade atribuída de maneira preliminar através do *software* Pangolin[46] (Fig. 9).

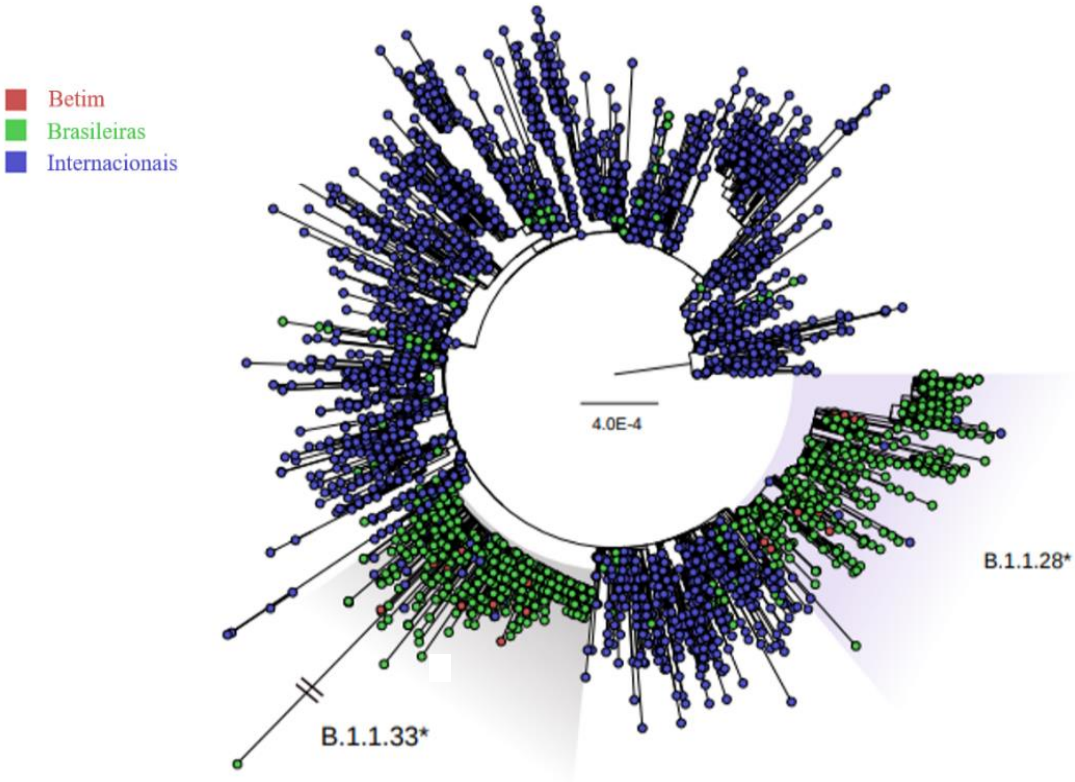

Figura 9. Reconstrução filogenética dos 35 genomas de SARS-CoV-2 de Betim-MG. A árvore não-enraizada foi construída através de um modelo de máxima verossimilhança com 3814 sequencias-referência brasileiras e internacionais. Os genomas de Betim estão representados pelos ramos vermelhos, outros genomas nacionais em verde e genomas internacionais em azul.

O modelo apontou ainda a existência de diversos ramos entre as sequências de Betim para ambas as linhagens, sugerindo que a cidade passou por múltiplos eventos de introdução do SARS-CoV-2. Estas evidências serão melhor exploradas a partir do modelo filogeográfico.

Os resultados demonstram que as linhagens circulantes de SARS-CoV-2 na cidade de Betim -MG refletem a situação da pandemia de COVID-19 no estado de Minas Gerais e estados vizinhos nos meses de junho a julho de 2020, ao confirmarmos as linhagens B.1.1.28 e B.1.1.33, prevalentes no país no momento do estudo. Além disto sugerem a predominância da transmissão comunitária a nível nacional nas introduções do SARS-CoV-2 em Betim-MG.

4.2. DISPERSÃO DO SARS-CoV-2 EM BETIM-MG

Para avaliar a dispersão do SARS-CoV-2 no município de Betim-MG, os resultados positivos para o diagnostico molecular de COVID-19 e do sequenciamento foram incluídos juntamente com os dados de latitude e longitude em modelos de interpolação (Fig. 10).

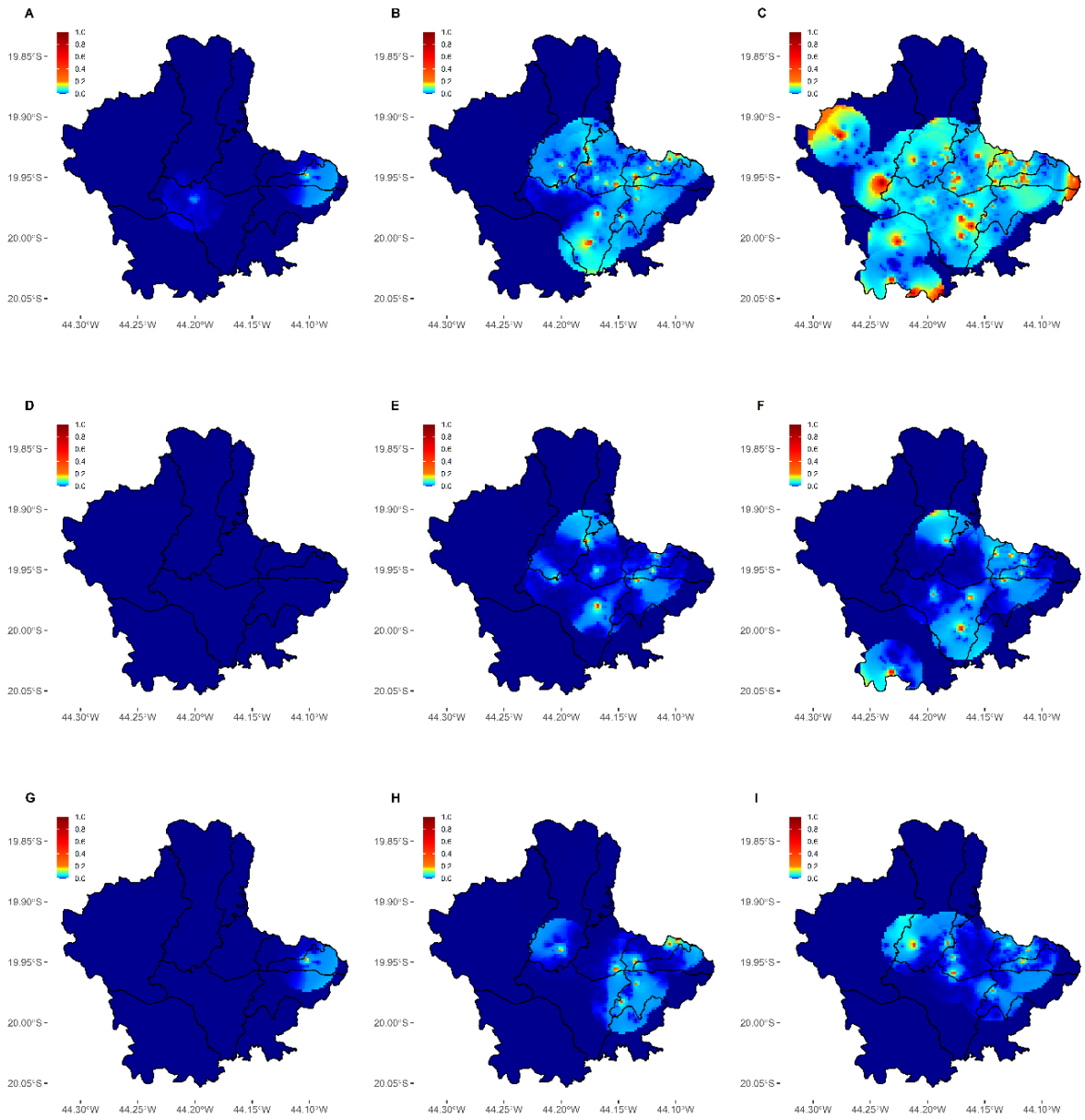

Figura 10. Dispersão do SARS CoV-2 nas macrorregiões de Betim ao longo das 3 rodadas do estudo. Mapas de calor mostrando a prevalência (0-1 em escala de azul marinho à vermelho) do SARS-CoV-2 nas macrorregiões de Betim-MG ao longo das 3 rodadas dispostas em ordem cronológica da esquerda para a direita (A, D e G: de 03 a 05 de junho; B, E e H: de 23 a 25 de junho e C, F e I: de 13 a 15 de julho). Os modelos mostram a dispersão do SARS-CoV-2 a partir de amostras: detectáveis (A-C), classificadas como B.1.1.28 (D-F) e classificadas como B.1.1.33 (G-I), na cidade de Betim-MG.

Os modelos generalistas incluindo todas as amostras detectáveis, bem como os modelos incluindo as linhagens B.1.1.28 e B.1.1.33 nas 3 rodadas de coleta mostraram um avanço geral das infecções que coincide com o curso da pandemia ao longo do estudo. Os

modelos relativos às linhagens, no entanto, não sugerem a existência de padrões de dispersão específicos para cada clado de SARS-CoV-2, uma vez que há considerável sobreposição entre os modelos de ambas as linhagens

4.3. ESTIMATIVA DE INTRODUCAO DAS LINHAGENS DE SARS-CoV-2 EM BETIM-MG

Visando traçar a história temporal e espacial das introduções de cada clado de SARS-CoV-2 em Betim-MG modelos filogeográficos datados com estatística Bayesiana (BEAST v.1.10.4[91] - HKY+I+G4[92,93]) foram construídos para as linhagens identificadas na cidade dando origem a 2 árvores filogenéticas enraizadas.

A árvore montada a partir de sequências da linhagem B.1.1.28 aponta para pelo menos 7 introduções distintas desta linhagem, sendo: 1 com provável origem em São Paulo, 4 com provável origem em outros estados Brasileiros e 2 com provável origem internacional. O modelo mostra ainda 2 clusters, cada um contendo 6 sequências obtidas em Betim, o que sugere que para esta linhagem, predominou a transmissão comunitária (66,67%) em comparação com introduções externas (33,33%) do SARS-CoV-2. A datação do modelo sugere ainda que a introdução mais antiga da linhagem B.1.1.28 na cidade de Betim-MG se deu na segunda metade mês de abril, com intervalo de confiança entre 17 de abril e 11 de maio de 2020, pouco mais de um mês depois do primeiro caso positivo reportado na cidade[43] (Fig. 11).

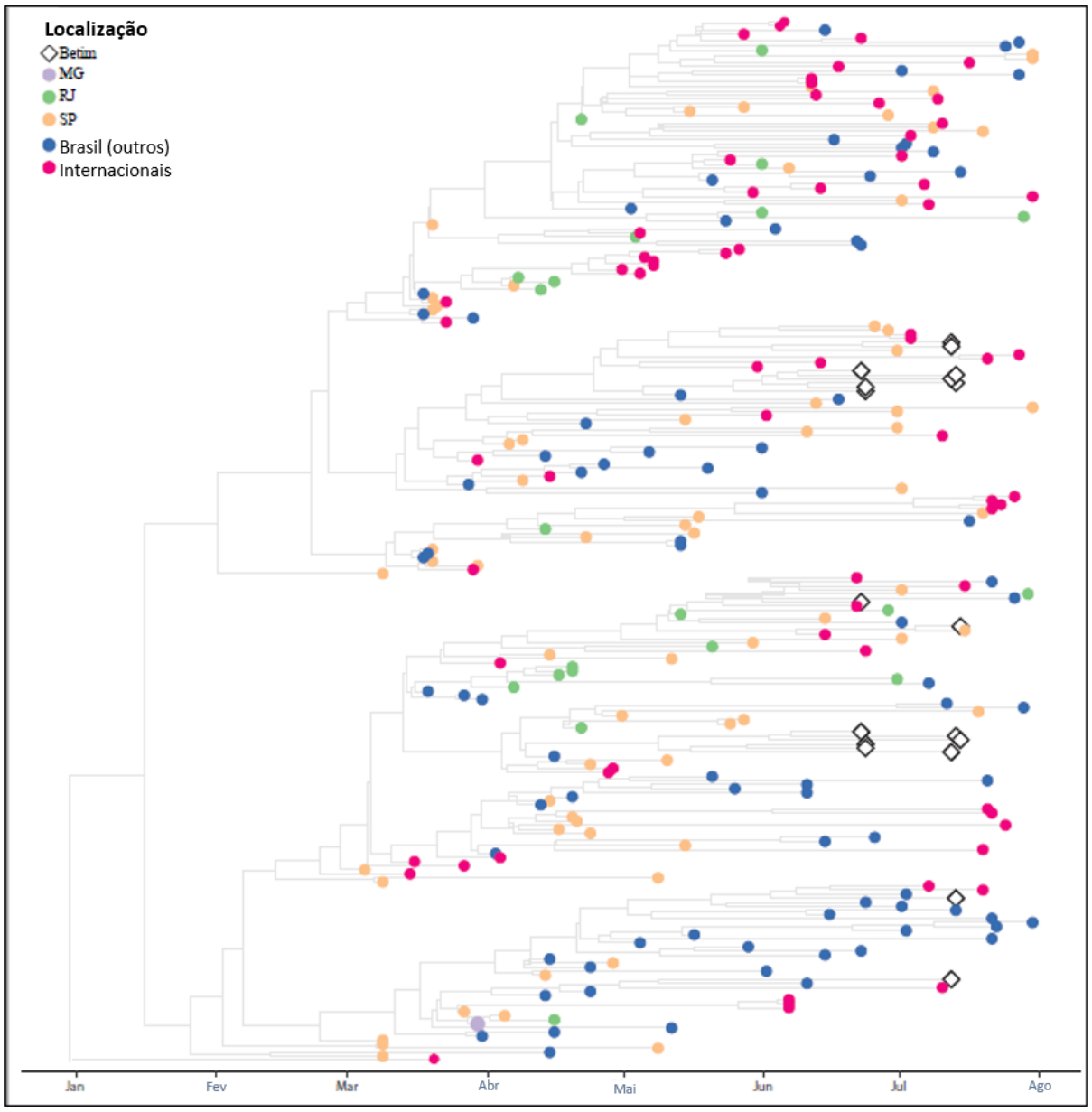

Figura 11. Reconstrução filogeográfica por estatística Bayesiana dos genomas da linhagem B.1.1.28 obtidos em Betim-MG. As 251 sequências que compõem a base de dados do modelo estão representadas nas seguintes cores: Betim (18) em losango branco, MG (2) em lilás, RJ (21) em verde, SP (70) em laranja, Outros estados brasileiros (70) em azul e internacionais (70) em rosa.

Para a linhagem B.1.1.33, o modelo aponta para pelo menos 12 introduções distintas desta linhagem, sendo: 2 com provável origem em São Paulo, 5 com provável origem no Rio de Janeiro, 2 com provável origem em outros estados Brasileiros e 3 com provável origem internacional. O modelo mostra 3 clusters: 2 contendo cada um contendo 2 sequências e 1 contendo 4 sequências obtidas em Betim. Portanto, para esta linhagem o modelo sugere que predominou a ocorrência de introduções externas (52,94%) em relação à transmissão

comunitária (47,06%) do SARS-CoV-2. A datação do modelo sugere ainda que a introdução mais antiga da linhagem B.1.1.33 na cidade de Betim-MG se deu no início do mês de fevereiro, com intervalo de confiança entre 14 de janeiro de 2020 e 25 de fevereiro de 2020, curiosamente, um mês antes do primeiro caso positivo reportado na cidade[43]. O genoma de B.1.1.33 cujo agrupamento apresenta a datação mais antiga, Betim_2964, chamou a atenção pelo tamanho do seu ramo e sua posição em relação às demais obtidas no nosso estudo, se agrupando com outra sequência de B.1.1.33 oriunda do Rio de Janeiro em um ramo distante de todas as outras sequencias desta linhagem (Fig. 12).

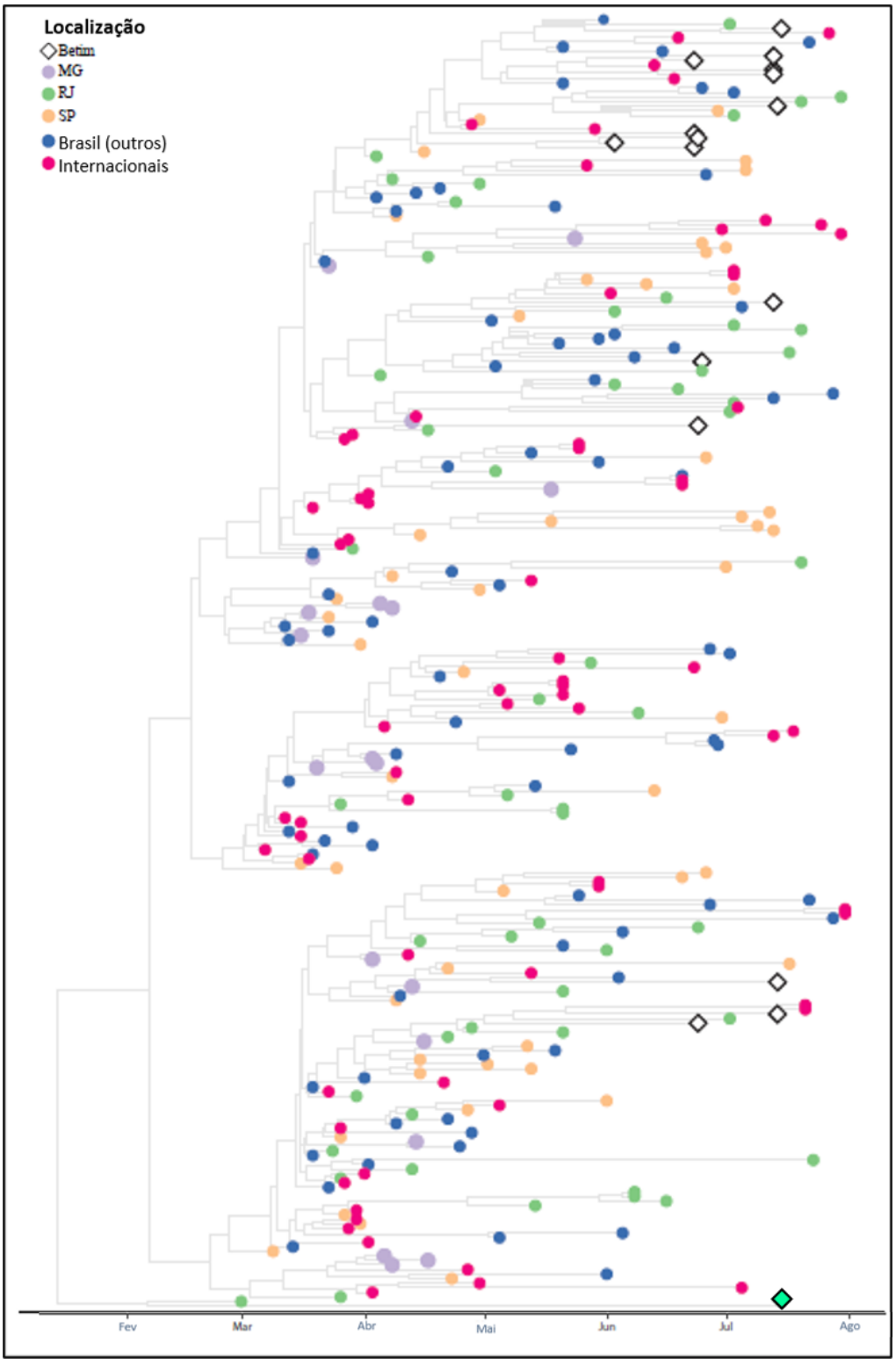

Figura 12. Reconstrução filogeográfica por estatística Bayesiana dos genomas da linhagem B.1.1.33 obtidos em Betim-MG. As 249 sequências que compõem a base de dados do modelo estão representadas nas seguintes cores: Betim (17) em losango branco, MG (20) em lilás, RJ (53) em verde, SP (53) em laranja, Outros estados brasileiros (53) em azul e internacionais (53) em rosa. O genoma Betim_2964 está destacado (losango verde-água) devido ao tamanho do seu ramo e à sua posição singular.

4.4. GENOMA ENCONTRADO EM BETIM-MG TEM DOZE MUTAÇÕES EXCLUSIVAS

A comparação das sequências geradas com o genoma de referência do SARS-CoV-2 (NC_045512.2) revelou a presença das mutações características de cada linhagem[46] em todas as 35 sequências obtidas. Além das mutações definidoras, 58 outros polimorfismos de nucleotídeo único (SNP) foram encontrados nas 18 sequências da linhagem B.1.1.28 (cobertura média: 96,72% do genoma). Para a linhagem B.1.1.33, 70 outros polimorfismos deste tipo foram encontrados (cobertura média: 96,84% do genoma). A lista completa dos polimorfismos encontrados nas 35 sequências obtidas neste estudo está representada ao lado da frequência de aparecimento destes polimorfismos (Anexo 2). A sequência Betim_2964 (cobertura: 94,3% do genoma), que agrupou-se em um longo ramo distante dos demais genomas de B.1.1.33, apresentou 12 polimorfismos exclusivos espalhados por todo o genoma - 8 a mais que a média de polimorfismos exclusivos em outras sequências desta linhagem (3,625) - sendo 7 no gene da replicase (Orf1ab), 2 no gene da proteína *Spike* (S), 1 na Orf3, 1 na Orf8b e 1 no gene do núcleocapsídeo (N). Destes 12 polimorfismos, 8 dão origem a mutações não-sinônimas e 4 a mutações sinônimas (Fig. 13).

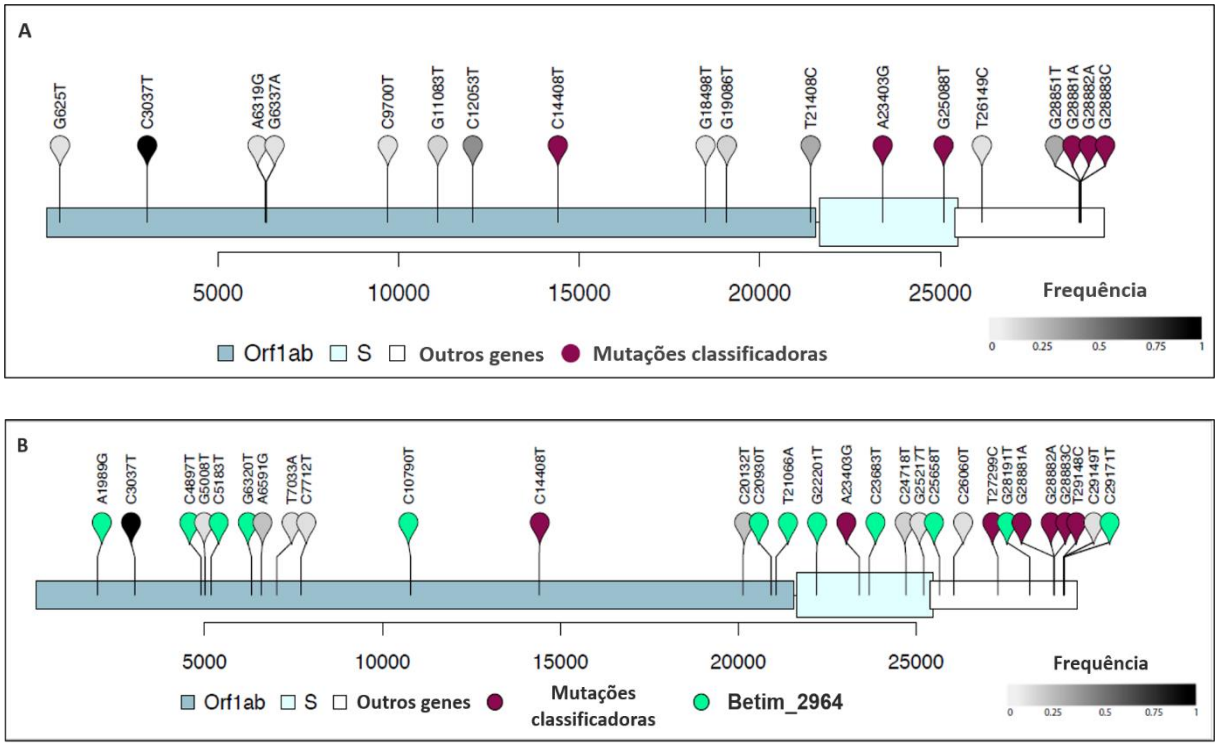

Figura 13. Localização dos polimorfismos presentes nas sequências encontradas em Betim-MG no genoma do SARS-CoV-2. O mapa mostra os polimorfismos presentes em pelo menos 2 sequências das linhagens B.1.1.28 (A) e B.1.1.33 (B). As mutações classificadoras estão representadas em vinho. A frequência (0-1) dos demais polimorfismos está representada em

escala de cinza. O mapa dos genomas da linhagem B.1.1.33 destaca ainda os 12 polimorfismos exclusivos da sequência Betim_2964 em verde-água.

Os resultados da análise de polimorfismos em nossas sequências corroboraram para a confirmação das linhagens encontradas em Betim-MG a partir da análise das mutações de sítio único definidoras de cada variante. A análise revelou ainda um genoma da linhagem B.1.1.33 com mutações estruturais em todo o genoma, a maioria não-sinônimas. O achado ajuda a elucidar a distância desta sequência em relação às demais no modelo de B.1.1.33, e levanta a necessidade de investigar futuramente a prevalência desta variante no cenário local.

5. DISCUSSÃO

O presente trabalho teve como objetivo realizar um estudo de vigilância genômica do vírus SARS-CoV-2 durante a fase inicial da pandemia de COVID-19 na cidade de Betim-MG. As amostras empregadas neste estudo são oriundas de um trabalho anterior que avaliou a soro-prevalência do vírus em 3 rodadas a partir da sétima[43] semana epidemiológica da cidade no ano de 2020. A comunicação dos resultados deste estudo ocorreu em boletins epidemiológicos e por meio de transmissões ao vivo em parceria com a prefeitura da cidade de maneira concomitante ao término de cada rodada[98]. As amostras em que foi detectado o RNA do SARS-CoV-2 foram incluídas em um fluxo de triagem, montagem de bibliotecas, sequenciamento de genoma completo e destinadas à vigilância genômica propriamente dita. O estudo sequenciou 35 genomas completos, classificados posteriormente nas linhagens B.1.1.28 (18) e B.1.1.33 (17). A análise filogeográfica revelou múltiplas introduções para ambas as linhagens, que se dispersaram de maneira semelhante, de acordo com modelos de dispersão. Um dos genomas de B.1.1.33 encontrados apresentou 12 mutações estruturais exclusivas.

A tecnologia de sequenciamento empregada neste estudo gera fragmentos de leitura curtos (*short reads*). Os fragmentos foram combinados por meio de *softwares* da bioinformática no processo de montagem das sequências-consenso de SARS-CoV-2. Diferentes metodologias são empregadas em estudos de vigilância genômica de SARS-CoV-2 e outros patógenos, a exemplo das tecnologias Oxford Nanopore© (ONT), que geram fragmentos de leitura longos (*long reads*). Entretanto, tecnologias *short read* são consideradas ideais para estudos que visam realizar a filogenia do SARS-CoV-2, uma vez que a baixa taxa de mutação dos coronavírus faz com que vírus de diferentes origens tenham poucas trocas nucleotídicas entre si, fato que exige que os genomas sejam montados com alta profundidade (> 100x) de leitura para cada nucleotídeo[99]. A alta profundidade é um atributo vantajoso das tecnologias *short read* que se origina da inerente complexidade da montagem de sequências-

consenso com fragmentos gerados por este tipo de metodologia. O controle de qualidade levando em conta a cobertura genômica de sequências montadas a partir de dados de *Next-Generation Sequencing* (NGS) é um problema clássico dos estudos de filogenia viral[100], o que faz com que diferentes trabalhos atribuam pontos de corte qualitativos diversos (75-90%) para sequências empregadas em análises filogenéticas[50,101].Os 35 genomas gerados neste estudo utilizados para reconstruções de filogenia apresentaram profundidade média mínima de 239x para cada nucleotídeo e cobertura de 96,78% do genoma total do vírus, o que são considerados bons parâmetros de acordo com a literatura consolidada. Dentre os 35 genomas obtidos neste estudo, 8 apresentaram cobertura inferior a 95% (79,4%-94,8%), sendo adequados para construção de filogenias e classificação de linhagens, mas não para anotação de novas variantes[46] (Tabela 3). Entretanto, a identificação de mutações classificadoras de cada linhagem e outros polimorfismos típicos, como C241T foi possível também para estes genomas (Anexo 3), o que sugere que apesar da cobertura total ser inferior ao limiar desejável, os valores de cobertura e profundidade podem ter sido superiores nas posições genômicas que marcam estas trocas nucleotídicas. Uma das limitações da estratégia de sequenciamento utilizada reside no fato das amostras terem sido selecionadas para montagem de bibliotecas e sequenciamento com base no CT de amplificação da reação de RT-qPCR (≤ 30), o que resultou na exclusão de 45 amostras positivas para SARS-CoV-2 das etapas finais do estudo, potencializando o viés de seleção. A decisão de selecionar as amostras com base no CT se justifica tendo em vista o fato desta métrica estar inversamente relacionada à carga viral da amostra após a extração[102]. A carga viral da amostra por sua vez, está relacionada à eficiência e qualidade do sequenciamento em protocolos de montagem de genomas virais completos como os protocolos de sequenciamento do SARS-CoV-2[103]. Esta decisão se mostrou acertada tendo em vista que 4 das 39 bibliotecas que não apresentaram boas métricas de sequenciamento, possuíam CT entre 29 e 30, o que reforça a correlação deste parâmetro com a qualidade dos dados de sequenciamento. Espera-se que em estudos futuros de vigilância genômica, estratégias extração de RNA, montagem de bibliotecas e sequenciamento capazes de lidar com amostras que apresentam baixa carga viral sejam consideradas, de modo a minimizar o viés de seleção aumentando a representatividade da amostra.

Para avaliar a dispersão geográfica do SARS-CoV-2 na cidade de Betim-MG, a coleta das amostras foi realizada através de busca domiciliar ativa de maneira randomizada e estratificada geograficamente, de modo a minimizar o viés espacial e propiciar uma boa cobertura da área urbanizada do município. O estudo não levou em conta o quadro clínico dos participantes antes do cadastramento dos domicílios, sendo assim, um dos primeiros estudos do país que obteve amostras detectáveis de indivíduos sintomáticos e

assintomáticos, evitando desta forma o viés de seleção. O desenho do estudo levou em conta ainda o fator temporal e distribuiu o mesmo número de coletas (1080) em 3 rodadas de 3 dias de trabalho cada com duração de 3 de junho a 15 de julho de 2020 respeitando um espaçamento de 17 dias entre o fim de uma rodada e o início da próxima, visando respeitar o período necessário para a soro-conversão de novos casos. O estudo foi assim estruturado visando capturar o provável crescimento numérico dos casos positivos acompanhado da propagação espacial da pandemia entre as rodadas, o que de fato foi verificado nos resultados e pode ser observado nos painéis A-C dos modelos de dispersão (Fig. 10). A presença de supostos padrões de dispersão para cada linhagem pode estar relacionada à sub-representação de determinadas regiões que surge, dentre outros fatores, em decorrência da triagem de amostras positivas para sequenciamento com base no CT de amplificação que não levou em conta o viés espacial. Assim, o fato do sul da região Citrolândia (limite sudoeste do município) apresentar um foco apenas no painel da terceira rodada da linhagem B.1.1.33 (F) não implica dizer que esta região não foi afetada pela introdução de outras linhagens no momento do estudo. A ausência de padrões específicos de cada linhagem sugere que outros fatores, como o grau de urbanização das diferentes regiões do município (Fig. 7) foi mais determinante na prevalência do vírus na cidade. A comunicação dos resultados de soro-prevalência nas diferentes regiões foi importante para a adoção de NPIs à nível local no momento do estudo (início da pandemia na cidade). Entretanto, com o avanço da COVID-19, a detecção do SARS-CoV-2 foi recorrentemente reportada em todas as regiões do município, sendo os casos positivos distribuídos de acordo com a densidade demográfica na cidade[43]. Mapeamentos constantes da frequência de linhagens do SARS-CoV-2 com coleta de amostras representativas em todas regiões de Betim são necessários para uma aproximação mais fidedigna da dispersão do vírus no município.

As reconstruções filogenéticas identificaram as linhagens de SARS-CoV-2 B.1.1.28 e B.1.1.33 em circulação no município de Betim-MG entre junho e julho de 2020. Ambas as linhagens são derivadas do clado B.1, que apresenta a mutação S:D614G, que se consolidou evolutivamente por conferir aos vírus portadores deste polimorfismo uma infectividade significativamente maior[47]. Este resultado é fortemente corroborado pela literatura através de trabalhos que realizaram a vigilância genômica do vírus no país com intervalo de coleta de amostras compatível ao do presente estudo[53-56]. No entanto, o número de amostras efetivamente sequenciadas pode não ter sido suficiente para inferir adequadamente a frequência das linhagens circulantes na cidade. O viés amostral pode, por exemplo, ter omitido linhagens que porventura estivessem circulando em frequência suficientemente baixa ao ponto de não serem representadas por uma amostra de 35 genomas. A história temporal e espacial da introdução do SARS-CoV-2 no município de Betim foi estimada através de

modelos filogeográficos independentes para cada linhagem. Ambos os modelos sugerem que as linhagens encontradas em Betim são oriundas de introduções majoritariamente nacionais, o que corrobora com a observação de que estas sequências circulavam amplamente pelo país no momento do estudo.[50-56]. Um fato que chama a atenção é a diferença no grau de agrupamento dos genomas de Betim nas diferentes linhagens. As sequências da linhagem B.1.1.28 se apresentam majoritariamente em *clusters*, o que contrasta com as sequências da linhagem B.1.1.33, que aparecem muito mais dispersas (Fig. 11; Fig. 12). Esta disposição sugere uma transmissão mais comunitária para B.1.1.28 e mais externa para B.1.1.33. Os modelos filogeográficos mostraram que a linhagem B.1.1.28 apresentou menos eventos de introdução (7) do que a linhagem B.1.1.33 (12). Os modelos apontaram ainda que a linhagem B.1.1.28 teria sido introduzida 3 meses mais tarde na cidade de Betim que a linhagem B.1.1.33. Curiosamente, a despeito do histórico de ambas as linhagens, encontramos sequências classificadas como B.1.1.28 (18) na mesma proporção que sequências classificadas como B.1.1.33 (17) no nosso estudo. Espera-se que posteriormente a consolidação dos dados a respeito da evolução da frequência de ambas as linhagens nos cenários estadual e nacional em uma janela de tempo compatível com a duração do nosso estudo ajudem a elucidar a dinâmica destas variantes. Os levantamentos de frequência de variantes do SARS-CoV-2 realizados após o pico da segunda onda da pandemia no estado de Minas Gerais mostram o crescimento e consolidação da prevalência da VOC P.1 (Gama) de janeiro a julho de 2021 e sua dinâmica com as linhagens B.1.1.7 (Alfa) e P.2 (Zeta)[104,105]. Posteriormente, após a variante B.1.617.2 (Delta) ter sido reportada no país[68], a proporção de infectados por esta linhagem rapidamente aumentou até superar todas as demais no cenário nacional atual[106]. Estudos constantes de vigilância genômica são necessários para manter atualizadas as frequências relativas das variantes do SARS-CoV-2 em diferentes contextos.

O modelo filogeográfico da linhagem B.1.1.33 evidenciou ainda um dos genomas de Betim (Betim_2964), que aparece em um longo ramo como grupo irmão de uma amostra do Rio de Janeiro. As duas amostras estão posicionadas como o grupo mais basal de todas as sequências incluídas no modelo (Fig. 12). Visando interpretar as posições das amostras de ambas as linhagens identificadas em Betim, foram construídos dois mapas de variantes que incluem a frequência e posição dos polimorfismos que aparecem em pelo menos 2 sequências de cada linhagem. Especial destaque foi dado aos polimorfismos da amostra Betim_2964, uma vez que a análise revelou 12 polimorfismos específicos desta amostra distribuídos por todo o genoma. Estes resultados ajudaram a explicar a distância desta amostra em relação às outras no modelo filogeográfico. Embora os requisitos mínimos[46] para que se possa inferir quaisquer afirmações não tenham sido encontrados, a presença de um genoma estável e com boa cobertura apresentando um significativo número de trocas nucleotídicas levanta a

hipótese de que uma introdução muito antiga de B.1.1.33 tenha acontecido na cidade ou ainda uma nova variante em circulação. Novos estudos de vigilância genômica focados na região são necessários para investigar ambas as hipóteses.

6. CONCLUSÃO

O presente trabalho identificou as linhagens B.1.1.28 e B.1.1.33 de SARS-CoV-2 circulando na cidade de Betim-MG entre 3 de junho e 15 de julho de 2020. A dispersão do vírus aumentou significativamente durante o período do estudo, mas sem a presença de padrões geográficos relacionados às linhagens. Os resultados dos modelos filogeográficos elaborados sugerem que múltiplas introduções ocorreram na cidade, embora para a linhagem B.1.1.28 a transmissão tenha se dado de maneira mais comunitária ao passo que para a linhagem B.1.1.33 ocorreram mais introduções externas do vírus. No modelo da linhagem B.1.1.33, identificamos ainda um genoma que possui doze mutações exclusivas além das mutações classificadoras da linhagem e outros polimorfismos compartilhados. Para trazer uma representação mais apropriada do panorama da evolução das linhagens do SARS-CoV-2 em Betim, a condução de estudos futuros de vigilância genômica robustos será necessária. Visando contornar o viés de seleção, sugere-se o investimento em estratégias de sequenciamento capazes de lidar com amostras com valores de CT altos. Apesar das barreiras aqui descritas, espera-se que o resultado das frequências das linhagens virais encontrado neste estudo possa melhorar a resolução da vigilância do SARS-CoV-2 nos meses de junho e julho a nível local e nacional. Espera-se ainda que os resultados dos modelos filogeográficos ajudem a revelar o papel das estradas interestaduais na dispersão do vírus em um município. A interpretação cuidadosa destes resultados pode guiar autoridades competentes na adoção de medidas não-farmacológicas importantes, minimizando o impacto da pandemia e deve trazer à tona a importância prática dos estudos de vigilância genômica viral no país. Os resultados deste trabalho foram compilados em um artigo recentemente submetido no arquivo de pré-prints *medRxiv* e na revista *Frontiers of Microbiology* e estão em revisão para publicação[107] (Anexo 4).

7. REFERÊNCIAS

1. Almeida JD, Berry DM, Cunningham CH, et al. Coronaviruses. **Nature**; 220:650, 1968.

2. Corman VM, Muth D, Niemeyer D, Drosten C. Hosts and Sources of Endemic Human Coronaviruses. **Adv Virus Res.** 2018;100:163-188, fev. 2018.

3. Masters PS, Perlman S. Coronaviridae. In: **Fields virology**. 6th ed. Lippincott Williams & Wilkins: Knipe DM, Howley PM, eds.: 2013. 825-58.

4. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., . . . Tan, W .A Novel Coronavirus from Patients with Pneumonia in China, 2019. **New England Journal of Medicine**, 382(8), pp.727-733, 2020.

5. Smith, E., Sexton, N., & Denison, M. Thinking Outside the Triangle: Replication Fidelity of the Largest RNA Viruses. **Annual Review Of Virology**, 1(1), 111-132, 2014.

6. Woo, P., Huang, Y., Lau, S., & Yuen, K. Coronavirus Genomics and Bioinformatics Analysis. **Viruses**, 2(8), 1804-1820, 2010.

7. Woo, P., Lau, S., Huang, Y., & Yuen, K. Coronavirus Diversity, Phylogeny and Interspecies Jumping. **Experimental Biology And Medicine**, 234(10), 1117-1127, 2009.

8. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. **Nat. Microbiol.** 5, 536–544, 2020.

9. Hamre, D., Procknow, J.J., A new virus isolated from the human respiratory tract. **Proc. Soc. Exp. Biol. Med.** 121, 190–193, 1966.

10. McIntosh, K., Dees, J.H., Becker, W.B., Kapikian, A.Z., Chanock, R.M. Recovery in tracheal organ cultures of novel viruses from patients with respiratory disease. **Proc. Natl. Acad. Sci. U. S. A.** 57, 933–940, 1967.

11. Tyrrell, D., & Bynoe, M. Cultivation of a Novel Type of Common-cold Virus in Organ Cultures. **BMJ**, 1(5448), 1467-1470, 1965.

12. Fehr, A., & Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. **Coronaviruses**, 1-23, 2015.

13. Drosten C, Günther S, Preiser W, van der Werf S, Brodt HR, Becker S, . . . Doerr HW. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. **N Engl J Med**; 348(20):1967-76, mai. 2003.

14. Rabaan AA, Al-Ahmed SH, Haque S, Sah R, Tiwari R, Malik YS, . . . Rodriguez-Morales AJ. SARS-CoV-2, SARS-CoV, and MERS-COV: A comparative overview. **Infez Med**. Ahead Of Print;28(2):174-184, jun. 2020.

15. van der Hoek, L., Pyrc, K., Jebbink, M.F., Vermeulen-Oost, W., Berkhout, R.J., Wolthers, K.C., . . . Berkhout, B. Identification of a new human coronavirus. **Nat. Med.** 10, 368–373, 2004.

16. Woo, P.C., Lau, S.K., Chu, C.M., Chan, K.H., Tsoi, H.W., Huang, Y., . . . Yuen, K.Y. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. **J. Virol.** 79, 884–895, 2005.

17. Nassar MS, Bakhrebah MA, Meo SA, Alsuabeyl MS, Zaher WA. Middle East Respiratory Syndrome Coronavirus (MERS-CoV) infection: epidemiology, pathogenesis and clinical characteristics. **Eur Rev Med Pharmacol Sci.** (15):4956-4961, ago. 2018.

18. Zaki, A.M., van Boheemen, S., Bestebroer, T.M., Osterhaus, A.D., Fouchier, R.A., Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. **N. Engl. J. Med.** 367, 1814–1820, 2012.

19. Volpato G, Fontefrancesco MF, Gruppuso P, Zocchi DM, Pieroni A. Baby pangolins on my plate: possible lessons to learn from the COVID-19 pandemic. **J Ethnobiol Ethnomed.**16(1):19, abr. 2020.

20. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, . . . Shi Z-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. **Nature.** 579(7798):270–273, 2020.

21. Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, . . . Shi W. A novel bat coronavirus reveals natural insertions at the s1/s2 cleavage site of the spike protein and a possible recombinant origin of hcov-19. **Curr Biol.** 30(19):3896, 2020.

22. Pipes L, Wang H, Huelsenbeck JP, Nielsen R. Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. **Mol Biol Evol.** 2020.

23. Yao H, Song Y, Chen Y, Wu N, Xu J, Sun C, . . . Li S. Molecular Architecture of the SARS-CoV-2 Virus. **Cell.** 183(3), 730-738.e13, 2020.

24. Tortorici, M., & Veesler, D. Structural insights into coronavirus entry. **Advances In Virus Research.** 93-116, 2019.

25. Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., . . . Tan, W. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. **The Lancet.** 395(10224), 565-574, 2020.

26. Kim, D., Lee, J., Yang, J., Kim, J., Kim, V., & Chang, H. The Architecture of SARS-CoV-2 Transcriptome. **Cell**. 181(4), 914-921.e10, 2020.

27. Bai C, Zhong Q, Gao GF. Overview of SARS-CoV-2 genome-encoded proteins. Sci China **Life Sci.** 10:1–15, 2021.

28. Walls, A., Park, Y., Tortorici, M., Wall, A., McGuire, A., & Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. **Cell.** 181(2), 281-292.e6, 2020.

29. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. **JAMA.** 25;324(8):782-793, ago. 2020.

30. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. **Nat Rev Microbiol.** 19:155–170, 2021.

31. Astuti, I., & Ysrafil. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. **Diabetes & Metabolic Syndrome: Clinical Research & Reviews.** 14(4), 407-412, 2020.

32. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, . . . Qian Z. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. **Nature Communications.** 11(1), 2020.

33. Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., & Li, F. Cell entry mechanisms of SARS-CoV-2. **Proceedings Of The National Academy Of Sciences.** 117(21), 11727-11734, 2020.

34. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. **Antiviral Res.** 176:104742, 2020.

35. Baloch S, Baloch MA, Zheng T, Pei X. The Coronavirus Disease 2019 (COVID-19) Pandemic. **Tohoku J Exp Med.** 250(4):271-278, abr. 2020.

36. Promed Post - ProMED-mail, dez. 2019. Disponível em: https://promedmail.org/promed-post/?id=6864153+#COVID19. Acesso em 10 de janeiro de 2021.

37. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. **Lancet Inf Dis.** 20(5):533-534, 2020. Disponível em: https://github.com/CSSEGISandData/COVID-19. Acesso em 10 de janeiro de 2021.

38. Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, . . . Roser M. Coronavirus Pandemic (COVID-19), 2020. Disponível em: <https://ourworldindata.org/coronavirus>. Acesso em: 09/09/2021.

39. Araujo DB, Machado RRG, Amgarten DE, Malta FM, de Araujo GG, Monteiro CO, . . . Durigon EL. SARS-CoV-2 isolation from the first reported patients in Brazil and establishment of a coordinated task network. **Mem Inst Oswaldo Cruz.** 23;115:e200342, out. 2020.

40. Gatto M, Bertuzzo E, Mari L, Miccoli S, Carraro L, Casagrandi R, Rinaldo A,. Spread and dynamics of the COVID-19 epidemic in Italy: effects of emergency containment measures. **Proc Natl Acad Sci USA.** 117(19):10484-91, 2020.

41. Moscadelli A, Albora G, Biamonte MA, Giorgetti D, Innocenzio M, Paoli S, Lorini C, . . . Bonaccorsi G. Fake News and Covid-19 in Italy: Results of a Quantitative Observational Study. **International Journal of Environmental Research and Public Health.** 17(16):5850, 2020.

42. Coronavírus Brasil, 2021. Disponível em: <https://covid.saude.gov.br/>. Acesso em: 11 de janeiro de 2021.

43. Coronavírus. SES-MG, 2021. Disponível em: <https://coronavirus.saude.mg.gov.br/>. Acesso em: 23 de novembro de 2021.

44. ICTV Code: the International Code of Virus Classification and Nomenclature. International Committee on Taxonomy of Viruses (ICTV), 2018. Acesso em: <https://talk.ictvonline.org/information/w/ictv-information/383/ictv-code/>. Disponível em: 10/10/2021.

45. Domingo E, Perales C. Viral quasispecies. **PLoS Genet.** 15(10):e1008271, out. 2019.

46. Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., . . . Pybus, O.G. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. **Nat Microbiol.** 5, 1403–1407, 2020.

47. Alteri C, Cento V, Piralla A, Costabile V, Tallarita M, Colagrossi L, . . . , Baldanti F. Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. **Nat Commun.** 12(1):434, 2021.

48. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, . . . Connor TR. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. **Cell.** 184(1):64-75.e11, jan. 2021.

49. Jesus JG, Sacchi C, Candido DDS, Claro IM, Sales FCS, Manuli ER, . . . Faria NR. Importation and early local transmission of COVID-19 in Brazil, 2020. **Rev Inst Med Trop Sao Paulo.** 62:e30, 2020.

50. Candido DS, Claro IM, de Jesus JG, Souza WM, Moreira FRR, Dellicour S, Mellan TA, . . . Faria NR. Evolution and epidemic spread of SARS-CoV-2 in Brazil. **Science.** 369(6508):1255-1260, set. 2020.

51. Xavier J, Giovanetti M, Adelino T, Fonseca V, Barbosa da Costa AV, Ribeiro AA, . . . Assunção Oliveira MA. The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. **Emerg Microbes Infect.** (1):1824-1834, dez. 2020.

52. Botelho-Souza LF, Nogueira-Lima FS, Roca TP, Naveca FG, de Oliveria Dos Santos A, Maia ACS, . . . Vieira DS. SARS-CoV-2 genomic surveillance in Rondônia, Brazilian Western Amazon. **Sci Rep.** Feb 12;11(1):3770, 2021.

53. Resende PC, Delatorre E, Gräf T, Mir D, Motta FC, Appolinario LR, . . . Siqueira MM. Evolutionary Dynamics and Dissemination Pattern of the SARS-CoV-2 Lineage B.1.1.33 During the Early Pandemic Phase in Brazil. **Front Microbiol.** 17(11):615280, 2021.

54. Dos Santos CA, Bezerra GVB, de Azevedo Marinho ARRA, Alves JC, Tanajura DM, Martins-Filho PR. SARS-CoV-2 Genomic Surveillance in Northeast Brazil: Timing of Emergence of the Brazilian Variant of Concern P1. **J Travel Med.** 5:taab066, mai. 2021.

55. Paiva MHS, Guedes DRD, Docena C, Bezerra MF, Dezordi FZ, Machado LC, . . . Wallau GL. Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2

at One of the Largest Metropolitan Areas of Northeast Brazil. **Viruses.** 12(12):1414, 2020.

56. Voloch CM, da Silva Francisco R Jr, de Almeida LGP, Cardoso CC, Brustolini OJ, Gerber AL, . . . de Vasconcelos ATR. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. **J Virol.** 1:JVI.00119-21, 2021.

57. World Health Organization. COVID-19 Weekly Epidemiological Update, fev. 2021. "Special edition: Proposed working definitions of SARS-CoV-2 Variants of Interest and Variants of Concern". Disponível em: <https://www.who.int/docs/default-source/coronaviruse/situation-reports/20210225_weekly_epi_update_voc-special-edition.pdf>. Acesso em: 29/07/2021.

58. Centers for Disease Control and Prevention (CDC). SARS-CoV-2 Variant Classifications and Definitions. 2021. Disponível em: <https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html>. Acesso em: 09/09/2021.

59. Bittar C, Possebon FS, Ullmann LS, de Almeida, LGP, Banho CA, Campos C, . . . Araújo Jr, JP. Potential new B.1.1.28 sub-lineage with L452R in Brazil. Disponível em: <https://github.com/charlespwd/project-title>. Acesso em: 29/07/2021.

60. de Siqueira IC, Camelier AA, Maciel EAP, Nonaka CKV, Neves MCLC, Macêdo YSF, . . . Gräf T. Early detection of P.1 variant of SARS-CoV-2 in a cluster of cases in Salvador, Brazil. **Int J Infect Dis.** 108:252-255, jul. 2021.

61. Faria NR, Mellan TA, Whittaker C, Claro IM, Candido DDS, Mishra S, . . . Sabino EC. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. **Science.** 372(6544):815-821, mai. 2021.

62. Francisco RDS Jr, Benites LF, Lamarca AP, de Almeida LGP, Hansen AW, Gularte JS, . . . Spilki FR. 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.** abr. 2021.

63. Resende PC, Gräf T, Paixão ACD, Appolinario L, Lopes RS, Mendonça ACDF, . . . Siqueira MM. 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.** 13(5):724, abr. 2021.

64. Slavov SN, Giovanetti M, Dos Santos Bezerra R, Fonseca V, Santos EV, Rodrigues ES, . . . Kashima S. Molecular surveillance of the on-going SARS-COV-2 epidemic in Ribeirao Preto City, Brazil. **Infect Genet Evol.** 93:104976, jun. 2021.

65. Claro IM, da Silva Sales FC, Ramundo MS, Candido DS, Silva CAM, de Jesus JG, . . . Levi JE. Local Transmission of SARS-CoV-2 Lineage B.1.1.7, Brazil, December 2020. **Emerg Infect Dis.** 27(3):970-972, 2021.

66. Dos Santos CA, Bezerra GVB, de Azevedo Marinho ARRA, Sena LOC, Alves JC, de Souza MSF, . . . Martins-Filho PR. First Report of SARS-CoV-2 B.1.1.251 lineage in Brazil. **J Travel Med.** 28(4):taab033, jun. 2021.

67. Pereira F, Tosta S, Lima MM, da Silva LRO, Nardy VB, Gómez MKA, . . . Leal A. Genomic surveillance activities unveil the introduction of the SARS-CoV-2 B.1.525 variant of interest in Brazil: Case report. **J Med Virol.** 93(9):5523-5526. set. 2021.

68. Lamarca AP, de Almeida LGP, da Silva Francisco Junior R, Cavalcante L, Machado DT, Brustolini O, . . . Vasconcelos ATR. Genomic Surveillance Tracks the First Community Outbreak of the SARS-CoV-2 Delta (B.1.617.2) Variant in Brazil. **J Virol.** 3:JVI0122821, nov. 2021.

69. Sintchenko V, Iredell JR, Gilbert GL. Pathogen profiling for disease management and surveillance. **Nat Rev Microbiol.** 5(6):464-70, jun. 2020.

70. Chan, J. M., & Rabadan, R. Quantifying pathogen surveillance using temporal genomic data. **mBio.** 4(1), e00524-12, 2013.

71. Hu T, Li J, Zhou H, Li C, Holmes EC, Shi W. Bioinformatics resources for SARS-CoV-2 discovery and surveillance. **Brief Bioinform.** bbaa386, jan. 2021.

72. Lo, S.W., Jamrozy, D. Genomics and epidemiological surveillance. **Nat Rev Microbiol** 18, 478, 2020.

73. Departamento De Edificações e Estradas De Rodagem De Minas Gerais - DER/MG. Rodovias Estaduais-MG. 2021. Disponível em: <www.der.mg.gov.br/>. Acesso em: 11 de janeiro de 2021.

74. Departamento Nacional De Infraestrutura De Transportes – DNIT. Rodovias Federais. 2021. Disponível em: <www.gov.br/dnit/pt-br/rodovias/rodovias-federais>. Acesso em: 11 de janeiro de 2021.

75. Instituto Brasileiro de Geografia e Estatística (IBGE), 2020. Disponível em: <https://www.ibge.gov.br/cidades-e-estados/mg/betim.html>. Acesso em: 11 de janeiro de 2021.

76. Centers for Disease Control and Prevention (CDC). Research Use Only 2019-Novel Coronavirus (2019-nCoV) Real-time RT-PCR Primers and Probes. 2020. Disponível

em:<https://www.cdc.gov/coronavirus/2019-ncov/lab/rt-pcr-panel-primer-probes.html>. Acesso em: 09/09/2021.

77. Artic Network. PrimalSeq Sequencing Primers v3, Artic 2019-nCoV, mar. 2020. Disponível em: <https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3>. Acesso em: 11/09/2021.

78. Bolger AM, Lohse M, Usadel B, Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**. 30(15):2114–2120, 2014.

79. Langmead B & Salzberg S. Fast gapped-read alignment with Bowtie 2. **Nat Methods.** p.357–359, 2012.

80. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, . . . Durbin R. 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. **Bioinformatics.** 25(16):2078–2079, 2009.

81. Quinlan AR & Hall IM, BEDTools: a flexible suite of utilities for comparing genomic features, **Bioinformatics.** 26(6):841–842, 2010.

82. Shu, Y., McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. **EuroSurveillance.** 22(13), 2017.

83. Katoh K & Standley DM, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. **Molecular Biology and Evolution.** 30(4):772–780, 2013.

84. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. **Molecular Biology and Evolution.** 37(5):1530–1534, 2020.

85. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. American Mathematical Society: **Lectures on Mathematics in the Life Sciences.** v.17:57–86, 1986.

86. Guindon S, Dufayard S-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. **Systematic Biology.** 59(3):307–321, 2010.

87. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2021. Disponível em: <https://www.R-project.org/>. Acesso em 11/09/2021.

88. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, et al. Software for Computing and Annotating Genomic Ranges. **PLoS Comput Biol.** 9(8): e1003118, 2013.

89. Ou J & Zhu L. trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. **Nature Methods.** 16, 453–454, 2019.

90. Gräler B, Pebesma E, Heuvelink G. Spatio-Temporal Interpolation using gstat. **The R Journal.** 8(1), 204-218, 2016.

91. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. **Virus Evol.** 4(1), vey016, 2018.

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

93. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. **J Mol Evol.** 39:306–314, 1994.

94. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci, **Molecular Biology and Evolution**. 30(3):713–724, 2013.

95. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. **PLoS Comput Biol** 5(9): e1000520, 2009.

96. Rambaut A, Drummond AJ, Xie D, Baele G and Suchard MA. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. **Systematic Biology.** syy032, 2018.

97. IBGE. Instituto Brasileiro de Geografia e Estatística. Coleção de Mapas Municipais. Betim-MG, 2020. Disponível em: https://geoftp.ibge.gov.br/cartas_e_mapas/mapas_municipais/colecao_de_mapas_municipais/2020/MG/betim/3106705_MM.pdf. Acesso em: 10/10/2021.

98. Assessoria de Imprensa UFMG. UFMG vai detectar comportamento e alcance do coronavírus em Betim, 2020. Disponível em: https://ufmg.br/comunicacao/assessoria-de-imprensa/release/ufmg-vai-detectar-comportamento-e-alcance-do-coronavirus-em-betim. Acesso em: 23/11/2021.

99. Bull RA, Adikari TN, Ferguson JM, Hammond JM, Stevanovski I, Beukers AG, . . . Deveson IW. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. **Nat Commun.** 11(1):6272, dec. 2020.

100. Huang H, Knowles LL. Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. **Syst Biol.** 65(3):357-65, 2016.

101. da Silva Filipe A, Shepherd JG, Williams T, Hughes J, Aranday-Cortes E, Asamaphan P, . . . Thomson EC. Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. **Nat Microbiol.** 6(1):112-122, jan. 2021.

102. Moody A, Sellers S, Bumstead N. Measuring infectious bursal disease virus RNA in blood by multiplex real-time quantitative RT-PCR. **J Virol Methods.** 85(1-2):55-64, mar. 2000.

103. Jacot D, Pillonel T, Greub G, Bertelli C. Assessment of SARS-CoV-2 Genome Sequencing: Quality Criteria and Low-Frequency Variants. **J Clin Microbiol.** 59(10):e0094421, set. 2021.

104. Assessoria de imprensa UFMG. UFMG divulga balanço de caracterização de variantes do coronavírus em BH, 2021. Disponível em: https://ufmg.br/comunicacao/noticias/ufmg-divulga-balanco-de-caracterizacao-de-variantes-do-coronavirus-neste-ano-em-bh. Acesso em 23/11/2021.

105. MCTI. Rede Corona-ômica-MCTI conclui estudo que analisou a estimativa da frequência de variantes de SARS-CoV-2 no estado de Minas Gerais, 2021. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/noticias/2021/06/rede-corona-omica-mcti-conclui-estudo-que-analisou-a-estimativa-da-frequencia-de-variantes-de-sars-cov-2-no-estado-de-minas-gerais. Acesso em 23/11/2021.

106. Ferraz, A. Proportion of Delta variant samples triples in Brazil in one month. **The Brazilian Report,** 2021. Disponível em: https://brazilian.report/liveblog/2021/08/11/delta-variant-samples-triples/. Acesso em 24/11/2021.

107. Gilson Silva AV, Menezes D, Moreira FR, Torres OA, Fonseca PL, Moreira RG, . . . de Souza RP. Seroprevalence, prevalence, and genomic surveillance: monitoring the initial phases of the SARS-CoV-2 pandemic in Betim, Brazil. **medRxiv, (),** 2021.10.21.21265140, 2021.

8. ANEXOS

Anexo 1. Lista de iniciadores utilizados na amplificação do genoma total de SARS-CoV-2. Inclui a identificação do iniciador, primeiro (*pool* 1) ou segundo (*pool* 2) conjunto de iniciadores sequenciais, sequência e tamanho (adaptado de Artic Network, 2020)[77].

ID	*Pool*	Sequência	Tamanho (pb)
nCoV-2019_1_ESQUERDO	1	ACCAACCAACTTTCGATCTCTTGT	24
nCoV-2019_1_DIREITO	1	CATCTTTAAGATGTTGACGTGCCTC	25
nCoV-2019_2_ESQUERDO	2	CTGTTTTACAGGTTCGCGACGT	22
nCoV-2019_2_DIREITO	2	TAAGGATCAGTGCCAAGCTCGT	22
nCoV-2019_3_ESQUERDO	1	CGGTAATAAAGGAGCTGGTGGC	22
nCoV-2019_3_DIREITO	1	AAGGTGTCTGCAATTCATAGCTCT	24
nCoV-2019_4_ESQUERDO	2	GGTGTATACTGCTGCCGTGAAC	22
nCoV-2019_4_DIREITO	2	CACAAGTAGTGGCACCTTCTTTAGT	25
nCoV-2019_5_ESQUERDO	1	TGGTGAAACTTCATGGCAGACG	22
nCoV-2019_5_DIREITO	1	ATTGATGTTGACTTTCTCTTTTTGGAGT	28
nCoV-2019_6_ESQUERDO	2	GGTGTTGTTGGAGAAGGTTCCG	22
nCoV-2019_6_DIREITO	2	TAGCGGCCTTCTGTAAAACACG	22
nCoV-2019_7_ESQUERDO	1	ATCAGAGGCTGCTCGTGTTGTA	22
nCoV-2019_7_ESQUERDO_alt0	1	CATTTGCATCAGAGGCTGCTCG	22
nCoV-2019_7_DIREITO	1	TGCACAGGTGACAATTTGTCCA	22
nCoV-2019_7_DIREITO_alt5	1	AGGTGACAATTTGTCCACCGAC	22
nCoV-2019_8_ESQUERDO	2	AGAGTTTCTTAGAGACGGTTGGGA	24
nCoV-2019_8_DIREITO	2	GCTTCAACAGCTTCACTAGTAGGT	24
nCoV-2019_9_ESQUERDO	1	TCCCACAGAAGTGTTAACAGAGGA	24
nCoV-2019_9_ESQUERDO_alt4	1	TTCCCACAGAAGTGTTAACAGAGG	24
nCoV-2019_9_DIREITO	1	ATGACAGCATCTGCCACAACAC	22
nCoV-2019_9_DIREITO_alt2	1	GACAGCATCTGCCACAACACAG	22
nCoV-2019_10_ESQUERDO	2	TGAGAAGTGCTCTGCCTATACAGT	24
nCoV-2019_10_DIREITO	2	TCATCTAACCAATCTTCTTCTTGCTCT	27
nCoV-2019_11_ESQUERDO	1	GGAATTTGGTGCCACTTCTGCT	22
nCoV-2019_11_DIREITO	1	TCATCAGATTCAACTTGCATGGCA	24
nCoV-2019_12_ESQUERDO	2	AAACATGGAGGAGGTGTTGCAG	22
nCoV-2019_12_DIREITO	2	TTCACTCTTCATTTCCAAAAAGCTTGA	27
nCoV-2019_13_ESQUERDO	1	TCGCACAAATGTCTACTTAGCTGT	24
nCoV-2019_13_DIREITO	1	ACCACAGCAGTTAAAACACCCT	22

nCoV-2019_14_ESQUERDO	2	CATCCAGATTCTGCCACTCTTGT	23
nCoV-2019_14_ESQUERDO_alt4	2	TGGCAATCTTCATCCAGATTCTGC	24
nCoV-2019_14_DIREITO	2	AGTTTCCACACAGACAGGCATT	22
nCoV-2019_14_DIREITO_alt2	2	TGCGTGTTTCTTCTGCATGTGC	22
nCoV-2019_15_ESQUERDO	1	ACAGTGCTTAAAAAGTGTAAAAGTGCC	27
nCoV-2019_15_ESQUERDO_alt1	1	AGTGCTTAAAAAGTGTAAAAGTGCCT	26
nCoV-2019_15_DIREITO	1	AACAGAAACTGTAGCTGGCACT	22
nCoV-2019_15_DIREITO_alt3	1	ACTGTAGCTGGCACTTTGAGAGA	23
nCoV-2019_16_ESQUERDO	2	AATTTGGAAGAAGCTGCTCGGT	22
nCoV-2019_16_DIREITO	2	CACAACTTGCGTGTGGAGGTTA	22
nCoV-2019_17_ESQUERDO	1	CTTCTTTCTTTGAGAGAAGTGAGGACT	27
nCoV-2019_17_DIREITO	1	TTTGTTGGAGTGTTAACAATGCAGT	25
nCoV-2019_18_ESQUERDO	2	TGGAAATACCCACAAGTTAATGGTTTAAC	29
nCoV-2019_18_ESQUERDO_alt2	2	ACTTCTATTAAATGGGCAGATAACAACTGT	30
nCoV-2019_18_DIREITO	2	AGCTTGTTTACCACACGTACAAGG	24
nCoV-2019_18_DIREITO_alt1	2	GCTTGTTTACCACACGTACAAGG	23
nCoV-2019_19_ESQUERDO	1	GCTGTTATGTACATGGGCACACT	23
nCoV-2019_19_DIREITO	1	TGTCCAACTTAGGGTCAATTTCTGT	25
nCoV-2019_20_ESQUERDO	2	ACAAAGAAAACAGTTACACAACAACCA	27
nCoV-2019_20_DIREITO	2	ACGTGGCTTTATTAGTTGCATTGTT	25
nCoV-2019_21_ESQUERDO	1	TGGCTATTGATTATAAACACTACACACCC	29
nCoV-2019_21_ESQUERDO_alt2	1	GGCTATTGATTATAAACACTACACACCCT	29
nCoV-2019_21_DIREITO	1	TAGATCTGTGTGGCCAACCTCT	22
nCoV-2019_21_DIREITO_alt0	1	GATCTGTGTGGCCAACCTCTTC	22
nCoV-2019_22_ESQUERDO	2	ACTACCGAAGTTGTAGGAGACATTATACT	29
nCoV-2019_22_DIREITO	2	ACAGTATTCTTTGCTATAGTAGTCGGC	27

nCoV-2019_23_ESQUERDO	1	ACAACTACTAACATAGTTACACGGTGT	27
nCoV-2019_23_DIREITO	1	ACCAGTACAGTAGGTTGCAATAGTG	25
nCoV-2019_24_ESQUERDO	2	AGGCATGCCTTCTTACTGTACTG	23
nCoV-2019_24_DIREITO	2	ACATTCTAACCATAGCTGAAATCGGG	26
nCoV-2019_25_ESQUERDO	1	GCAATTGTTTTTCAGCTATTTTGCAGT	27
nCoV-2019_25_DIREITO	1	ACTGTAGTGACAAGTCTCTCGCA	23
nCoV-2019_26_ESQUERDO	2	TTGTGATACATTCTGTGCTGGTAGT	25
nCoV-2019_26_DIREITO	2	TCCGCACTATCACCAACATCAG	22
nCoV-2019_27_ESQUERDO	1	ACTACAGTCAGCTTATGTGTCAACC	25
nCoV-2019_27_DIREITO	1	AATACAAGCACCAAGGTCACGG	22
nCoV-2019_28_ESQUERDO	2	ACATAGAAGTTACTGGCGATAGTTGT	26
nCoV-2019_28_DIREITO	2	TGTTTAGACATGACATGAACAGGTGT	26
nCoV-2019_29_ESQUERDO	1	ACTTGTGTTCCTTTTTGTTGCTGC	24
nCoV-2019_29_DIREITO	1	AGTGTACTCTATAAGTTTTGATGGTGTG T	29
nCoV-2019_30_ESQUERDO	2	GCACAACTAATGGTGACTTTTTGCA	25
nCoV-2019_30_DIREITO	2	ACCACTAGTAGATACACAAACACCAG	26
nCoV-2019_31_ESQUERDO	1	TTCTGAGTACTGTAGGCACGGC	22
nCoV-2019_31_DIREITO	1	ACAGAATAAACACCAGGTAAGAATGAGT	28
nCoV-2019_32_ESQUERDO	2	TGGTGAATACAGTCATGTAGTTGCC	25
nCoV-2019_32_DIREITO	2	AGCACATCACTACGCAACTTTAGA	24
nCoV-2019_33_ESQUERDO	1	ACTTTTGAAGAAGCTGCGCTGT	22
nCoV-2019_33_DIREITO	1	TGGACAGTAAACTACGTCATCAAGC	25
nCoV-2019_34_ESQUERDO	2	TCCCATCTGGTAAAGTTGAGGGT	23
nCoV-2019_34_DIREITO	2	AGTGAAATTGGGCCTCATAGCA	22
nCoV-2019_35_ESQUERDO	1	TGTTCGCATTCAACCAGGACAG	22
nCoV-2019_35_DIREITO	1	ACTTCATAGCCACAAGGTTAAAGTCA	26
nCoV-2019_36_ESQUERDO	2	TTAGCTTGGTTGTACGCTGCTG	22
nCoV-2019_36_DIREITO	2	GAACAAAGACCATTGAGTACTCTGGA	26
nCoV-2019_37_ESQUERDO	1	ACACACCACTGGTTGTTACTCAC	23
nCoV-2019_37_DIREITO	1	GTCCACACTCTCCTAGCACCAT	22
nCoV-2019_38_ESQUERDO	2	ACTGTGTTATGTATGCATCAGCTGT	25
nCoV-2019_38_DIREITO	2	CACCAAGAGTCAGTCTAAAGTAGCG	25
nCoV-2019_39_ESQUERDO	1	AGTATTGCCCTATTTTCTTCATAACTGGT	29

nCoV-2019_39_DIREITO	1	TGTAACTGGACACATTGAGCCC	22
nCoV-2019_40_ESQUERDO	2	TGCACATCAGTAGTCTTACTCTCAGT	26
nCoV-2019_40_DIREITO	2	CATGGCTGCATCACGGTCAAAT	22
nCoV-2019_41_ESQUERDO	1	GTTCCCTTCCATCATATGCAGCT	23
nCoV-2019_41_DIREITO	1	TGGTATGACAACCATTAGTTTGGCT	25
nCoV-2019_42_ESQUERDO	2	TGCAAGAGATGGTTGTGTTCCC	22
nCoV-2019_42_DIREITO	2	CCTACCTCCCTTTGTTGTGTTGT	23
nCoV-2019_43_ESQUERDO	1	TACGACAGATGTCTTGTGCTGC	22
nCoV-2019_43_DIREITO	1	AGCAGCATCTACAGCAAAAGCA	22
nCoV-2019_44_ESQUERDO	2	TGCCACAGTACGTCTACAAGCT	22
nCoV-2019_44_ESQUERDO_alt3	2	CCACAGTACGTCTACAAGCTGG	22
nCoV-2019_44_DIREITO	2	AACCTTTCCACATACCGCAGAC	22
nCoV-2019_44_DIREITO_alt0	2	CGCAGACGGTACAGACTGTGTT	22
nCoV-2019_45_ESQUERDO	1	TACCTACAACTTGTGCTAATGACCC	25
nCoV-2019_45_ESQUERDO_alt2	1	AGTATGTACAAATACCTACAACTTGTGCT	29
nCoV-2019_45_DIREITO	1	AAATTGTTTCTTCATGTTGGTAGTTAGAGA	30
nCoV-2019_45_DIREITO_alt7	1	TTCATGTTGGTAGTTAGAGAAAGTGTGTC	29
nCoV-2019_46_ESQUERDO	2	TGTCGCTTCCAAGAAAAGGACG	22
nCoV-2019_46_ESQUERDO_alt1	2	CGCTTCCAAGAAAAGGACGAAGA	23
nCoV-2019_46_DIREITO	2	CACGTTCACCTAAGTTGGCGTA	22
nCoV-2019_46_DIREITO_alt2	2	CACGTTCACCTAAGTTGGCGTAT	23
nCoV-2019_47_ESQUERDO	1	AGGACTGGTATGATTTTGTAGAAAACCC	28
nCoV-2019_47_DIREITO	1	AATAACGGTCAAAGAGTTTTAACCTCTC	28
nCoV-2019_48_ESQUERDO	2	TGTTGACACTGACTTAACAAAGCCT	25
nCoV-2019_48_DIREITO	2	TAGATTACCAGAAGCAGCGTGC	22
nCoV-2019_49_ESQUERDO	1	AGGAATTACTTGTGTATGCTGCTGA	25
nCoV-2019_49_DIREITO	1	TGACGATGACTTGGTTAGCATTAATACA	28
nCoV-2019_50_ESQUERDO	2	GTTGATAAGTACTTTGATTGTTACGATGGT	30

nCoV-2019_50_DIREITO	2	TAACATGTTGTGCCAACCACCA	22
nCoV-2019_51_ESQUERDO	1	TCAATAGCCGCCACTAGAGGAG	22
nCoV-2019_51_DIREITO	1	AGTGCATTAACATTGGCCGTGA	22
nCoV-2019_52_ESQUERDO	2	CATCAGGAGATGCCACAACTGC	22
nCoV-2019_52_DIREITO	2	GTTGAGAGCAAAATTCATGAGGTCC	25
nCoV-2019_53_ESQUERDO	1	AGCAAAATGTTGGACTGAGACTGA	24
nCoV-2019_53_DIREITO	1	AGCCTCATAAAACTCAGGTTCCC	23
nCoV-2019_54_ESQUERDO	2	TGAGTTAACAGGACACATGTTAGACA	26
nCoV-2019_54_DIREITO	2	AACCAAAAACTTGTCCATTAGCACA	25
nCoV-2019_55_ESQUERDO	1	ACTCAACTTTACTTAGGAGGTATGAGCT	28
nCoV-2019_55_DIREITO	1	GGTGTACTCTCCTATTTGTACTTTACTGT	29
nCoV-2019_56_ESQUERDO	2	ACCTAGACCACCACTTAACCGA	22
nCoV-2019_56_DIREITO	2	ACACTATGCGAGCAGAAGGGTA	22
nCoV-2019_57_ESQUERDO	1	ATTCTACACTCCAGGGACCACC	22
nCoV-2019_57_DIREITO	1	GTAATTGAGCAGGGTCGCCAAT	22
nCoV-2019_58_ESQUERDO	2	TGATTTGAGTGTTGTCAATGCCAGA	25
nCoV-2019_58_DIREITO	2	CTTTTCTCCAAGCAGGGTTACGT	23
nCoV-2019_59_ESQUERDO	1	TCACGCATGATGTTTCATCTGCA	23
nCoV-2019_59_DIREITO	1	AAGAGTCCTGTTACATTTTCAGCTTG	26
nCoV-2019_60_ESQUERDO	2	TGATAGAGACCTTTATGACAAGTTGCA	27
nCoV-2019_60_DIREITO	2	GGTACCAACAGCTTCTCTAGTAGC	24
nCoV-2019_61_ESQUERDO	1	TGTTTATCACCCGCGAAGAAGC	22
nCoV-2019_61_DIREITO	1	ATCACATAGACAACAGGTGCGC	22
nCoV-2019_62_ESQUERDO	2	GGCACATGGCTTTGAGTTGACA	22
nCoV-2019_62_DIREITO	2	GTTGAACCTTTCTACAAGCCGC	22
nCoV-2019_63_ESQUERDO	1	TGTTAAGCGTGTTGACTGGACT	22
nCoV-2019_63_DIREITO	1	ACAAACTGCCACCATCACAACC	22
nCoV-2019_64_ESQUERDO	2	TCGATAGATATCCTGCTAATTCCATTGT	28
nCoV-2019_64_DIREITO	2	AGTCTTGTAAAAGTGTTCCAGAGGT	25
nCoV-2019_65_ESQUERDO	1	GCTGGCTTTAGCTTGTGGGTTT	22
nCoV-2019_65_DIREITO	1	TGTCAGTCATAGAACAAACACCAATAGT	28
nCoV-2019_66_ESQUERDO	2	GGGTGTGGACATTGCTGCTAAT	22
nCoV-2019_66_DIREITO	2	TCAATTTCCATTTGACTCCTGGGT	24

nCoV-2019_67_ESQUERDO	1	GTTGTCCAACAATTACCTGAAACTTACT	28
nCoV-2019_67_DIREITO	1	CAACCTTAGAAACTACAGATAAATCTTGGG	30
nCoV-2019_68_ESQUERDO	2	ACAGGTTCATCTAAGTGTGTGTGT	24
nCoV-2019_68_DIREITO	2	CTCCTTTATCAGAACCAGCACCA	23
nCoV-2019_69_ESQUERDO	1	TGTCGCAAAATATACTCAACTGTGTCA	27
nCoV-2019_69_DIREITO	1	TCTTTATAGCCACGGAACCTCCA	23
nCoV-2019_70_ESQUERDO	2	ACAAAAGAAAATGACTCTAAAGAGGGTTT	29
nCoV-2019_70_DIREITO	2	TGACCTTCTTTTAAAGACATAACAGCAG	28
nCoV-2019_71_ESQUERDO	1	ACAAATCCAATTCAGTTGTCTTCCTATTC	29
nCoV-2019_71_DIREITO	1	TGGAAAAGAAAGGTAAGAACAAGTCCT	27
nCoV-2019_72_ESQUERDO	2	ACACGTGGTGTTTATTACCCTGAC	24
nCoV-2019_72_DIREITO	2	ACTCTGAACTCACTTTCCATCCAAC	25
nCoV-2019_73_ESQUERDO	1	CAATTTTGTAATGATCCATTTTTGGGTGT	29
nCoV-2019_73_DIREITO	1	CACCAGCTGTCCAACCTGAAGA	22
nCoV-2019_74_ESQUERDO	2	ACATCACTAGGTTTCAAACTTTACTTGC	28
nCoV-2019_74_DIREITO	2	GCAACACAGTTGCTGATTCTCTTC	24
nCoV-2019_75_ESQUERDO	1	AGAGTCCAACCAACAGAATCTATTGT	26
nCoV-2019_75_DIREITO	1	ACCACCAACCTTAGAATCAAGATTGT	26
nCoV-2019_76_ESQUERDO	2	AGGGCAAACTGGAAAGATTGCT	22
nCoV-2019_76_ESQUERDO_alt3	2	GGGCAAACTGGAAAGATTGCTGA	23
nCoV-2019_76_DIREITO	2	ACACCTGTGCCTGTTAAACCAT	22
nCoV-2019_76_DIREITO_alt0	2	ACCTGTGCCTGTTAAACCATTGA	23
nCoV-2019_77_ESQUERDO	1	CCAGCAACTGTTTGTGGACCTA	22
nCoV-2019_77_DIREITO	1	CAGCCCCTATTAAACAGCCTGC	22
nCoV-2019_78_ESQUERDO	2	CAACTTACTCCTACTTGGCGTGT	23
nCoV-2019_78_DIREITO	2	TGTGTACAAAAACTGCCATATTGCA	25
nCoV-2019_79_ESQUERDO	1	GTGGTGATTCAACTGAATGCAGC	23
nCoV-2019_79_DIREITO	1	CATTTCATCTGTGAGCAAAGGTGG	24
nCoV-2019_80_ESQUERDO	2	TTGCCTTGGTGATATTGCTGCT	22
nCoV-2019_80_DIREITO	2	TGGAGCTAAGTTGTTTAACAAGCG	24
nCoV-2019_81_ESQUERDO	1	GCACTTGGAAAACTTCAAGATGTGG	25

nCoV-2019_81_DIREITO	1	GTGAAGTTCTTTTCTTGTGCAGGG	24
nCoV-2019_82_ESQUERDO	2	GGGCTATCATCTTATGTCCTTCCCT	25
nCoV-2019_82_DIREITO	2	TGCCAGAGATGTCACCTAAATCAA	24
nCoV-2019_83_ESQUERDO	1	TCCTTTGCAACCTGAATTAGACTCA	25
nCoV-2019_83_DIREITO	1	TTTGACTCCTTTGAGCACTGGC	22
nCoV-2019_84_ESQUERDO	2	TGCTGTAGTTGTCTCAAGGGCT	22
nCoV-2019_84_DIREITO	2	AGGTGTGAGTAAACTGTTACAAACAAC	27
nCoV-2019_85_ESQUERDO	1	ACTAGCACTCTCCAAGGGTGTT	22
nCoV-2019_85_DIREITO	1	ACACAGTCTTTTACTCCAGATTCCC	25
nCoV-2019_86_ESQUERDO	2	TCAGGTGATGGCACAACAAGTC	22
nCoV-2019_86_DIREITO	2	ACGAAAGCAAGAAAAAGAAGTACGC	25
nCoV-2019_87_ESQUERDO	1	CGACTACTAGCGTGCCTTTGTA	22
nCoV-2019_87_DIREITO	1	ACTAGGTTCCATTGTTCAAGGAGC	24
nCoV-2019_88_ESQUERDO	2	CCATGGCAGATTCCAACGGTAC	22
nCoV-2019_88_DIREITO	2	TGGTCAGAATAGTGCCATGGAGT	23
nCoV-2019_89_ESQUERDO	1	GTACGCGTTCCATGTGGTCATT	22
nCoV-2019_89_ESQUERDO_alt2	1	CGCGTTCCATGTGGTCATTCAA	22
nCoV-2019_89_DIREITO	1	ACCTGAAAGTCAACGAGATGAAACA	25
nCoV-2019_89_DIREITO_alt4	1	ACGAGATGAAACATCTGTTGTCACT	25
nCoV-2019_90_ESQUERDO	2	ACACAGACCATTCCAGTAGCAGT	23
nCoV-2019_90_DIREITO	2	TGAAATGGTGAATTGCCCTCGT	22
nCoV-2019_91_ESQUERDO	1	TCACTACCAAGAGTGTGTTAGAGGT	25
nCoV-2019_91_DIREITO	1	TTCAAGTGAGAACCAAAAGATAATAAGCA	29
nCoV-2019_92_ESQUERDO	2	TTTGTGCTTTTTAGCCTTTCTGCT	24
nCoV-2019_92_DIREITO	2	AGGTTCCTGGCAATTAATTGTAAAAGG	27
nCoV-2019_93_ESQUERDO	1	TGAGGCTGGTTCTAAATCACCCA	23
nCoV-2019_93_DIREITO	1	AGGTCTTCCTTGCCATGTTGAG	22
nCoV-2019_94_ESQUERDO	2	GGCCCCAAGGTTTACCCAATAA	22
nCoV-2019_94_DIREITO	2	TTTGGCAATGTTGTTCCTTGAGG	23
nCoV-2019_95_ESQUERDO	1	TGAGGGAGCCTTGAATACACCA	22
nCoV-2019_95_DIREITO	1	CAGTACGTTTTTGCCGAGGCTT	22

nCoV-2019_96_ESQUERDO	2	GCCAACAACAACAAGGCCAAAC	22
nCoV-2019_96_DIREITO	2	TAGGCTCTGTTGGTGGGAATGT	22
nCoV-2019_97_ESQUERDO	1	TGGATGACAAAGATCCAAATTTCAAAGA	28
nCoV-2019_97_DIREITO	1	ACACACTGATTAAAGATTGCTATGTGAG	28
nCoV-2019_98_ESQUERDO	2	AACAATTGCAACAATCCATGAGCA	24
nCoV-2019_98_DIREITO	2	TTCTCCTAAGAAGCTATTAAAATCACATGG	30

Anexo 2. Sequências utilizadas nas reconstruções filogeográficas. Inclui as sequências Betim-MG obtidas no estudo além das sequências amostradas a partir do Banco de dados GISAID[80].

Linhagem	IDs GISAID
B.1.1.28	EPI_ISL_5416106; EPI_ISL_5416105; EPI_ISL_5416108; EPI_ISL_5416101; EPI_ISL_5416109; EPI_ISL_5416100; EPI_ISL_5416088; EPI_ISL_5416121; EPI_ISL_5416117; EPI_ISL_5416116; EPI_ISL_5416119; EPI_ISL_5416112; EPI_ISL_5416115; EPI_ISL_5416114; EPI_ISL_5416092; EPI_ISL_5416093; EPI_ISL_5416095; EPI_ISL_1182550; EPI_ISL_904018; EPI_ISL_416036; EPI_ISL_431180; EPI_ISL_875541; EPI_ISL_875544; EPI_ISL_875545; EPI_ISL_875550; EPI_ISL_888671; EPI_ISL_1139067; EPI_ISL_3048755; EPI_ISL_431240; EPI_ISL_445380; EPI_ISL_470653; EPI_ISL_470654; EPI_ISL_476259; EPI_ISL_541344; EPI_ISL_541372; EPI_ISL_690818; EPI_ISL_693257; EPI_ISL_693278; EPI_ISL_708745; EPI_ISL_848624; EPI_ISL_848628; EPI_ISL_849680; EPI_ISL_861636; EPI_ISL_861879; EPI_ISL_861890; EPI_ISL_861900; EPI_ISL_861902; EPI_ISL_861906; EPI_ISL_861912; EPI_ISL_861913; EPI_ISL_978518; EPI_ISL_1068210; EPI_ISL_1084725; EPI_ISL_1139060; EPI_ISL_1172014; EPI_ISL_1511641; EPI_ISL_3243099; EPI_ISL_579220; EPI_ISL_581703; EPI_ISL_685539; EPI_ISL_722131; EPI_ISL_735399; EPI_ISL_848622; EPI_ISL_848623; EPI_ISL_861634; EPI_ISL_861896; EPI_ISL_861905; EPI_ISL_861914; EPI_ISL_1068082; EPI_ISL_1068394; EPI_ISL_1084723; EPI_ISL_1084724; EPI_ISL_1084727; EPI_ISL_1084728; EPI_ISL_1171871; EPI_ISL_2645638; EPI_ISL_3048762; EPI_ISL_493851; EPI_ISL_509500; EPI_ISL_511103; EPI_ISL_541386; EPI_ISL_792643; EPI_ISL_848620; EPI_ISL_848621; EPI_ISL_861867; EPI_ISL_906065; EPI_ISL_906066; EPI_ISL_1068378; EPI_ISL_1139056; EPI_ISL_1139057; EPI_ISL_1340764; EPI_ISL_3243101; EPI_ISL_456869; EPI_ISL_473651; EPI_ISL_493852; EPI_ISL_493868; EPI_ISL_572379; EPI_ISL_801403; EPI_ISL_848615; EPI_ISL_861655; EPI_ISL_861656; EPI_ISL_861659; EPI_ISL_861661; EPI_ISL_875543; EPI_ISL_875548; EPI_ISL_906067; EPI_ISL_1068242; EPI_ISL_1068244; EPI_ISL_1068377; EPI_ISL_1139054; EPI_ISL_1469808; EPI_ISL_1714591; EPI_ISL_1714616; EPI_ISL_1714620; EPI_ISL_2645637; EPI_ISL_2677096; EPI_ISL_2821285; EPI_ISL_478094; EPI_ISL_478111; EPI_ISL_591372; EPI_ISL_848571; EPI_ISL_861647; EPI_ISL_861648; EPI_ISL_861649; EPI_ISL_861657; EPI_ISL_861658; EPI_ISL_861660; EPI_ISL_861666; EPI_ISL_1068246; EPI_ISL_1068247; EPI_ISL_1117408; EPI_ISL_1300937; EPI_ISL_1511644; EPI_ISL_1583668; EPI_ISL_1712905; EPI_ISL_1712912; EPI_ISL_3048772; EPI_ISL_3243104; EPI_ISL_3243105; EPI_ISL_478249; EPI_ISL_490026; EPI_ISL_591537; EPI_ISL_591538; EPI_ISL_603023; EPI_ISL_667639; EPI_ISL_678320; EPI_ISL_693230; EPI_ISL_693233; EPI_ISL_735412; EPI_ISL_802105; EPI_ISL_802106; EPI_ISL_848611; EPI_ISL_850198; EPI_ISL_1063789; EPI_ISL_1068250; EPI_ISL_1139052; EPI_ISL_1340761; EPI_ISL_1469823; EPI_ISL_1469841; EPI_ISL_2821287; EPI_ISL_3243107; EPI_ISL_3316170; EPI_ISL_508266; EPI_ISL_551467; EPI_ISL_672205; EPI_ISL_690635; EPI_ISL_735429; EPI_ISL_735433; EPI_ISL_776765; EPI_ISL_792115; EPI_ISL_831660; EPI_ISL_861667; EPI_ISL_940608; EPI_ISL_943988; EPI_ISL_1068205; EPI_ISL_1117429; EPI_ISL_1121329; EPI_ISL_1340757; EPI_ISL_1469851; EPI_ISL_1714483; EPI_ISL_1714533; EPI_ISL_2161051; EPI_ISL_2466186; EPI_ISL_2466188; EPI_ISL_2466189; EPI_ISL_2497434; EPI_ISL_2758766; EPI_ISL_3266121; EPI_ISL_522491; EPI_ISL_527032; EPI_ISL_593687; EPI_ISL_593698; EPI_ISL_667561; EPI_ISL_667626; EPI_ISL_735428; EPI_ISL_735431; EPI_ISL_750176; EPI_ISL_792116; EPI_ISL_831688; EPI_ISL_833156; EPI_ISL_848565; EPI_ISL_875540; EPI_ISL_888672; EPI_ISL_1040847; EPI_ISL_1040849; EPI_ISL_1040850; EPI_ISL_1068209; EPI_ISL_1068254; EPI_ISL_1799499; EPI_ISL_2223909; EPI_ISL_2677308; EPI_ISL_3046161; EPI_ISL_3243110; EPI_ISL_3243111; EPI_ISL_3266085; EPI_ISL_547433; EPI_ISL_547434; EPI_ISL_547435; EPI_ISL_547437; EPI_ISL_593711; EPI_ISL_603037; EPI_ISL_603038; EPI_ISL_667634; EPI_ISL_667649; EPI_ISL_667650; EPI_ISL_693214; EPI_ISL_750175; EPI_ISL_750177; EPI_ISL_751185; EPI_ISL_751201; EPI_ISL_861875; EPI_ISL_861895; EPI_ISL_861901; EPI_ISL_884250; EPI_ISL_940924; EPI_ISL_961766; EPI_ISL_1181384; EPI_ISL_2603520; EPI_ISL_2731471; EPI_ISL_2731477; EPI_ISL_492036; EPI_ISL_513514; EPI_ISL_513532; EPI_ISL_513546; EPI_ISL_513557; EPI_ISL_513578; EPI_ISL_541354; EPI_ISL_541355; EPI_ISL_541359; EPI_ISL_623130; EPI_ISL_623131; EPI_ISL_717787; EPI_ISL_717794; EPI_ISL_717810; EPI_ISL_717811; EPI_ISL_717812; EPI_ISL_717813
B.1.1.33	EPI_ISL_5416107; EPI_ISL_5416102; EPI_ISL_5416104; EPI_ISL_5416103; EPI_ISL_5416087; EPI_ISL_5416120; EPI_ISL_5416089; EPI_ISL_5416118; EPI_ISL_5416113; EPI_ISL_5416090; EPI_ISL_5416091; EPI_ISL_5416098; EPI_ISL_5416097; EPI_ISL_5416111; EPI_ISL_5416099; EPI_ISL_5416110; EPI_ISL_5416094; EPI_ISL_5416096; EPI_ISL_2557345; EPI_ISL_470582; EPI_ISL_470583; EPI_ISL_470584; EPI_ISL_470585; EPI_ISL_470586; EPI_ISL_470587; EPI_ISL_470588; EPI_ISL_470589; EPI_ISL_470590; EPI_ISL_470591; EPI_ISL_470594; EPI_ISL_470595; EPI_ISL_470596; EPI_ISL_476194; EPI_ISL_623104; EPI_ISL_623105; EPI_ISL_672675; EPI_ISL_672679; EPI_ISL_904029; EPI_ISL_1739298; EPI_ISL_427294; EPI_ISL_427295; EPI_ISL_457796; EPI_ISL_465088; EPI_ISL_476221; EPI_ISL_486427; EPI_ISL_524798; EPI_ISL_541370; EPI_ISL_1181608; EPI_ISL_1181611; EPI_ISL_1181612; EPI_ISL_1181613; EPI_ISL_1181619; EPI_ISL_1201883; EPI_ISL_1396076; EPI_ISL_1482810; EPI_ISL_3105856; EPI_ISL_425373; EPI_ISL_427296; EPI_ISL_427297; EPI_ISL_427298; EPI_ISL_454057; EPI_ISL_454172; EPI_ISL_454311; EPI_ISL_457953; EPI_ISL_470613; EPI_ISL_476324; EPI_ISL_511042; EPI_ISL_511043; EPI_ISL_511190; EPI_ISL_511523; EPI_ISL_541376; EPI_ISL_636836; EPI_ISL_636838; EPI_ISL_639775; EPI_ISL_639787; EPI_ISL_755827; EPI_ISL_776754; EPI_ISL_776762; EPI_ISL_792105; EPI_ISL_801663; EPI_ISL_848598; EPI_ISL_848627; EPI_ISL_861635; EPI_ISL_861869; EPI_ISL_861893; EPI_ISL_861907; EPI_ISL_861915; EPI_ISL_978530; EPI_ISL_1139062; EPI_ISL_1181588; EPI_ISL_1181602; EPI_ISL_1181603; EPI_ISL_1181606; EPI_ISL_1181607; EPI_ISL_1678584; EPI_ISL_2196361; EPI_ISL_470570; EPI_ISL_470577; EPI_ISL_470611; EPI_ISL_470655; EPI_ISL_473841; EPI_ISL_524795; EPI_ISL_527017; EPI_ISL_538349; EPI_ISL_561935; EPI_ISL_578451; EPI_ISL_583762; EPI_ISL_593615; EPI_ISL_672751; EPI_ISL_690779; EPI_ISL_721992; EPI_ISL_722000; EPI_ISL_722030; EPI_ISL_722042; EPI_ISL_722043; EPI_ISL_722130; EPI_ISL_722136; EPI_ISL_722140; EPI_ISL_801696; EPI_ISL_1068392; EPI_ISL_1084729; EPI_ISL_1139063; EPI_ISL_1181592; EPI_ISL_1181595; EPI_ISL_1181622; EPI_ISL_1690551; EPI_ISL_1691459; EPI_ISL_1694627; EPI_ISL_1694628; EPI_ISL_1694631; EPI_ISL_1694632; EPI_ISL_458325; EPI_ISL_524794; EPI_ISL_591369; EPI_ISL_722041; EPI_ISL_722129; EPI_ISL_776759; EPI_ISL_792589; EPI_ISL_801853; EPI_ISL_833155; EPI_ISL_848584; EPI_ISL_848616; EPI_ISL_861638; EPI_ISL_861642; EPI_ISL_962063; EPI_ISL_1181578; EPI_ISL_1181579; EPI_ISL_1181582; EPI_ISL_1181583; EPI_ISL_1181585; EPI_ISL_1652278; EPI_ISL_1695909; EPI_ISL_2970372; EPI_ISL_2970374; EPI_ISL_3048768; EPI_ISL_514137; EPI_ISL_524470; EPI_ISL_534315; EPI_ISL_568790; EPI_ISL_623167; EPI_ISL_721991; EPI_ISL_747615; EPI_ISL_792393; EPI_ISL_792644; EPI_ISL_801833; EPI_ISL_861651; EPI_ISL_861653; EPI_ISL_1068241; EPI_ISL_1084726; EPI_ISL_1181569; EPI_ISL_1181571; EPI_ISL_1181572; EPI_ISL_1181573; EPI_ISL_1181574; EPI_ISL_1403125; EPI_ISL_1468434; EPI_ISL_1469830; EPI_ISL_1652450; EPI_ISL_2677312; EPI_ISL_2775496; EPI_ISL_480357; EPI_ISL_547571; EPI_ISL_591370; EPI_ISL_591371; EPI_ISL_591389; EPI_ISL_591390; EPI_ISL_717911; EPI_ISL_748144; EPI_ISL_748145; EPI_ISL_748667; EPI_ISL_792383; EPI_ISL_792594; EPI_ISL_801846; EPI_ISL_848574; EPI_ISL_848579; EPI_ISL_848614; EPI_ISL_861662; EPI_ISL_861664; EPI_ISL_918513; EPI_ISL_1167853; EPI_ISL_1181550; EPI_ISL_1181561; EPI_ISL_1181562; EPI_ISL_2557387; EPI_ISL_2603527; EPI_ISL_2677310; EPI_ISL_2775499; EPI_ISL_2970373; EPI_ISL_3190273; EPI_ISL_514138; EPI_ISL_527012; EPI_ISL_583495; EPI_ISL_603025; EPI_ISL_848613; EPI_ISL_1117440; EPI_ISL_1139053; EPI_ISL_1181519; EPI_ISL_1181521; EPI_ISL_1181523; EPI_ISL_1181726; EPI_ISL_1248878; EPI_ISL_1248884; EPI_ISL_1396053; EPI_ISL_2497433; EPI_ISL_2603525; EPI_ISL_2775471; EPI_ISL_2970375; EPI_ISL_3190274; EPI_ISL_478718; EPI_ISL_693247; EPI_ISL_708529; EPI_ISL_735407; EPI_ISL_735414; EPI_ISL_735415; EPI_ISL_735416; EPI_ISL_735425; EPI_ISL_735427; EPI_ISL_735430; EPI_ISL_776539; EPI_ISL_914395; EPI_ISL_977171; EPI_ISL_977172; EPI_ISL_1117442; EPI_ISL_1181509; EPI_ISL_1181510; EPI_ISL_1181511; EPI_ISL_1181512; EPI_ISL_1181513; EPI_ISL_1248893; EPI_ISL_1396054; EPI_ISL_2017449; EPI_ISL_2603519; EPI_ISL_2758797; EPI_ISL_2779435; EPI_ISL_2928141; EPI_ISL_523645; EPI_ISL_547578; EPI_ISL_547580; EPI_ISL_603031; EPI_ISL_603032; EPI_ISL_693227; EPI_ISL_693248; EPI_ISL_735410; EPI_ISL_837558; EPI_ISL_837560; EPI_ISL_1068234; EPI_ISL_1121328; EPI_ISL_1181500; EPI_ISL_1181501; EPI_ISL_1248889; EPI_ISL_1396059; EPI_ISL_1396288; EPI_ISL_1533980; EPI_ISL_2187989; EPI_ISL_2603518; EPI_ISL_2731474; EPI_ISL_2731475; EPI_ISL_3048790; EPI_ISL_603039; EPI_ISL_746642; EPI_ISL_751187; EPI_ISL_751188; EPI_ISL_861891; EPI_ISL_861897

Anexo 3. Polimorfismos encontrados nas sequências de Betim. Inclui a linhagem, a posição e troca, o gene e a frequência do polimorfismo na linhagem.

Linhagem	SNP	Gene	Frequência (%)
B.1.1.28	C241T	Orf1ab	1

B.1.1.33	C241T	Orf1ab	1
B.1.1.28	G246A	Orf1ab	0,06
B.1.1.33	G246A	Orf1ab	0,06
B.1.1.28	G625T	Orf1ab	0,11
B.1.1.33	A929G	Orf1ab	0,06
B.1.1.33	C1218T	Orf1ab	0,06
B.1.1.33	T1831C	Orf1ab	0,06
B.1.1.33	A1989G	Orf1ab	0,06
B.1.1.33	A2255T	Orf1ab	0,06
B.1.1.33	A2491G	Orf1ab	0,06
B.1.1.28	C2638T	Orf1ab	0,06
B.1.1.28	C3037T	Orf1ab	1
B.1.1.33	C3037T	Orf1ab	1
B.1.1.33	A3193C	Orf1ab	0,06
B.1.1.28	A3350G	Orf1ab	0,06
B.1.1.33	A3481G	Orf1ab	0,06
B.1.1.33	G3527T	Orf1ab	0,06
B.1.1.33	C3768T	Orf1ab	0,06
B.1.1.33	C4320T	Orf1ab	0,06
B.1.1.33	C4345T	Orf1ab	0,06
B.1.1.33	C4540T	Orf1ab	0,06
B.1.1.33	C4890T	Orf1ab	0,06
B.1.1.33	C4897T	Orf1ab	0,06
B.1.1.33	G4908A	Orf1ab	0,06
B.1.1.33	G5008T	Orf1ab	0,12
B.1.1.33	C5183T	Orf1ab	0,06
B.1.1.28	A5802G	Orf1ab	0,06
B.1.1.28	T5815C	Orf1ab	0,06
B.1.1.33	C6070T	Orf1ab	0,06
B.1.1.28	C6258T	Orf1ab	0,06
B.1.1.33	T6274C	Orf1ab	0,06
B.1.1.28	A6319G	Orf1ab	0,11
B.1.1.33	G6320T	Orf1ab	0,06

B.1.1.28	G6337A	Orf1ab	0,11
B.1.1.28	C6395A	Orf1ab	0,06
B.1.1.28	C6402T	Orf1ab	0,06
B.1.1.33	A6591G	Orf1ab	0,24
B.1.1.33	C6606T	Orf1ab	0,06
B.1.1.33	T7033A	Orf1ab	0,12
B.1.1.33	C7712T	Orf1ab	0,12
B.1.1.33	C7843T	Orf1ab	0,06
B.1.1.28	C9700T	Orf1ab	0,11
B.1.1.33	C10369T	Orf1ab	0,06
B.1.1.33	C10790T	Orf1ab	0,06
B.1.1.33	T10999C	Orf1ab	0,06
B.1.1.28	C11074T	Orf1ab	0,06
B.1.1.28	G11083T	Orf1ab	0,17
B.1.1.33	G11083T	Orf1ab	0,06
B.1.1.28	A11217G	Orf1ab	0,06
B.1.1.28	G11243T	Orf1ab	0,06
B.1.1.28	G11266T	Orf1ab	0,06
B.1.1.28	G11335T	Orf1ab	0,06
B.1.1.33	A11559G	Orf1ab	0,06
B.1.1.28	C11653G	Orf1ab	0,06
B.1.1.28	C12053T	Orf1ab	0,44
B.1.1.33	A12075G	Orf1ab	0,06
B.1.1.28	C12400T	Orf1ab	0,06
B.1.1.28	T12811G	Orf1ab	0,06
B.1.1.33	C13255T	Orf1ab	0,06
B.1.1.33	C13476T	Orf1ab	0,06
B.1.1.28	G13571T	Orf1ab	0,06
B.1.1.28	C14408T	Orf1ab	1
B.1.1.33	C14408T	Orf1ab	0,94
B.1.1.33	C15197T	Orf1ab	0,06
B.1.1.28	T15338C	Orf1ab	0,06
B.1.1.28	T15420A	Orf1ab	0,06

B.1.1.28	C15952T	Orf1ab	0,06
B.1.1.28	G16365T	Orf1ab	0,06
B.1.1.33	C16650T	Orf1ab	0,06
B.1.1.33	T16731A	Orf1ab	0,06
B.1.1.28	C17172T	Orf1ab	0,06
B.1.1.28	T18083C	Orf1ab	0,06
B.1.1.28	G18498T	Orf1ab	0,11
B.1.1.33	G18583T	Orf1ab	0,06
B.1.1.28	C18657T	Orf1ab	0,06
B.1.1.28	G19086T	Orf1ab	0,17
B.1.1.28	C19367T	Orf1ab	0,06
B.1.1.28	G19542T	Orf1ab	0,06
B.1.1.28	C20132T	Orf1ab	0,06
B.1.1.33	C20132T	Orf1ab	0,24
B.1.1.33	C20930T	Orf1ab	0,06
B.1.1.33	C20946T	Orf1ab	0,06
B.1.1.28	G20995T	Orf1ab	0,06
B.1.1.33	T21066A	Orf1ab	0,06
B.1.1.28	T21408C	Orf1ab	0,33
B.1.1.33	G21786T	S	0,06
B.1.1.33	G22201T	S	0,06
B.1.1.28	C22998T	S	0,06
B.1.1.33	C23029T	S	0,06
B.1.1.28	T23227C	S	0,06
B.1.1.33	T23296C	S	0,06
B.1.1.28	A23403G	S	1
B.1.1.33	A23403G	S	1
B.1.1.28	G23429T	S	0,06
B.1.1.28	C23525T	S	0,06
B.1.1.33	C23625T	S	0,06
B.1.1.33	C23683T	S	0,06
B.1.1.33	T24194C	S	0,06
B.1.1.33	C24718T	S	0,18

B.1.1.28	G25088T	S	1
B.1.1.33	G25217T	S	0,12
B.1.1.33	G25538A	3	0,06
B.1.1.33	C25658T	3	0,06
B.1.1.28	G25687T	3	0,06
B.1.1.33	T25719C	3	0,06
B.1.1.33	C26060T	3	0,12
B.1.1.33	C26111T	3	0,06
B.1.1.28	T26149C	3	0,11
B.1.1.28	G26529C	M	0,06
B.1.1.28	G26769A	M	0,06
B.1.1.33	C27092T	M	0,06
B.1.1.33	G27258T	6	0,06
B.1.1.33	T27299C	6	1
B.1.1.33	A27315T	6	0,06
B.1.1.33	C27371T	6	0,06
B.1.1.33	C27903T	8b	0,06
B.1.1.33	G28191T	8b	0,06
B.1.1.28	T28297C	N	0,06
B.1.1.33	T28297C	N	0,06
B.1.1.28	G28460C	N	0,06
B.1.1.33	A28615G	N	0,06
B.1.1.28	G28851T	N	0,33
B.1.1.28	G28881A	N	1
B.1.1.33	G28881A	N	1
B.1.1.28	G28882A	N	1
B.1.1.33	G28882A	N	1
B.1.1.28	G28883C	N	1
B.1.1.33	G28883C	N	1
B.1.1.28	C29064T	N	0,06
B.1.1.28	G29100A	N	0,06
B.1.1.33	T29148C	N	1
B.1.1.33	C29149T	N	0,12

B.1.1.33	C29171T	N	0,06
B.1.1.28	C29366A	N	0,06
B.1.1.28	G29402T	N	0,06
B.1.1.28	G29414T	N	0,06
B.1.1.33	C29585T	10	0,06
B.1.1.28	G29734T	10	0,06
B.1.1.28	C29741T	10	0,06
B.1.1.28	C29838T	10	0,11

Anexo 4. Artigo '*Seroprevalence, prevalence, and genomic surveillance: monitoring the initial phases of the SARS-CoV-2 pandemic in Betim, Brazil.*'. O documento anexado provém de um arquivo de pré-prints (*medRxiv*). O artigo está em revisão para publicação na revista científica *Frontiers of Microbiology*.

Seroprevalence, prevalence, and genomic surveillance: monitoring the initial phases of the SARS-CoV-2 pandemic in Betim, Brazil

Ana Valesca Fernandes Gilson Silva [1,#,*], Diego Menezes [2,3,#], Filipe Romero Rebello Moreira [4,#], Octavio Alcântara Torres [1], Paula Luize Camargos Fonseca [2,3], Rennan Garcias Moreira [5], Hugo José Alves [2,3], Vivian Ribeiro Alves [1], Tania Maria de Resende Amaral [1], Adriano Neves Coelho [1], Júlia Maria Saraiva Duarte [3],Augusto Viana da Rocha [1], Luiz Gonzaga Paula de Almeida [6], João Locke Ferreira de Araújo [2,3], Hilton Soares de Oliveira [1], Nova Jersey Claudio de Oliveira [1], Camila Zolini de Sá [4], Jôsy Hubner de Sousa [7], Elizângela Gonçalves de Souza [1], Rafael Marques de Souza [2,3], Luciana de Lima Ferreira [2,3], Alexandra Lehmkuhl Gerber [6], Ana Paula de Campos Guimarães [6], Paulo Henrique Silva Maia [1], Fernanda Martins Marim [2,3], Lucyene Miguita [8], Cristiane Campos Monteiro [1], Tuffi Saliba Neto [1], Fabricia Soares Freire Pugêdo [1], Daniel Costa Queiroz [2,3], Damares Nigia Alborguetti Cuzzuol Queiroz [1], Luciana Cunha Resende-Moreira [9], Franciele Martins Santos [7], Erika Fernanda Carlos Souza [1], Carolina Moreira Voloch [4], Ana Tereza Vasconcelos [6], Renato Santana de Aguiar [2,3,10,*], Renan Pedra de Souza [2,3,*]

[1] Escola de Saúde Pública de Betim, Betim, MG, Brazil
[2] Programa de Pós Graduação em Genética; Departamento de Genética, Ecologia e Evolução; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[3] Laboratório de Biologia Integrativa; Departamento de Genética, Ecologia e Evolução; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[4] Departamento de Genética, Instituto de Biologia, Universidade Federal do Rio de Janeiro, RJ, Brazil
[5] Centro de Laboratórios Multiusuários, Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[6] Laboratório Nacional de Computação Científica, Petropólis, RJ, Brazil
[7] Programa de Pós-graduação em Biologia Celular, Departamento de Morfologia; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[8] Departamento de Patologia; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[9] Departamento de Botânica; Instituto de Ciências Biológicas; Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[10] Instituto D'Or de Pesquisa e Ensino (IDOR), Rio de Janeiro, RJ, Brazil
[#] Authors contributed equally
[*] Corresponding author

Renan P. Souza (renanpedra@gmail.com) or Renato S. Aguiar (santanarnt@gmail.com) Universidade Federal de Minas Gerais. Av. Antônio Carlos, 6627 ICB – Pampulha, 31270901 – Belo Horizonte – Minas Gerais – Brazil. Phone: +553134092895.

Ana Valesca F. G. Silva (anavalescafernandes@hotmail.com) Escola de Saúde Pública de Betim. R. Para de Minas, 640 – Brasileia, 32600412 – Betim – Minas Gerais – Brazil. Phone: +553135123000

Abstract:

The Covid-19 pandemic has created an unprecedented need for epidemiological monitoring using diverse strategies. We conducted a project combining prevalence, seroprevalence, and genomic surveillance approaches to describe the initial pandemic stages in Betim City, Brazil. We collected 3239 subjects in a population-based age-, sex- and neighbourhood-stratified, household, prospective; cross-sectional study divided into three surveys 21 days apart sampling the same geographical area. In the first survey, overall prevalence (participants positive in serological or molecular tests) reached 0.46% (90% CI 0.12% – 0.80%), followed by 2.69% (90% CI 1.88% – 3.49%) in the second survey and 6.67% (90% CI 5.42% - 7.92%) in the third. The underreporting reached 11, 19.6, and 20.4 times in each survey, respectively. We observed increased odds to test positive in females compared to males (OR 1.88 95% CI 1.25 – 2.82), while the single best predictor for positivity was ageusia/ anosmia (OR 8.12, 95% CI 4.72 – 13.98). Thirty-five SARS-CoV-2 genomes were sequenced, of which 18 were classified as lineage B.1.1.28, while 17 were B.1.1.33. Multiple independent viral introductions were observed. Integration of multiple epidemiological strategies was able to describe Covid-19 dispersion in the city adequately. Presented results have helped local government authorities to guide pandemic management.

1. Introduction

Since its emergence in December 2019, the new human coronavirus has had a tremendous impact on humanity due to the pandemic nature of its infection, called Covid-19 [1]. The SARS-CoV-2 pathogen was described on January 24, 2020. In Brazil, the first case of Covid-19 was reported on February 26, 2020, in the city of São Paulo [2]. The virus spread rapidly, and the country had the highest number of cases and deaths in Latin America, experiencing its first peak wave in late July 2020. Although most cases were identified in the most prominent Brazilian cities, São Paulo and Rio de Janeiro, dispersion to other municipalities were quickly reported. Betim, a town located in the Minas Gerais State in Brazil with an estimated population of 439,340 in 2019, had its first reported SARS-CoV-2 case on March 23, 2020, in two patients returning from Europe. Two months later, on May 23, 2020, only 73 confirmed cases had been reported, although 4380 suspected cases were identified in public databases indicating limited testing availability.

Brazilian public healthcare system has prioritized testing subjects with symptoms due to scarce diagnostic tests, particularly in the early days of the pandemic. Since data suggest that symptomatic cases represent a fraction of persons infected with SARS-CoV-2, official statistics were expected to be underestimated [3]. Epidemiological surveillance using prevalence studies is needed to evaluate the true extent of SARS-CoV-2 dispersion, significantly extending testing to asymptomatic subjects. Combining serological and molecular tests may be a more robust strategy to uncover viral diffusion in a territory, avoiding each test's kinetic detection limitations. Valid prevalence and seroprevalence estimates for a population rely on two major factors: (i) a representative population sample and (ii) accurate diagnostic testing [4].

While the epidemiological investigation is essential for controlling Covid-19, genomic surveillance is equally crucial. Robust SARS-CoV-2 variant monitoring can track viral evolution, detect new variants, describe patterns and clusters of transmission, outbreak tracking, among others. Therefore, it can provide actionable information on implementing a more targeted public health strategy that addresses local priorities through stakeholder engagement and mitigation efforts [5]. Here, we conducted a study combining seroprevalence, prevalence, and genomic surveillance approaches to understand the SARS-CoV-2 epidemic spread in Betim city.

2 - Materials and Methods:

2.1 – Seroprevalence and prevalence

The Research Ethics Committee approved the present experiment under protocol CAAE 31459220.2.0000.5651. We conducted a population-based age-, sex- and neighbourhood-stratified, household, prospective; cross-sectional study repeated every 21 days in the same geographic area to determine the extent of SARS-CoV-2 transmission in Betim, Minas Gerais, Brazil. Three surveys were held: June 3-5, June 23-25, and July 13-15, 2020. The sample size (n = 1,080 each survey) was estimated considering dichotomous outcome (positive or negative), the population of 439,340 inhabitants, the confidence level of 90%, the maximum margin of error of 2.5%, and lack of a priori information on the prevalence of SARS-COV-2 in the municipality's population (the latter represented by p = q = 0.5) and using the equation below:

$$n = \frac{z^2_{\frac{\alpha}{2}}.\hat{p}.\hat{q}.N}{E^2.(N-1) + z^2_{\frac{\alpha}{2}}.\hat{p}.\hat{q}}$$

Random sampling was employed to ensure representativeness of the population, stratified by sex, age (0 to 5; 6 to 19; 20 to 39; 40 to 59 and 60 years or older) and city neighbourhoods (Centro, Alterosas, Imbiruçu, Norte, Teresópolis, PTB, Citrolândia, Vianópolis, Icaivera, and Petrovale). Every census tract (population stratum created by Governmental agencies) was sampled with at least one address. In case of refusal or closed households, the closest home was selected. Thirty-six teams (one driver, one nurse, and one community health worker) worked on active sampling subjects in 1080 addresses during three days. Clinical and epidemiological data were obtained using a questionnaire during interviews with participants or their legal guardians who signed the Informed Consent. Biological samples were collected using a nasal swab to conduct RT-PCR and capillary blood obtained by fingerstick for the serological test.

RT-PCR to detect SARS-CoV-2 RNA was initially conducted in pools of ten samples [6]. Whenever pools were positive, individual samples were examined. Molecular diagnosis was established according to the CDC 2019-Novel Coronavirus Real-Time RT-PCR Diagnostic Panel (N1, N2 and RNP primers). Serological tests were conducted using the SARS-CoV-2 Antibody Test (Guangzhou Wondfo Biotech Co., Ltd.) that detects IgM/IgG antibodies. The same test was used in a previous study in Brazil [7].

Reported sensitivity is 86.43% (95% CI: 82.41% ~ 89.58%) and specificity 99.57% (95% CI: 97.63% ~ 99.92%). We have validated antibody tests using serum samples from subjects who were SARS-CoV-2 positive confirmed with RT-PCR.

Associations of each variable of interest with surveys (Table 1) and positive status (Table 2) were assessed using chi-square tests. Odds ratios were estimated using logistic regression with the *glm* function. Spatial geostatistical modelling and prediction were carried out using the *gstat* and *predict* functions from the *gstat* package. All analyses were carried out in R software (version 4.1.1).

2.2 – Genomic surveillance

Whole viral genome amplification and DNA library preparation was carried out as described elsewhere [8]. Briefly, QIAseq SARS-CoV-2 Primer Panel - QIAGEN kit was used to amplify positive samples, following manufacturer instructions. In total, 39 of the 84 detectable samples were eligible for library preparation based on their CTs ≤ 30. Library concentration was measured using the QIAseq Library Quant Assay - QIAGEN kit, and the fragment integrity and size were evaluated using Bioanalyzer (Agilent Technologies, Waldbronn, DE). Sequencing was carried out on a MiSeq (Illumina, San Diego, CA, USA).

The raw data generated were filtered by Trimmomatic v0.39 [9], which trimmed low-quality bases (Phred score < 30) and removed short reads (<50 nucleotides) as well as adapters and primer sequences. Reads were then mapped against the SARS-CoV-2 reference genome (accession number: NC_045512.2) with Bowtie2 [10]. The resulting BAM files were manipulated with SAMtools, BCFtools [11], and BEDtools [12] to generate consensus genome sequences. Bases with less than 10x sequencing depth were masked. In total, 35 of the 39 genome sequences presented coverage greater than 79% and average sequencing depth greater than 200x. Sequencing metadata is available in **Table S1**. The 35 consensus genome sequences were submitted to the PANGOLIN 2.0 lineage classification tool (database version February 2, 2021) [13].

To confirm the PANGOLIN identification and further contextualize the diversity of lineages circulating in Betim, we performed a set of phylogenetic analyses. First, a global dataset was assembled from a subset of high-quality data available on GISAID and the newly generated genomes (n = 3,814). This dataset contained all Brazilian sequences and one per week for each country, as available on GISAID until January 12, 2021. These sequences were aligned with MAFFT v7.475[14], and a maximum likelihood tree was

inferred on IQ-Tree 2 [15], under the GTR+F+I+G4 model [16], [17]. Shimoidara-Hasegawa approximate likelihood ratio test (SH-aLRT) was used to assess branches' statistical support [18].

Two subsets of the previous dataset were assembled to explore the temporal dynamics of introduction and circulation of SARS-CoV-2 in Betim, comprehending sequences belonging to lineages B.1.1.28 (n = 258) and B.1.1.33 (n = 284). The parameterization of the phylogeographic model was set to be primarily informative concerning introductions of SARS-CoV-2 in Betim. Therefore, we set the model with six discrete categories: Betim City, Minas Gerais State, Rio de Janeiro State, São Paulo State, other Brazilian States, and foreign sequences. These locations were represented by 18, 2, 22, 71, 79, and 66 sequences in dataset B.1.1.28 while B.1.1.33 dataset composition was 17, 20, 53, 52, 73, and 69 sequences from each region, respectively.

Maximum likelihood trees were inferred from these datasets, and their temporal signal was evaluated with tempest v1.5.3 [19]. Time scaled phylogenies were then inferred from these datasets with BEAST v1.10.4 [20], using: (***i***) the HKY+I+G4 nucleotide substitution model [17], (***ii***) the strict molecular clock model, (***iii***) the non-parametric coalescent skygrid tree prior [21] and (***iv***) a symmetric discrete phylogeographic model [22]. A normal prior distribution (mean = 1.13×10^{-3}; std= 5.1×10^{-4}) on clock rate was assumed, based on a previous estimate [23]. The cutoff values of the skygrid tree prior were set based on the previously estimated dates for the emergence of each lineage [23]. The number of grids of the tree priors was set to match the approximate number of weeks comprehended between the estimated dates for lineages' emergence and the dates of the most recently sampled sequences (41 weeks, both datasets). Two and three independent chains of 200 million generations sampling every 10,000 states were performed for datasets B.1.1.33 and B.1.1.28, respectively. Tracer v1.7.1 [24] was used to verify mixing and convergence of chains (effective sample size > 200 for all parameters), which were then combined with logcombiner v1.10.4 after 10% burning removal. Maximum clade credibility trees were generated with treeannotator v1.10.4. All logs and trees are available in https://github.com/LBI-lab/SARS-CoV-2_phylogenies.git.

3 - Results

3.1 – Seroprevalence and prevalence

Table 1 presents clinical and epidemiological data obtained from participants. No significant difference was observed for the presence of any prior health condition across surveys (pneumopathy, chronic neurological disease, pregnant, postpartum, chronic cardiovascular disease, chronic kidney disease, obesity, asthma, immunodepression, chronic liver disease, diabetes, hypertension, transplanted, cancer or any comorbidity) indicating proper sampling was conducted since there was no reason to find significant differences in the period. Four symptoms (cough, sore throat, myalgia, and rhinorrhea) and contact with a symptomatic person increased while international travel decreased. Prevalence and seroprevalence increased across surveys.

Pandemic progression in Betim city is presented in **Figure 1**. Confirmed cases underestimation was found in all three surveys. In the first survey, overall prevalence (participants positive in serological or molecular tests) reached 0.46% (90% CI 0.12% – 0.80%), followed by 2.69% (90% CI 1.88% – 3.49%) in the second survey and 6.67% (90% CI 5.42% - 7.92%) in the third. The underreporting was obtained by the difference between survey prevalence and official data, and its magnitude reached 11, 19.6, and 20.4 times (distance between black dots and red curve in Figure 1B). Active transmission areas (RT-PCR positive participants) were observed increasing across time (**Figure 1C-E**). By the third survey, almost all populated city areas were likely to have viral circulation (**Figure 1E**). The same pattern of increase was observed in overall prevalence for most administrative regions (**Figure 1F-G**).

We have also evaluated whether clinical and epidemiological variables were associated with molecular or serological test positivity (**Table 2**). Several significant results were observed, mostly with reported symptoms (fever, cough, sore throat, dyspnoea, myalgia, rhinorrhea, respiratory discomfort, nausea/ vomit, headache, prostration, ageusia/ anosmia). We also observed increased odds to test positive in females compared to males (OR 1.88 95% CI 1.25 – 2.82) and clear enrichment of positive cases in certain city regions (e.g., Imbiruçu and Terezópolis). Surprisingly, people with obesity were more likely to be positive (OR 3.33, 95% CI 1.68 – 6.59). The single best predictor for positivity was ageusia/ anosmia (OR 8.12, 95% CI 4.72 – 13.98). Non-significant results can be found in **Table S2**.

3.2 – Genomic viral surveillance

In total, 35 novel SARS-CoV-2 genome sequences were obtained (GISAID EPI_ISL_5416087-5416121). The sequences were classified by PANGOLIN 2.0 to assess the genetic diversity of SARS-CoV-2 circulating in Betim. 18 of the 35 genomes were classified as lineage B.1.1.28, while 17 were B.1.1.33 (*Probability* = 1.0). Further, a maximum likelihood tree was inferred from the global dataset GISAID [25].

The analysis supported these results, revealing sequences from the Betim cluster within several clades of these lineages confirming the circulation of (B.1.1.28 and B.1.1.33 during the first wave of COVID-19 pandemics in the city (**Figure 2**). The spread of Betim sequences across the tree suggests multiple independent introductions occurred in the town. Further, eight clades majorly composed by Betim sequences were inferred with variable degrees of statistical support (median SH-aLRT = 82.75, range: 0 - 100), suggesting the occurrence of local transmission in the city after initial introduction events. In addition to these clusters, nine introductions supported by single sequences have also been detected. Most Betim sequences or clusters are closely related to sequences from Rio de Janeiro and São Paulo, two neighbouring States connected by highways to Minas Gerais. To formally assess the dynamics of introduction and spread of SARS-CoV-2 in Betim, separated datasets for lineages B.1.1.28 and B.1.1.33 were evaluated. Regression between sampling times and genetic distances revealed both datasets had moderate temporal signal (B.1.1.28: $R^2 = 0.49$; B.1.1.33: $R^2 = 0.58$), justifying molecular clock analysis.

The time-scaled phylogeographic analysis performed with dataset B.1.1.28 suggests this lineage emerged on February 22, 2020, in São Paulo (95% highest posterior density, HPD: February 11, 2020 - March 05, 2020; geographic model posterior probability, PP = 0.91), later spreading to other Brazilian states (**Figure 3A**). The phylogeny reveals that two introduction events, dated between April 19, 2020 (95% HPD: April 17, 2020 - May 11, 2020) and April 22, 2020 (95% HPD: April 20, 2020 – May 27, 2020), led to the emergence of Betim clusters (harbouring between two and six sequences). Additionally, four introductions related to single sequences have been detected. The phylogeographic model suggests that three introductions occurred from another Brazilian region to Betim, in addition to other single events from RJ, another one from SP, and another from foreign sequences. All events presented high statistical support (PP > 92% for all events).

The phylogeographic reconstruction performed for dataset B.1.1.33 infers the origin of this lineage on February 06, 2020, in Rio de Janeiro (95% HPD: January 14, 2020 – February 25, 2020, PP = 0.78). The model supports the occurrence of many Betim clusters. One cluster comprises four sequences, dating to May 27, 2020 (95% HPD: May 01, 2020 - June 03, 2020) grouped with other sequences from other Brazilian regions and foreign. The model has also estimated eight introductions supported by single sequences. According to our phylogeny, the B.1.1.33 introductions came from different locations, such as the states of Rio de Janeiro, São Paulo, Minas Gerais, other Brazilian regions, and foreign sequences (PP > 0.81 for all events) (**Figure 3B**). The patterns reconstructed by both phylogeographic inferences are consistent, indicating the emergence of lineages B.1.1.28 and B.1.1.33 was followed by multiple importation events to diverse regions within the country, likely driven by human mobility. Additionally, evolutionary rate estimates also differed between datasets (B.1.1.28: 8.6372×10^{-4}, 95% HPD: 7.8379×10^{-4} - 9.4559×10^{-4}; B.1.1.33: 6.8743×10^{-4}, 95% HPD 6.1784×10^{-4} - 7.5446×10^{-4}).

4 - Discussion

Betim is a medium-sized Brazilian city (439,340 inhabitants, 343 thousand square kilometres) crossed by national roads connecting major Brazilian cities and serving as a local hub for the Brazilian Public Health System. Understanding its pandemic dynamic may provide relevant information for municipalities with similar features. Here, we estimated the overall prevalence of active infections, seroprevalence and conducted genomic surveillance before the first pandemic wave in Betim. Brazilian molecular diagnostic capacity was insufficient in the first months of the pandemic [26]. Therefore, Covid-19 cases may have been included in the official statistics as severe acute respiratory infection cases with unknown aetiology. Data until May 2020 indicated a positive association between higher per-capita income and molecular Covid-19 diagnosis, while the severe acute respiratory infection cases with unknown aetiology were associated with lower per-capita income, suggesting a possible diagnosis bias related to economic status [27]. Inadequate diagnosis availability may lead to underreporting [28]. Our data estimated underreporting rates up to 20 times.

No studies have been conducted in Brazil evaluating active infection prevalence using adequate sampling. Our study design was inspired by previous research conducted in Santa Clara, USA, using pooled samples [29]. Pooled PCR tests were initially suggested to be used in asymptomatic people [6] and later were recommended for surveillance studies in populations with low infection prevalence [30]. Seroprevalence studies were conducted during the first wave in Brazil that peaked in July 2020. Two of the highest city seroprevalences reported during the period were Boa Vista (25.4% in June 2020) [7] and São Luiz (40.4% between the end of July and August 2020) [31], both in the northern area of the country. A nationwide survey carried out in May and June 2020 presented seroprevalence lower than two per cent during both surveys in all sampled cities neighbouring Betim (less than 200km), corroborating our findings [7]. Furthermore, seroprevalences higher than ten per cent were solely found in towns in the North Region [7]. In December 2020, Manaus, the largest city in the North Region, experienced a resurgence of Covid-19 [32] despite high seroprevalence [33], likely due to the gamma variant [34].

Previous seroprevalence studies have indicated ethnic and socioeconomic bias for SARS-CoV-2 infection in Brazil since the pandemic's beginning [35], [36]. Results from Rio de Janeiro in April 2020 indicated that younger blood donors with lower education

levels were more likely to test positive for SARS-CoV-2 antibodies [35]. A nationwide study revealed that the poorest quintile was 2.16 times more likely to test positive with the lowest risks among white, educated, and wealthy individuals [36]. Likewise, we found one of the highest prevalences in the poorest neighbourhood, Terezópolis, that include the largest slum of the city where more than 23 thousand people live.

Further modelling results showed higher infection rates among young adults, lower socioeconomic status, and people without healthcare access in the less developed North and Northeast areas until August 2020 [37]. Betim also presents most of its inhabitants with less than 59 years (90.7%), but no age effect was observed in the infectivity rates. Increased female infection odds were observed, although previous reports indicated a gender predisposition towards death in some Brazilian regions with higher male risk [38]. One possible explanation could be that 70% of the global health workforce are women [39] and a gender bias of pandemic perception and attitude [40].

Covid-19 diffusion presents strong socio-spatial determinants. Relocation diffusion from more- to less-developed regions and hierarchical diffusion from countries with higher population and density were relevant since early 2020 [41]. Data indicated a similar pattern in the São Paulo State with contiguous diffusion from the capital metropolitan area and hierarchical with long-distance spread through major highways that connects São Paulo city with cities of regional relevance [42]. Modelling results revealed that São Paulo city may have accounted for more than 85% of the initial case spread in the entire country [43]. Betim is directly connected to São Paulo city by a main national highway which may have contributed to Covid-19 diffusion.

Genomic surveillance is a powerful tool to elucidate viral dispersion patterns. The first sequencing work conducted in Brazil evaluated the first six positive individuals and reported the same predominant lineages found in Italy [44]. Later, a study with samples collected until late April 2020 from different country areas showed the dominance of clade B-derived lineages. At the national level, the respective frequency of these clades was seen in a 98.98%/1.02% ratio [23]. In Minas Gerais State, A lineages represented 2.5% of the infections, B.1 appeared in 92.5% of the samples, and B was responsible for 5% of the cases [45]. The exclusivity of lineages B.1.1.28 and B.1.1.33 circulating in Betim-MG from June to July 2020, given that multiple introductions from different country regions were demonstrated, is representative of the extent of these lineages' dominance in the Brazilian scenario at the moment. Independent introductions also

emphasize the importance of inter-state mobility barriers as a measure to control the epidemic.

Our study presents some limitations. First, the household survey is less likely to sample severe cases, thus underestimating symptomatic Covid-19. Second, all clinical data were self-reported, which may lead to reporting bias [46]. Third, we could not sequence all PCR positive samples due to the low viral load and sequencing technology employed. Nevertheless, our study shows the potential to integrate different epidemiological inquiries (prevalence, seroprevalence, and genomic surveillance) to describe pandemic dispersion adequately. Moreover, our findings present original and relevant evidence that has helped local government authorities to guide pandemic management.

Conflict of interest

None

Acknowledgement

We want to thank nurses, community health workers, drivers and management personnel who collaborated in this project. We also thank Mr. Guilherme Carvalho da Paixão for his support. We gratefully acknowledge the authors from the originating laboratories responsible for obtaining the specimens and the submitting laboratories where genetic sequence data were generated and shared via the GISAID Initiative, on which this research is based (**Table S3**).

Funding

We acknowledge support from the Fundo Municipal de Saúde de Betim, Rede Corona-ômica BR MCTI/FINEP affiliated to RedeVírus/MCTI (FINEP 01.20.0029.000462/20, CNPq 404096/2020-4), CNPq (A.T.R.V. 303170/2017-4; R.S.A.: 312688/2017-2 and 439119/2018-9; R.P.S.: 310627/2018-4), MEC/CAPES (14/2020 - 23072.211119/2020-10), FINEP (0494/20 01.20.0026.00 and UFMG-NB3 1139/20), FAPEMIG (R.P.S.: APQ-00475-20) and FAPERJ (A.T.R.V. E-26/202.903/20 and Corona-ômica-RJ E-26/210.179/2020; C.M.V: 26/010.002278/2019; R.S.A 202.922/2018).

References:

[1] P. Zhou *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, 2020, doi: 10.1038/s41586-020-2012-7.

[2] D. B. Araujo *et al.*, "SARS-CoV-2 isolation from the first reported patients in brazil and establishment of a coordinated task network," *Mem. Inst. Oswaldo Cruz*, vol. 115, 2020, doi: 10.1590/0074-02760200342.

[3] S. L. Wu *et al.*, "Substantial underestimation of SARS-CoV-2 infection in the United States," *Nat. Commun.*, vol. 11, no. 1, 2020, doi: 10.1038/s41467-020-18272-4.

[4] O. Byambasuren *et al.*, "Comparison of seroprevalence of SARS-CoV-2 infections with cumulative and imputed COVID-19 cases: Systematic review," *PLoS ONE*, vol. 16, no. 4 April. 2021, doi: 10.1371/journal.pone.0248946.

[5] J. D. Robishaw *et al.*, "Genomic surveillance to combat COVID-19: challenges and opportunities," *The Lancet Microbe*, vol. 2, no. 9, 2021, doi: 10.1016/s2666-5247(21)00121-x.

[6] S. Lohse *et al.*, "Pooling of samples for testing for SARS-CoV-2 in asymptomatic people," *The Lancet Infectious Diseases*, vol. 20, no. 11. 2020, doi: 10.1016/S1473-3099(20)30362-5.

[7] P. C. Hallal *et al.*, "SARS-CoV-2 antibody prevalence in Brazil: results from two successive nationwide serological household surveys," *Lancet Glob. Heal.*, vol. 8, no. 11, 2020, doi: 10.1016/S2214-109X(20)30387-9.

[8] F. R. R. Moreira *et al.*, "Epidemic spread of sars-cov-2 lineage b.1.1.7 in Brazil," *Viruses*, vol. 13, no. 6. 2021, doi: 10.3390/v13060984.

[9] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, 2014, doi: 10.1093/bioinformatics/btu170.

[10] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-

efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, 2009, doi: 10.1186/gb-2009-10-3-r25.

[11] H. Li *et al.*, "The Sequence Alignment / Map (SAM) Format and SAMtools 1000 Genome Project Data Processing Subgroup," *Bioinformatics*, vol. 25, no. 16, 2009.

[12] A. R. Quinlan and I. M. Hall, "BEDTools: A flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, 2010, doi: 10.1093/bioinformatics/btq033.

[13] A. Rambaut *et al.*, "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," *Nat. Microbiol.*, vol. 5, no. 11, 2020, doi: 10.1038/s41564-020-0770-5.

[14] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: Improvements in performance and usability," *Mol. Biol. Evol.*, vol. 30, no. 4, 2013, doi: 10.1093/molbev/mst010.

[15] B. Q. Minh *et al.*, "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Mol. Biol. Evol.*, vol. 37, no. 5, 2020, doi: 10.1093/molbev/msaa015.

[16] S. Tavaré, "Some probabilistic and statistical problems in the analysis of DNA sequences," *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, vol. 17. 1986.

[17] Z. Yang, "Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods," *J. Mol. Evol.*, vol. 39, no. 3, 1994, doi: 10.1007/BF00160154.

[18] S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0," *Syst. Biol.*, vol. 59, no. 3, 2010, doi: 10.1093/sysbio/syq010.

[19] A. Rambaut, T. T. Lam, L. M. Carvalho, and O. G. Pybus, "Exploring the

temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)," *Virus Evol.*, vol. 2, no. 1, 2016, doi: 10.1093/ve/vew007.

[20] M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut, "Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10," *Virus Evol.*, vol. 4, no. 1, 2018, doi: 10.1093/ve/vey016.

[21] M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, and M. A. Suchard, "Improving bayesian population dynamics inference: A coalescent-based model for multiple loci," *Mol. Biol. Evol.*, vol. 30, no. 3, 2013, doi: 10.1093/molbev/mss265.

[22] P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard, "Bayesian phylogeography finds its roots," *PLoS Comput. Biol.*, vol. 5, no. 9, 2009, doi: 10.1371/journal.pcbi.1000520.

[23] D. S. Candido *et al.*, "Evolution and epidemic spread of SARS-CoV-2 in Brazil," *Science (80-.).*, vol. 369, no. 6508, pp. 1255–1260, 2020, doi: 10.1126/SCIENCE.ABD2161.

[24] A. Rambaut, A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard, "Posterior summarization in Bayesian phylogenetics using Tracer 1.7," *Syst. Biol.*, vol. 67, no. 5, 2018, doi: 10.1093/sysbio/syy032.

[25] Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data – from vision to reality," *Eurosurveillance*, vol. 22, no. 13. 2017, doi: 10.2807/1560-7917.ES.2017.22.13.30494.

[26] R. M. T. Grotto *et al.*, "Increasing molecular diagnostic capacity and COVID-19 incidence in Brazil," *Epidemiol. Infect.*, 2020, doi: 10.1017/S0950268820001818.

[27] W. M. de Souza *et al.*, "Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil," *Nat. Hum. Behav.*, vol. 4, no. 8, 2020, doi: 10.1038/s41562-020-0928-4.

[28] E. Kupek, "How many more? Under-reporting of the COVID-19 deaths in Brazil

in 2020," *Trop. Med. Int. Heal.*, vol. 26, no. 9, 2021, doi: 10.1111/tmi.13628.

[29] C. A. Hogan, M. K. Sahoo, and B. A. Pinsky, "Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2," *JAMA - Journal of the American Medical Association*, vol. 323, no. 19. 2020, doi: 10.1001/jama.2020.5445.

[30] L. Mutesa *et al.*, "A pooled testing strategy for identifying SARS-CoV-2 at low prevalence," *Nature*, vol. 589, no. 7841, 2021, doi: 10.1038/s41586-020-2885-5.

[31] A. A. M. da Silva *et al.*, "Population-based seroprevalence of SARS-CoV-2 and the herd immunity threshold in Maranhão," *Rev. Saude Publica*, vol. 54, 2020, doi: 10.11606/s1518-8787.2020054003278.

[32] E. C. Sabino *et al.*, "Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence," *The Lancet*, vol. 397, no. 10273. 2021, doi: 10.1016/S0140-6736(21)00183-5.

[33] L. F. Buss *et al.*, "Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic," *Science (80-.).*, vol. 371, no. 6526, 2021, doi: 10.1126/science.abe9728.

[34] N. R. Faria *et al.*, "Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil," *Science (80-.).*, vol. 372, no. 6544, 2021, doi: 10.1126/science.abh2644.

[35] L. A. Filho *et al.*, "Seroprevalence of anti-SARS-CoV-2 among blood donors in Rio de Janeiro, Brazil," *Rev. Saude Publica*, vol. 54, 2020, doi: 10.11606/s1518-8787.2020054002643.

[36] B. L. Horta *et al.*, "Prevalence of antibodies against SARS-CoV-2 according to socioeconomic and ethnic status in a nationwide Brazilian survey," *Rev. Panam. Salud Publica/Pan Am. J. Public Heal.*, vol. 40, 2020, doi: 10.26633/RPSP.2020.135.

[37] E. E. Campos de Lima, E. Gayawan, E. A. Baptista, and B. L. Queiroz, "Spatial pattern of COVID-19 deaths and infections in small areas of Brazil," *PLoS One*,

vol. 16, no. 2 Febuary, 2021, doi: 10.1371/journal.pone.0246808.

[38] P. Baqui, I. Bica, V. Marra, A. Ercole, and M. van der Schaar, "Ethnic and regional variations in hospital mortality from COVID-19 in Brazil: a cross-sectional observational study," *Lancet Glob. Heal.*, vol. 8, no. 8, 2020, doi: 10.1016/S2214-109X(20)30285-0.

[39] G. Lotta, M. Fernandez, D. Pimenta, and C. Wenham, "Gender, race, and health workers in the COVID-19 pandemic," *The Lancet*, vol. 397, no. 10281. 2021, doi: 10.1016/S0140-6736(21)00530-4.

[40] V. Galasso, V. Pons, P. Profeta, M. Becher, S. Brouard, and M. Foucault, "Gender differences in COVID-19 attitudes and behavior: Panel evidence from eight countries," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 44, 2020, doi: 10.1073/pnas.2012520117.

[41] T. Sigler *et al.*, "The socio-spatial determinants of COVID-19 diffusion: the impact of globalization, settlement characteristics and population," *Global. Health*, vol. 17, no. 1, 2021, doi: 10.1186/s12992-021-00707-2.

[42] C. M. C. Branco Fortaleza *et al.*, "The use of health geography modeling to understand early dispersion of COVID-19 in São Paulo, Brazil," *PLoS One*, vol. 16, no. 1 January, 2021, doi: 10.1371/journal.pone.0245051.

[43] M. A. L. Nicolelis, R. L. G. Raimundo, P. S. Peixoto, and C. S. Andreazzi, "The impact of super-spreader cities, highways, and intensive care availability in the early stages of the COVID-19 epidemic in Brazil," *Sci. Rep.*, vol. 11, no. 1, 2021, doi: 10.1038/s41598-021-92263-3.

[44] J. G. de Jesus *et al.*, "Importation and early local transmission of covid-19 in brazil, 2020," *Rev. Inst. Med. Trop. Sao Paulo*, vol. 62, 2020, doi: 10.1590/S1678-9946202062030.

[45] J. Xavier *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.*, vol. 9, no. 1, 2020, doi: 10.1080/22221751.2020.1803146.

[46] M. Baker, M. Stabile, and C. Deri, "What do self-reported, objective, measures of health measure?," *J. Hum. Resour.*, vol. 39, no. 4, 2004, doi: 10.2307/3559039.

Table 1: Clinical and epidemiological data obtained from participants. Bolded p values indicate $p < 0.05$.

Variable	Level	Overall n (%)	First survey n (%)	Second survey n (%)	Third survey n (%)	p-value
Administrative Regions	Alterosas	634 (19.6%)	198 (18.4%)	218 (20.2%)	218 (20.2%)	0.9584
	Citrolândia	219 (6.8%)	83 (7.7%)	68 (6.3%)	68 (6.3%)	
	Icaivera	62 (1.9%)	20 (1.9%)	21 (1.9%)	21 (1.9%)	
	Imbiruçu	565 (17.4%)	183 (17.0%)	191 (17.7%)	191 (17.7%)	
	Norte	333 (10.3%)	111 (10.3%)	111 (10.3%)	111 (10.3%)	
	Petrovale	41 (1.3%)	13 (1.2%)	14 (1.3%)	14 (1.3%)	
	PTB	290 (9.0%)	108 (10.0%)	91 (8.4%)	91 (8.4%)	
	Sede	583 (18.0%)	201 (18.6%)	191 (17.7%)	191 (17.7%)	
	Terezópolis	319 (9.8%)	109 (10.1%)	105 (9.7%)	105 (9.7%)	
	Vianópolis	193 (6.0%)	53 (4.9%)	70 (6.5%)	70 (6.5%)	
Sex	Female	1628 (50.3%)	548 (50.8%)	536 (49.6%)	544 (50.4%)	0.8619
Age range	0 - 5	217 (6.7%)	71 (6.6%)	73 (6.8%)	73 (6.8%)	1.0000
	6 - 19	650 (20.1%)	218 (20.2%)	217 (20.1%)	215 (19.9%)	
	20-39	1067 (32.9%)	354 (32.8%)	355 (32.9%)	358 (33.1%)	
	40-59	871 (26.9%)	291 (27.0%)	289 (26.8%)	291 (26.9%)	
	Above 60	434 (13.4%)	145 (13.4%)	146 (13.5%)	143 (13.2%)	
Pneumopathy	Yes	30 (0.9%)	7 (0.6%)	13 (1.2%)	10 (0.9%)	0.4042
Chronic neurological disease	Yes	39 (1.2%)	16 (1.5%)	10 (0.9%)	13 (1.2%)	0.4948
Pregnant	Yes	28 (0.9%)	10 (0.9%)	11 (1.0%)	7 (0.6%)	0.6257
Postpartum	Yes	9 (0.3%)	2 (0.2%)	3 (0.3%)	4 (0.4%)	0.7165
Chronic cardiovascular disease	Yes	96 (3.0%)	34 (3.2%)	39 (3.6%)	23 (2.1%)	0.1154
Chronic kidney disease	Yes	50 (1.5%)	24 (2.2%)	12 (1.1%)	14 (1.3%)	0.0799
Obesity	Yes	105 (3.2%)	33 (3.1%)	37 (3.4%)	35 (3.2%)	0.8903
Asthma	Yes	173 (5.3%)	65 (6.0%)	58 (5.4%)	50 (4.6%)	0.3537
Immunodepression	Yes	22 (0.7%)	9 (0.8%)	5 (0.5%)	8 (0.7%)	0.5507
Chronic liver disease	Yes	15 (0.5%)	4 (0.4%)	7 (0.6%)	4 (0.4%)	0.5478
Diabetes	Yes	228 (7.0%)	78 (7.2%)	74 (6.9%)	76 (7.0%)	0.9430
Hypertension	Yes	563 (17.4%)	190 (17.6%)	186 (17.2%)	187 (17.3%)	0.9698
Transplanted	Yes	4 (0.1%)	2 (0.2%)	1 (0.1%)	1 (0.1%)	0.7780
Cancer	Yes	23 (0.7%)	10 (0.9%)	8 (0.7%)	5 (0.5%)	0.4342
Any comorbidity	Yes	955 (29.5%)	327 (30.3%)	320 (29.6%)	308 (28.5%)	0.6552
Fever	Yes	224 (6.9%)	66 (6.1%)	70 (6.5%)	88 (8.1%)	0.1398
Cough	Yes	648 (20.0%)	185 (17.1%)	213 (19.7%)	250 (23.1%)	**0.0022**
Sore throat	Yes	397 (12.3%)	112 (10.4%)	125 (11.6%)	160 (14.8%)	**0.0051**
Dyspnoea	Yes	141 (4.4%)	49 (4.5%)	46 (4.3%)	46 (4.3%)	0.9336
Myalgia	Yes	284 (8.8%)	74 (6.9%)	99 (9.2%)	111 (10.3%)	**0.0165**
Rhinorrhea	Yes	717 (22.1%)	205 (19.0%)	240 (22.2%)	272 (25.2%)	**0.0025**
Respiratory discomfort	Yes	188 (5.8%)	63 (5.8%)	58 (5.4%)	67 (6.2%)	0.7084
Nausea/ vomit	Yes	120 (3.7%)	37 (3.4%)	39 (3.6%)	44 (4.1%)	0.7156
Headache	Yes	790 (24.4%)	244 (22.6%)	259 (24.0%)	287 (26.6%)	0.0936
Prostration	Yes	188 (5.8%)	60 (5.6%)	51 (4.7%)	77 (7.1%)	0.0523
Diarrhea	Yes	211 (6.5%)	59 (5.5%)	76 (7.0%)	76 (7.0%)	0.2336
Conjunctivitis	Yes	32 (1.0%)	13 (1.2%)	11 (1.0%)	8 (0.7%)	0.5478
Ageusia/ anosmia	Yes	101 (3.1%)	30 (2.8%)	30 (2.8%)	41 (3.8%)	0.2914
Loss of voice	Yes	56 (1.7%)	18 (1.7%)	13 (1.2%)	25 (2.3%)	0.1381
Sought health assistance	Hospital	138 (4.3%)	41 (3.8%)	41 (3.8%)	56 (5.2%)	0.1492
	Basic Health Unit	129 (4.0%)	42 (3.9%)	41 (3.8%)	46 (4.3%)	
	Emergency Care Unit	127 (3.9%)	38 (3.5%)	35 (3.2%)	54 (5.0%)	
	None	2845 (87.8%)	958 (88.8%)	963 (89.2%)	924 (85.6%)	
Admitted to a health institution	Yes	38 (1.2%)	11 (1.0%)	12 (1.1%)	15 (1.4%)	0.7085
International travel	Yes	14 (0.4%)	10 (0.9%)	4 (0.4%)	0 (0.0%)	**0.0043**
Household contact with symptomatic person	Yes	640 (19.8%)	157 (14.6%)	193 (17.9%)	290 (26.9%)	**< 0.0001**
Sorological test	Reactive	39 (1.2%)	3 (0.3%)	8 (0.7%)	28 (2.6%)	**< 0.0001**
	Non-reactive	3200 (98.8%)	1076 (99.7%)	1072 (99.3%)	1052 (97.4%)	
PCR test	Detected	84 (2.6%)	2 (0.2%)	22 (2.0%)	60 (5.6%)	**< 0.0001**
	Undetected	3112 (96.1%)	1035 (95.9%)	1057 (98.0%)	1020 (94.4%)	
	Indeterminate	42 (1.3%)	42 (3.9%)	0 (0.0%)	0 (0.0%)	
Prevalence	Sorological reacive and/ or PCR detected	106 (3.3%)	5 (0.5%)	29 (2.7%)	72 (6.7%)	**< 0.0001**

Table 2: Significant associations of clinical and epidemiological data with positive test (serological or molecular). Non-significant associations are presented in Table S1. Bolded p values indicate $p < 0.05$.

Variable	Level	Positive	Negative	p-value
Survey	First	5 (4.7%)	1074 (34.3%)	**< 0.0001**
	Second	29 (27.4%)	1051 (33.5%)	
	Third	72 (67.9%)	1008 (32.2%)	
Administrative Regions	Alterosas	18 (17.0%)	616 (19.7%)	**0.0024**
	Citrolândia	4 (3.8%)	215 (6.9%)	
	Icaivera	0 (0.0%)	62 (2.0%)	
	Imbiruçu	32 (30.2%)	533 (17.0%)	
	Norte	11 (10.4%)	322 (10.3%)	
	Petrovale	0 (0.0%)	41 (1.3%)	
	PTB	8 (7.5%)	282 (9.0%)	
	Sede	15 (14.2%)	568 (18.1%)	
	Terezópolis	17 (16.0%)	302 (9.6%)	
	Vianópolis	1 (0.9%)	192 (6.1%)	
Sex	Female	69 (65.1%)	1559 (49.8%)	**0.0026**
Fever	No	88 (83.0%)	2927 (93.4%)	**< 0.0001**
	Yes	18 (17.0%)	206 (6.6%)	
Cough	No	73 (68.9%)	2518 (80.4%)	**0.0053**
	Yes	33 (31.1%)	615 (19.6%)	
Sore throat	No	77 (72.6%)	2765 (88.3%)	**< 0.0001**
	Yes	29 (27.4%)	368 (11.7%)	
Dyspnoea	No	96 (90.6%)	3002 (95.8%)	**0.0180**
	Yes	10 (9.4%)	131 (4.2%)	
Myalgia	No	72 (67.9%)	2883 (92.0%)	**< 0.0001**
	Yes	34 (32.1%)	250 (8.0%)	
Rhinorrhea	No	70 (66.0%)	2452 (78.3%)	**0.0041**
	Yes	36 (34.0%)	681 (21.7%)	
Respiratory discomfort	No	90 (84.9%)	2961 (94.5%)	**< 0.0001**
	Yes	16 (15.1%)	172 (5.5%)	
Nausea/ vomit	No	94 (88.7%)	3025 (96.6%)	**< 0.0001**
	Yes	12 (11.3%)	108 (3.4%)	
Headache	No	50 (47.2%)	2399 (76.6%)	**< 0.0001**
	Yes	56 (52.8%)	734 (23.4%)	
Prostration	No	83 (78.3%)	2968 (94.7%)	**< 0.0001**
	Yes	23 (21.7%)	165 (5.3%)	
Ageusia/ anosmia	No	87 (82.1%)	3051 (97.4%)	**< 0.0001**
	Yes	19 (17.9%)	82 (2.6%)	
Obesity	No	96 (90.6%)	3038 (97.0%)	**< 0.0001**
	Yes	10 (9.4%)	95 (3.0%)	
Sought health assistance	Hospital	8 (7.5%)	130 (4.1%)	**0.0032**
	None	81 (76.4%)	2764 (88.2%)	
	Basic Health Unit	8 (7.5%)	121 (3.9%)	
	Emergency Care Unit	9 (8.5%)	118 (3.8%)	
Household contact with symptomatic person	No	71 (67.0%)	2528 (80.7%)	**0.0007**
	Yes	35 (33.0%)	605 (19.3%)	

Figure 1: Covid-19 pandemic progression in Betim. (A) Absolute number of new cases according to official city statistics. (B) Cumulative number of cases according to official city statistics. Black dots indicate estimated overall prevalence (immunological and molecular tests) in the current study with its 95% confidence interval. Distance from black dots and red curve represent underreporting. (C-E) Dispersion of positive molecular tests across each survey. In the third survey (panel E), most populated areas of the city already had a non-null probability of presenting residents with a positive molecular test. (F-G) Overall prevalence (immunological and molecular tests) comparison in each of the ten administrative regions of the city across successive surveys. An increase was observed in most areas from the first to the second survey and, more substantially, from the second to the third survey.

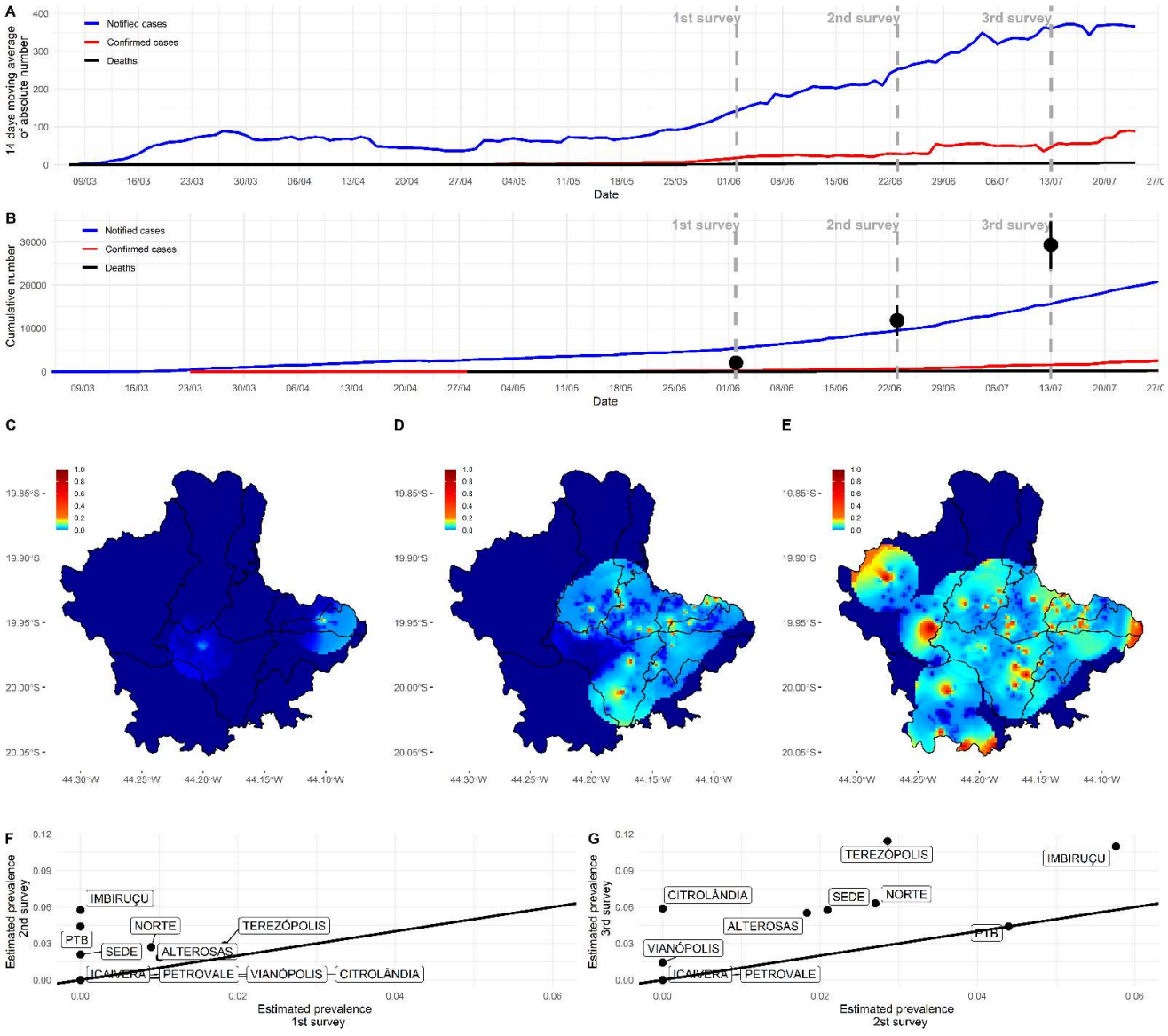

Figure 2: Phylogenetic characterization of SARS-CoV-2 genomes characterized in Betim. A maximum-likelihood tree was inferred on IQ-Tree under the GTR+F+I+G4 model with a comprehensive reference dataset, encompassing all Brazilian sequences plus one international sequence per country per week, from late 2019 to January 12 2021 (n = 3,814). The phylogeny depicted exhibits a subtree of 2,023 tips that harbours all relevant diversity considered for this study, mainly lineages B.1.1.28 (light salmon) and B.1.1.33 (light blue) where the novel genome sequences sparsely clustered. Tip shapes mark sequences characterized in this study. The scale bar indicates average nucleotide substitutions per site.

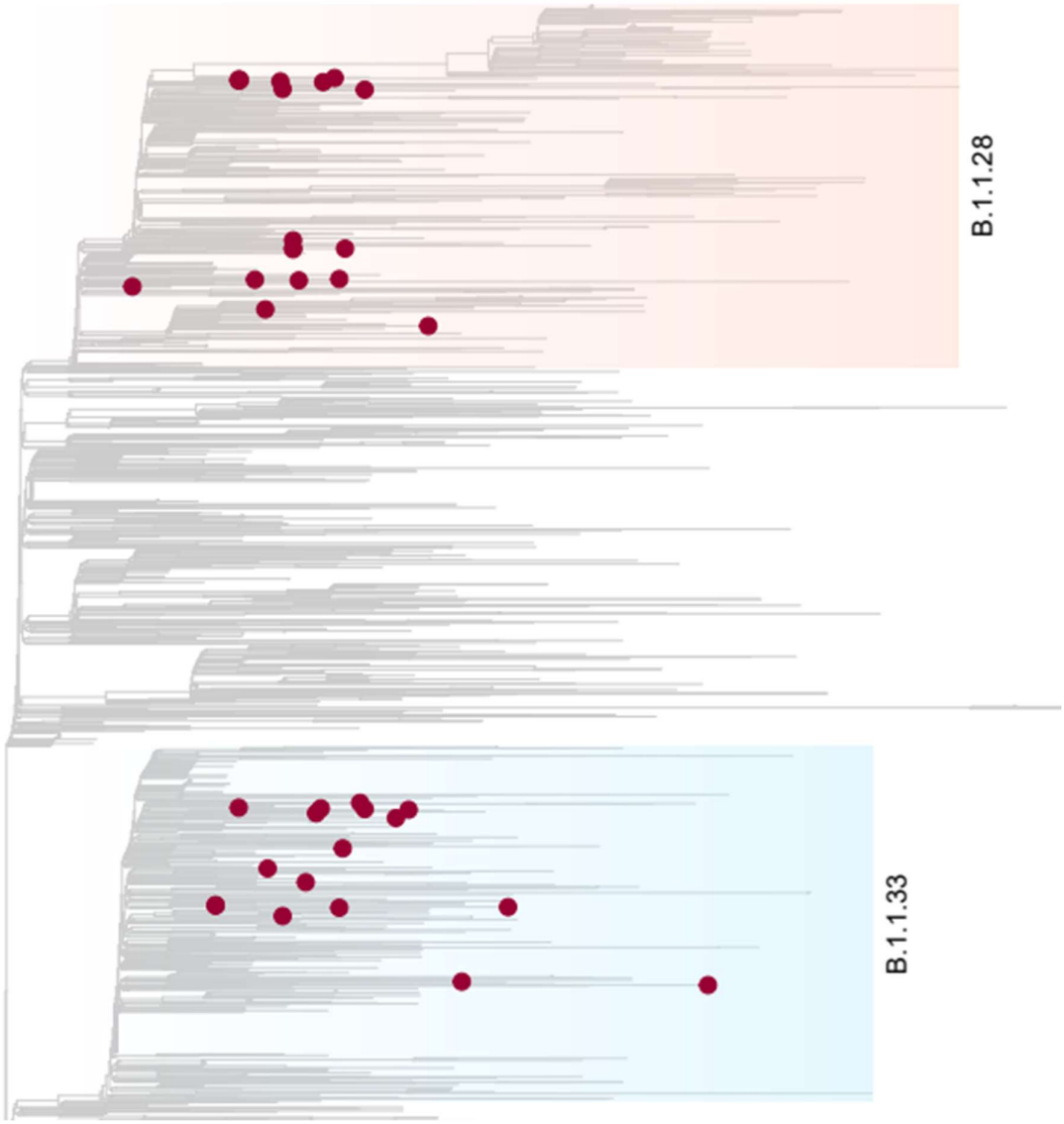

Figure 3: Spread of B.1.1.28 and B.1.1.33 lineages in Betim city. (A) Time-resolved maximum clade credibility phylogeny from a dataset comprehending 240 publicly available B.1.1.28 sequences and the 18 genomes generated in this study. **(B)** Time-resolved maximum clade credibility phylogeny from a dataset including 267 publicly available B.1.1.33 sequences and the 17 genomes generated in this study. For both analyses, the HKY+I+G4 nucleotide substitution model was used. The diamond indicates sequences from Betim city obtained in this study. The trees inferred are available on https://github.com/LBI-lab/SARS-CoV-2_phylogenies.git.

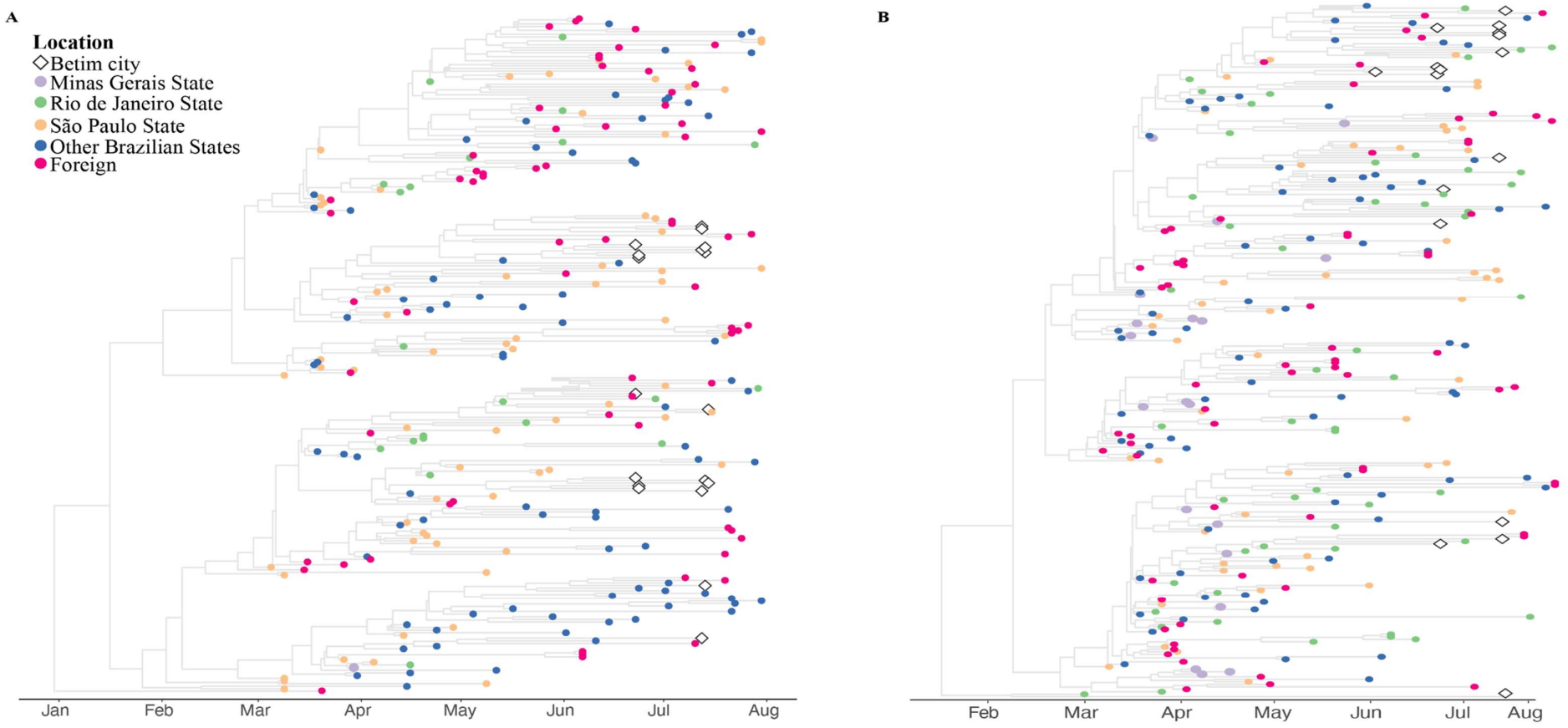

Table S1: Sequencing statistics.

ID	Raw	Paired_filtered	Unpaired_filtered	Mapped	Average_depth	Coverage
LBIb1520	790922	614298	86891	680628	2670.83	0.998562065
LBIb1756	895726	698392	96497	756729	2972.99	0.998495185
LBIb1853	595660	110010	221855	312410	996.48	0.949438202
LBIb1521	970750	684820	140045	787387	3056	0.998361423
LBIb1367	898558	520784	172820	321874	1262.4	0.998327983
LBIb1730	604232	162584	211009	118340	449543	0.992409042
LBIb1834	713926	506904	100255	519339	2027.59	0.998361423
LBIb1338	443152	218098	99202	144518	562251	0.996956929
LBIb1155	587778	276476	135914	60791	239	0.973481808
LBIb1128	589594	280462	142500	64184	250	0.976056715
LBIb1957	1096024	975312	32986	854776	4308.04	0.877407705
LBIb0013	288132	240058	3544	233727	1080.49	0.991873997
LBIb2769	555052	486824	11649	250565	1210.8	0.806046014
LBIb2626	312768	260432	8929	210031	1129.37	0.943151418
LBIb2405	931682	830622	12449	812147	4070.16	0.993077849
LBIb2427	245794	200684	9298	102113	532031	0.790730337
LBIb2421	970744	866408	19461	797757	3964.35	0.935995185
LBIb2791	410522	373574	5169	370680	2036.86	0.990001338
LBIb2224	2768498	2618126	30325	2600398	13991.7	0.993010968
LBIb2933	680422	622700	7978	613536	3254.58	0.973147405
LBIb2624	556428	511752	7062	507512	2796	0.9179374
LBIb2621	802890	747922	10676	717957	4050.13	0.977962814
LBIb2905	1291410	1067698	50017	1000586	5093.82	0.955557785
LBIb3231	3115324	2888612	39358	2854534	15183.4	0.99705725
LBIb2256	949972	892682	10741	883323	4972.17	0.97950107
LBIb2808	834356	739114	14535	723412	3741.48	0.943820225
LBIb2964	822522	722510	12942	314131	1762.05	0.934691011
LBIb3167	1060064	972588	12679	965926	5015.61	0.993846977
LBIb2674	1398626	1279472	28774	1236914	6484.77	0.9941145
LBIb1905	2295810	1661904	174732	1370724	6910.15	0.97752809
LBIb1806	2111520	1906066	61989	1885222	10040.7	0.989332531
LBIb1706	2368546	1554024	290025	1170912	6275.13	0.976290797
LBIb2296	2508976	1884492	202645	1600481	7961.24	0.991372392
LBIb2892	2369302	2159526	63471	2149739	11672.1	0.991673355
LBIb2494	2285012	2088736	52678	1919668	10072	0.986824505

Table S2: Association of clinical and epidemiological data with a positive test (serological or molecular). Bolded p values indicate $p < 0.05$.

Variable	Level	Positive	Negative	p-value
Survey	First	5 (4.7%)	1074 (34.3%)	
	Second	29 (27.4%)	1051 (33.5%)	
	Third	72 (67.9%)	1008 (32.2%)	**< 0.0001**
Administrative Regions	Alterosas	18 (17.0%)	616 (19.7%)	
	Citrolândia	4 (3.8%)	215 (6.9%)	
	Icaivera	0 (0.0%)	62 (2.0%)	
	Imbiruçu	32 (30.2%)	533 (17.0%)	
	Norte	11 (10.4%)	322 (10.3%)	
	Petrovale	0 (0.0%)	41 (1.3%)	
	PTB	8 (7.5%)	282 (9.0%)	
	Sede	15 (14.2%)	568 (18.1%)	
	Terezópolis	17 (16.0%)	302 (9.6%)	
	Vianópolis	1 (0.9%)	192 (6.1%)	**0.002424**
Sex	Female	69 (65.1%)	1559 (49.8%)	**0.002642**
	Male	37 (34.9%)	1574 (50.2%)	
Age range	0-5	3 (2.8%)	214 (6.8%)	
	06-19	15 (14.2%)	635 (20.3%)	
	20-39	42 (39.6%)	1025 (32.7%)	
	40-59	31 (29.2%)	840 (26.8%)	
	Above60	15 (14.2%)	419 (13.4%)	0.190538
International travel	No	106 (100.0%)	3119 (99.6%)	
	Yes	0 (0.0%)	14 (0.4%)	1
Fever	No	88 (83.0%)	2927 (93.4%)	
	Yes	18 (17.0%)	206 (6.6%)	**0.000075**
Cough	No	73 (68.9%)	2518 (80.4%)	
	Yes	33 (31.1%)	615 (19.6%)	**0.005304**
Sore throat	No	77 (72.6%)	2765 (88.3%)	
	Yes	29 (27.4%)	368 (11.7%)	**0.000003**
Dyspnoea	No	96 (90.6%)	3002 (95.8%)	
	Yes	10 (9.4%)	131 (4.2%)	**0.018051**
Myalgia	No	72 (67.9%)	2883 (92.0%)	
	Yes	34 (32.1%)	250 (8.0%)	**< 0.0001**
Rhinorrhea	No	70 (66.0%)	2452 (78.3%)	
	Yes	36 (34.0%)	681 (21.7%)	**0.004198**
Respiratory discomfort	No	90 (84.9%)	2961 (94.5%)	
	Yes	16 (15.1%)	172 (5.5%)	**0.000079**
Nausea/ vomit	No	94 (88.7%)	3025 (96.6%)	
	Yes	12 (11.3%)	108 (3.4%)	**0.000075**
Headache	No	50 (47.2%)	2399 (76.6%)	
	Yes	56 (52.8%)	734 (23.4%)	**< 0.0001**
Prostration	No	83 (78.3%)	2968 (94.7%)	
	Yes	23 (21.7%)	165 (5.3%)	**< 0.0001**
Diarrhea	No	95 (89.6%)	2933 (93.6%)	
	Yes	11 (10.4%)	200 (6.4%)	0.150267
Conjunctivitis	No	105 (99.1%)	3102 (99.0%)	
	Yes	1 (0.9%)	31 (1.0%)	1
Ageusia/ anosmia	No	87 (82.1%)	3051 (97.4%)	
	Yes	19 (17.9%)	82 (2.6%)	**< 0.0001**
Loss of voice	No	102 (96.2%)	3081 (98.3%)	
	Yes	4 (3.8%)	52 (1.7%)	0.206495
Pneumopathy	No	104 (98.1%)	3105 (99.1%)	
	Yes	2 (1.9%)	28 (0.9%)	0.593166
Chronic neurological disease	No	106 (100.0%)	3094 (98.8%)	
	Yes	0 (0.0%)	39 (1.2%)	0.482095
Pregnant	No	104 (98.1%)	3107 (99.2%)	
	Yes	2 (1.9%)	26 (0.8%)	0.533509
Postpartum	No	105 (99.1%)	3125 (99.7%)	
	Yes	1 (0.9%)	8 (0.3%)	0.699884
Chronic cardiovascular disease	No	106 (100.0%)	3037 (96.9%)	
	Yes	0 (0.0%)	96 (3.1%)	0.123958
Chronic kidney disease	No	103 (97.2%)	3086 (98.5%)	
	Yes	3 (2.8%)	47 (1.5%)	0.489013
Obesity	No	96 (90.6%)	3038 (97.0%)	
	Yes	10 (9.4%)	95 (3.0%)	**0.000721**
Asthma	No	102 (96.2%)	2964 (94.6%)	
	Yes	4 (3.8%)	169 (5.4%)	0.609912
Immunodepression	No	104 (98.1%)	3113 (99.4%)	
	Yes	2 (1.9%)	20 (0.6%)	0.348298
Chronic liver disease	No	106 (100.0%)	3118 (99.5%)	
	Yes	0 (0.0%)	15 (0.5%)	1
Diabetes	No	100 (94.3%)	2911 (92.9%)	
	Yes	6 (5.7%)	222 (7.1%)	0.71047
Hypertension	No	91 (85.8%)	2585 (82.5%)	0.445923

	Yes	15 (14.2%)	548 (17.5%)	
Transplanted	No	106 (100.0%)	3129 (99.9%)	1
	Yes	0 (0.0%)	4 (0.1%)	
Cancer	No	106 (100.0%)	3110 (99.3%)	0.766303
	Yes	0 (0.0%)	23 (0.7%)	
Any comorbidity	No	73 (68.9%)	2211 (70.6%)	0.787174
	Yes	33 (31.1%)	922 (29.4%)	
Sought health assistance	Hospital	8 (7.5%)	130 (4.1%)	**0.003285**
	None	81 (76.4%)	2764 (88.2%)	
	Basic Health Unit	8 (7.5%)	121 (3.9%)	
	Emergency Care Unit	9 (8.5%)	118 (3.8%)	
Admitted in health institution	No	103 (97.2%)	3098 (98.9%)	0.249183
	Yes	3 (2.8%)	35 (1.1%)	
Household contact with symptomatic person	No	71 (67.0%)	2528 (80.7%)	**0.000774**
	Yes	35 (33.0%)	605 (19.3%)	

Table S3: Acknowledgement to sequences obtained from GSAID.