GMR
The Original by FUNPEC-RP

Optimization of Illumina AmpliSeq protocol for SARS-CoV-2 and detection of circulating variants in Goiás State, Brazil from November 2020 to July 2021

C.P. Targueta[1], R.S. Braga-Ferreira[1], A.A. de Melo[1], J.S. de Curcio[2], R. Nunes[1], R.O. Dias[1], F. Mello-Andrade[3,4], D.M. Silva[5], E.P. Silveira-Lacerda[2], T.G. Castro[1], T.M.A. Pedroso[5], L.A. Pereira[6], A.F. Mendonça[6], R.M. Almeida[6], V.L. Silva[6] and M.P.C. Telles[1,7]

[1] Laboratório de Genética & Biodiversidade, Departamento de Genética, Instituto de Ciências Biológicas I, Universidade Federal de Goiás UFG, Goiânia, GO, Brasil
[2] Laboratório de Genética Molecular e Citogenética, Departamento de Genética, Instituto de Ciências Biológicas I, Universidade Federal de Goiás, Goiânia, GO, Brasil
[3] Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Goiânia, GO, Brasil
[4] Laboratório de Biotecnologia Ambiental e Ecotoxicologia, Instituto de Patologia Tropical e Saúde Pública, Universidade Federal de Goiás, Goiânia, GO, Brasil
[5] Laboratório de Mutagênese, Instituto de Ciências Biológicas I, Departamento de Genética, Universidade Federal de Goiás, Goiânia, GO, Brasil
[6] Laboratório Estadual de Saúde Pública Dr. Giovanni Cysneiros, Goiânia, GO, Brasil
[7] Escola de Ciências Médicas e da Vida, Pontifícia Universidade Católica de Goiás, Goiânia, GO, Brasil

Corresponding author: M.P.C. Telles
E-mail: tellesmpc@gmail.com

Genet. Mol. Res. 21 (1): gmr19018
Received January 24, 2022
Accepted March 02, 2022
Published March 30, 2022
DOI http://dx.doi.org/10.4238/gmr19018

ABSTRACT. The SARS-CoV-2 pandemic has demonstrated the need for genomic epidemiology surveillance. To date, various methodologies have been applied, including metagenomic

approaches and amplicon-based sequencing associated with high-throughput sequencing platforms. We adapted some details in amplicon-based sequencing using a SARS-CoV-2 community panel (Illumina AmpliSeq), with additional modifications for balanced and high-quality sequencing using the MiSeq platform. The modified protocol was used to detect circulating SARS-CoV-2 variants in Goiás state, Brazil. Initially, RNA samples were obtained from swab samples from 15 patients from the state of Goiás, Brazil, in November/2020 and February/2021 to validate protocol steps. The libraries were prepared following AmpliSeq for Illumina workflow with modifications; subsequently, we analyzed 305 positive samples collected from the state of Goiás from December 2020 to July 2021. For protocol improvement, we removed the need to treat samples with DNAse and demonstrated the importance of quantification by qPCR before and after library dilution. No fragmentation pattern was observed in the samples analyzed with Bioanalyzer. The libraries returned sequencing results that were used for genome assembly and variant detection. We were able to assemble SARS-CoV-2 genomes from 318 samples, which were used to identify 13 variants of coronavirus circulating in Goiás throughout those months. Variants of concern, such as Alpha (B.1.1.7), Gamma (P.1) and Delta (B.1.617.2) were detected; the latter was detected at first in Goiás in April 2021. The modifications in the workflow we developed were successfully applied to detect SARS-CoV-2 variants, resulting in high coverage genome assembly, and they can be used to increase the number of genome sequences and aid in epidemiological surveillance in Brazil.

Key words: AmpliSeq; Amplicon sequencing; Coronavirus; COVID-19; High-throughput sequencing; Viral genome

INTRODUCTION

The coronavirus pandemic emergence in 2019 brought the need to apply molecular tools for genomic surveillance (Okada et al., 2020). SARS-CoV-2 total genome sequencing sheds light on mutations that could define the emergence of new variants. Various methods of whole genome sequencing are available and can cover more than 99% of the SARS-CoV-2 genome (Charre et al., 2020). Advantages of whole-genome sequencing include the identification and monitoring of new viral mutations, providing information about transmission routes on a spatiotemporal scale, and improving strategies of diagnostic and epidemiological control (WHO, 2021).

However, it is necessary to establish an experimental and analysis workflow that brings reliable and effective results for monitoring new mutations in the SARS-CoV-2 genome. The World Health Organization (WHO, 2021) recommends practical planning for obtaining new genomes related to sample selection, sequencing strategy and bioinformatics analysis, in addition to data sharing for public health and scientific resources.

The first sequenced SARS-CoV-2 genomes were achieved using a metagenomic approach in early 2020 (Wu et al., 2020; Zhou et al., 2020). Since then, specific tools have been developed for amplicon or enrichment sequencing of this virus to increase the amount of data and reduce costs per sample, thereby increasing the number of genomes analyzed and coverage (Chiara et al., 2020).

Resequencing amplicon methods are based on the development of specific primers designed to amplify multiple regions throughout the genomes, followed by insertion of adapters to construct the sequencing libraries. This technique increases the amount of specific nucleic acid molecules, decreasing the need for an exhaustive sequencing effort (Chiara et al., 2020). Although this generally allows for good genome coverage, it is still important to establish a good workflow coupled with good laboratory procedures.

Since the first sequencing of the first lineage identified as SARS-CoV-2 (WIV04) (Zhou et al., 2020), thousands of genome sequences have become available in public databases, such as GISAID (Global Initiative on Sharing All Influenza Data - https://www.gisaid.org/). The main lineages of SARS-CoV-2 have been proposed from phylogenetic analyses and are classified into two major clades (A and B) (Rambaut et al., 2020). Several variants of SARS-CoV2 were described throughout the pandemic and classified into related groups. Some of these are classified as variants of concern (VOCs) by the WHO, due to their impact on global health and higher transmission rate (Cascella et al., 2021).

Retrospectively, variant Alpha (B.1.1.7) was described as the first VOC in December 2020 with its origin in the United Kingdom. During the same period, two new VOCs were reported, Beta (B.1.351) and Delta (B.1.617.2) with origins in South Africa and India, respectively. At the end of 2020 and the beginning of 2021, the variant Gamma (P.1) was identified in Brazil, and the rapid dissemination and transmission of this VOC placed the country as the epicenter of the pandemic for several months. Recently, in November 2021, the Omicron (B.1.1.529) VOC was described in many countries (WHO, www.who.int/en/activities/tracking-SARS-CoV-2-variants, Cascella et al., 2021). This rapid emergence of VOCs from SARS-CoV2 highlights the importance of genomic surveillance studies for the identification of new variants that may increase community transmission.

In Brazil, some studies of SARS-CoV-2 at the genomic level were performed aiming at identifying the circulating variants (Candido et al., 2020; Xavier et al., 2020; Botelho-Souza et al., 2021; de Souza et al., 2021; Faria et al., 2021; LaMarcus et al., 2022). In March and April 2020, about 500 samples, mostly from the states of Rio de Janeiro and São Paulo (southeast Brazil), were classified into three distinct clades; lineage B was predominant (Candido et al., 2020). In the same year, between the months of November and December, 184 samples of SARS-CoV2 collected in Manaus state (North Brazil) were sequenced, and the analysis indicated the predominance of variant Gamma (P.1) (Faria et al., 2021). Throughout 2020 until February 2021 in Brazil, the profile of the evolution of SARS-CoV2 indicated fast dissemination of the two variants Gamma (P.1) and P.2, responsible for the second wave of COVID-19 in the country. The P.2 variant gradually disseminated in the country during the months of September 2020 to January 2021, in contrast to the rapid spread of P.1 between December of 2020 to February 2021 (Wolf, 2021). Later in 2021, efforts were made to track the circulation of VOCs in Brazil.

Challenges for obtaining good genome sequences are related to better analysis of samples and libraries prior to sequencing. Therefore, we developed a methodology for

SARS-CoV-2 genome sequencing using Illumina AmpliSeq technology, standardizing important checkup points throughout the procedures. We applied the methodology to sequence 305 samples as an effort of epidemiological surveillance in central Brazil (State of Goiás, Brazil) in 2021.

MATERIAL AND METHODS

Samples used for protocol standardization

Initially, we used samples of nasal swabs from 15 human patients that showed flu symptoms in the state of Goiás, Brazil. Those patients were COVID-19 positive, tested by RT-qPCR in a partner diagnostic laboratory showing quantification cycle (Cq) values < 28. The swab samples were stored at -80ºC until RNA extraction. Of the 15 samples obtained, 13 were used for further analysis of genomic sequencing protocol optimization. These samples belong to resident patients of three cities in Goiás: five patients from Anápolis, seven from Goiânia and one from Catalão. The samples were collected in November 2020 (one sample) and February 2021 (12 samples), the youngest patient was 24 years old and the oldest 82 (Table 1; Table S1). All participants were fully informed about the procedures and the aims of the study and signed informed written consent before participation. We obtained approval for this study from the Research Ethics Committee of the Federal University of Goiás (number CAE 39289520.000.5083; reference:5.160.191). All research procedures were according to the principles of the regulatory guidelines and standards described in Resolution No. 466/12 of the National Health Council, which approves the regulatory guidelines and standards for research involving human beings in Brazil. The research was conducted in accordance with the ethics committee (4.365.579). In addition, the study was registered at Sistema Nacional de Gestão do Patrimônio Genético (SISGEN A379C84).

Table 1. Description of the samples used for protocol optimization according to the patient's municipalities, age, collection date and average of Cq for the N gene detected by RT-qPCR. Description of RNA quantification and cDNA mass used for library preparation, the average size of fragments of the library and quantification by qPCR.

Samples [a]	Brazilian municipality	Patients age (years)	Collection Date	Average Cq_N [b]	RNA quantification (ng/µL) [c]	DNA mass (ng) [d]	Library average size (bp) [e]	Library qPCR quantification (nM) [f]
Cov2-24	Anápolis	41	Feb/04/21	13.2	4.0	20.2	344	24.8
Cov2-25	Anápolis	24	Feb/03/21	14.6	5.1	20.4	310	32.6
Cov2-26	Anápolis	44	Feb/04/21	14.1	4.4	20.9	397	27.6
Cov2-29	Anápolis	29	Feb/03/21	12.8	4.6	20.9	334	9.3
Cov2-46	Goiânia	39	Feb/19/21	13.2	4.7	21.1	364	39.8
Cov2-47	Goiânia	37	Feb/20/21	17.0	6.1	21.4	348	41.7
Cov2-49	Goiânia	28	Feb/19/21	13.8	4.6	20.6	341	27.5
Cov2-51	Catalão	41	Feb/21/21	139	2.4	16.8	355	22.9
Cov2-32	Anápolis	51	Feb/04/21	15.5	2.1	14.7	368	31.1
Cov2-02	Goiânia	50	Nov/17/20	18.7	7.4	20.8	339	42.4
Cov2-44	Goiânia	55	Feb/19/21	14.0	2.5	17.6	353	8.2
Cov2-48	Goiânia	52	Feb/20/21	17.1	6.0	21.0	365	75.6
Cov2-50	Goiânia	82	Feb/20/21	19.4	2.4	16.8	365	8.2

a) sample registration code. b) value of quantification cycle for the gene encoding the nucleocapsid protein of SARS-CoV-2 with the 2019-nCov RUO kit (IDT- INTEGRATED DNA TECHNOLOGIES). c) RNA concentration value obtained with the Qubit RNA High Sensitivity Assay Kit. d) DNA concentration value obtained with the kit Qubit DNA High Sensitivity Assay Kit. e) Average of libraries sizes obtained by Bioanalyzer. Libraries molarity obtained by qPCR assay using a KAPA Library Quantification Kit (Roche).

Nucleic acids isolation

Total RNA extraction from swab samples was performed using the QIAamp Viral RNA Mini Kit (Qiagen, Germany), as described by the manufacturer with few modifications. In brief, the samples were thawed and placed at 65ºC for 1 hour for virus inactivation. Then, samples were centrifuged at 14,000 rpm for 1 minute for precipitation of the drops at the bottom of the tubes to minimize chances of contamination between samples. Each sample was prepared in duplicate, resulting in a total input of 280 uL per sample. For each replicate, the Buffer AVL and Carrier RNA were individually added as described by the protocol. Then the samples in duplicate were concentrated in a single column and RNA extraction was performed. Subsequently, the RNAs were submitted to RT-qPCR to detect SARS-CoV-2 and Cq values under 28. Singleplex qPCR was performed with hydrolysis probes 2019-nCoV RUO kit catalog no. 10006713 (Integrated DNA Technologies IDT, Iowa, USA) for two target genes of SARS-CoV-2, N1 and N2, and internal control RNase P human.

We tested DNAse treatment in two samples. In this case, we added 1 uL of 10X DNAse I Reaction Buffer and 1 uL of DNase I Amplification Grade (1 U/uL) (Invitrogen, California, USA) to 8 uL of total extracted RNA. The treated samples were incubated for 15 minutes at room temperature (22 ºC) and inactivation of DNase was achieved by heating the samples to 70ºC for 15 minutes. The inactivation was carried out using heating to avoid the addition of EDTA.

Total RNA was quantified with Qubit RNA High Sensitivity Assay Kit (Thermo Fisher Scientific, Massachusetts, USA) in Qubit and diluted to 20 ng/uL, the concentration recommended by Illumina's AmpliSeq™ for the SARS-CoV-2 panel (Illumina, California, USA).

Library preparation

Samples of total RNA were reverse transcribed to cDNA with AmpliSeq™ cDNA Synthesis for Illumina using random primers following the manufacturer's protocol. We quantified the cDNA of each sample to confirm the positive transcription to complementary DNA with Qubit dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific, Massachusetts, USA).

The cDNA was used to prepare each library for amplicon-based sequencing. We used an AmpliSeq Library PLUS kit following the manufacturer's recommendation combined with a specific SARS-CoV-2 panel (Illumina, California, USA). The SARS-CoV-2 panel was designed containing two primer pools, totaling 247 primer pairs. Pool 1 generates 125 amplicons and pool 2 generates 122 amplicons, totaling 237 viral-specific SARS-CoV-2 targets and five human gene expression controls that range from 125-275 bp in length. The samples were individually indexed using AmpliSeq™ CD Indexes (Illumina, California, USA). Each library was checked for its quality and quantity on the Agilent DNA1000 kit on Bioanalyzer (Agilent Technologies, California, USA) and qPCR with KAPA Library Quantification Kit (Roche). For the latter quantification step, the samples were serially diluted to 1:10,000 in triplicate.

The libraries that showed Bioanalyzer profiles demonstrating amplicons were diluted to 4 nM and quantified again using qPCR (Step One Plus ™ Real Time PCR

System). Indexed and pooled libraries with final concentration of 4 nM were also quantified by qPCR for sequencing in the MiSeq platform.

Genome sequencing

Pooled 4 nM libraries were diluted to a final concentration of 20 pM and 18 pM (for the first and second sequence runs, respectively) to be loaded on MiSeq Reagent v2 Nano (Illumina, California, USA), with 1% PhiX. The libraries' denaturation followed two steps. First, we added sodium hydroxide (Sigma-Aldrich, Missouri, USA) 0.2 N to the pooled library. Second, right before loading on the sequence cartridge, we heated the library to 96ºC for 2 minutes and then kept it on ice for 5 minutes. The denatured libraries were loaded and sequenced using MiSeq Reagent v2 300 cycles. The sequence was performed with 150 paired-end cycles.

Consensus calling and lineage assignment

Sequencing quality was evaluated using FastQC v0.11.4 (Andrews et al., 2010), and Trimmomatic v0.39 (Bolger et al., 2014) was used to remove low sequencing quality bases and adapter sequences (ILLUMINACLIP:NexteraPE-PE.fa:2:30:10, SLIDINGWINDOW:4:20 and MINLEN:75). High-quality reads were mapped to the SARS-CoV-2 reference genome sequence (isolate Wuhan-Hu-1, NCBI accession ID: NC_045512.2) using Burrows-Wheeler Aligner v0.7.17 (Li and Durbin, 2009). Reads were sorted by mapping coordinates and non-mapped reads were removed from the bam file using Samtools v1.10 (Li et al., 2009). Duplicated reads were removed using PicardTools v2.0.1. Variant calling, filtering and consensus sequence generation were conducted using bcftools v1.10.2 (QUAL >20 and DP >5) (Li et al., 2011). Genome consensus sequence quality was assessed using Nextclade beta v0.13.0 and assigned to SARS-CoV-2 lineages using Pangolin COVID-19 Lineage Assigner webserver (available in: https://pangolin.cog-uk.io). Genome sequencing coverage was measured using bedtools v2.30.0 (Quinlan and Hall, 2010) and plotted using the *ggplot2* R package.

Characterization of SARS-CoV-2 circulating variants in the state of Goiás

After protocol standardization, 305 positive samples from December of 2020 to July 2021 were evaluated to determine the circulating variants in the state of Goiás as described above. The samples were from 69 municipalities in Goiás, including the capital Goiânia. All samples had Cq <30 with mean Cq = 18.56 (see Table S1). The sequencing of the 305 samples was performed in four separate runs with about 60 samples each using MiSeq Reagent v2 300 cycles.

RESULTS AND DISCUSSION

Standardization protocol Illumina AmpliSeq SARS-CoV-2

In order to recover the whole genome of SARS-CoV-2 from nasal swab samples, we performed a simple protocol based on amplicons, with some modifications, using the

Illumina MiSeq platform (Figure 1). Checkpoints were defined and added throughout the initial protocol. After sequencing the first 13 samples, it was possible to assemble the coronavirus genome with an average of 95.23% and sequencing quality score above Q30, which allowed us to identify four different variants circulating in the state of Goiás, Brazil and to apply the protocol to sequence an additional 305 samples from Goiás.

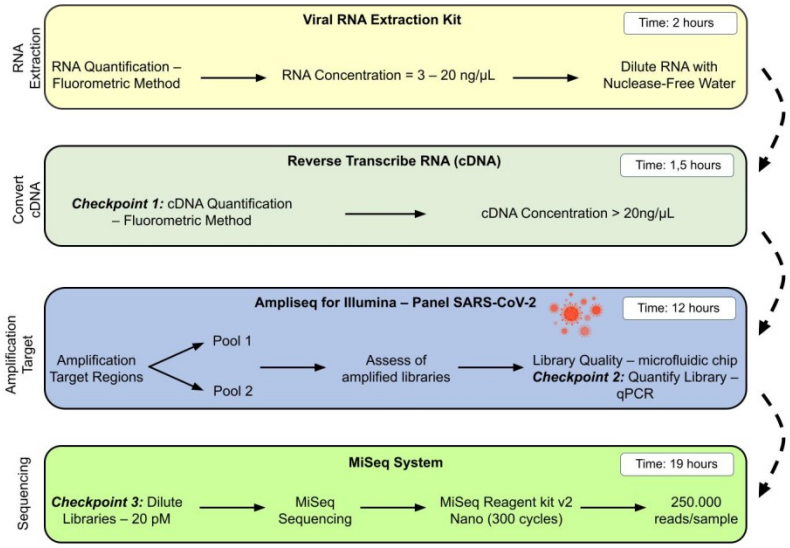

Figure 1. Workflow of adapted Amplicon based protocol for SARS-CoV-2 genome sequencing with the AmpliSeq Library Plus Illumina protocol. Checkpoints indicate important steps to obtain libraries for high-quality sequencing on MiSeq.

The workflow with the main steps and critical points are described in Figure 1. To standardize the procedure, we extracted RNA from 15 samples, although we only sequenced 13 samples (see discussion below). We were able to acquire total RNA with concentration ranging from 2.1 to 7.4 ng/uL from nasal swab samples from human patients collected in November 2020 and February 2021. Although it is known that samples collected from higher respiratory tracts, as nasal swabs, give less viral load than samples from lower respiratory tracts (Chiara et al., 2021), we optimized our amplicon-based protocol to make it possible to use samples already collected from partner diagnostic laboratories.

Sample storage and inactivation methods were efficient to maintain the integrity and give more security to handle samples after the decrease in infectivity. We chose to inactivate at 65ºC for 1 hour rather than 95ºC for 15 minutes, as the high temperature may affect RNA integrity even for diagnostic purposes (Pastorino et al., 2020). The time spent between the collection and RNA extraction may be a factor that requires caution in order to maintain RNA quality as well (Chiara et al., 2021). Although we had a sample from November, we were able to extract RNA at a concentration of 7.4 ng/uL, which allowed sequencing of this older sample.

The extracted RNA concentration ranged from 2.1 ng/ul to 7.4 ng/ul. The small quantity does not affect the efficiency of the methodology used for library construction based on amplicons because this workflow is based on producing DNA fragments of the

target region through PCR, followed by indexed adaptors ligation and amplification of this library. Higher amounts of RNA are necessary especially when using a metatranscriptomic approach, while amplicon-based approaches enable an increase from the target region, using PCR even with small RNA concentrations (Chiara et al., 2021). Also, the addition of Carrier RNA to increase the amount of RNA extracted did not affect the generation of SARS-CoV-2 genome reads, which would only affect metatranscriptomic approaches (Chiara et al., 2021).

RNA treated with DNase was one of the requirements listed as input on Illumina's AmpliSeq protocol. Therefore, we treated two samples with DNAse to deplete possible DNA contaminants. In these samples, we quantified the RNA after the treatment and in one of the treated samples, the RNA concentration decreased. Again, the use of DNAse treated samples is more important when using metatranscriptomic approaches to eliminate contamination in assembled genomes (Chiara et al., 2021). In the library quantification step, we noticed that samples treated with DNAse showed very weak fluorescence signals, while untreated samples showed strong fluorescence signals, demonstrating amplification of the desired fragments (Figure S1). For this reason, we eliminated the two treated libraries from the sequencing rounds. Although we discarded this step of the protocol, we were able to maintain good SARS-CoV-2 genome assemblage without DNAse treatment, since we achieved high quality reads using untreated RNA as well.

Following the protocol, we used about 20 ng (14.7 to 21.4 ng) of input for reverse transcription reactions of treated and untreated libraries. After the generation of cDNA, we quantified the product to check if the reaction had occurred (Checkpoint 1; Figure 1). This quantification led to a range of 17.8 to 83 ng/µL. After that, we followed the AmpliSeq (Illumina) protocol until libraries were indexed and ready for evaluation.

Quantification by qPCR is known as the preferred method for library quantification because it quantifies only DNA fragments harboring specific adapters (Checkpoint 2; Figure 1). Although it is more accurate than other methods, it is more expensive, which makes it a challenge to implement as a routine procedure. However, its use is necessary to improve the library's evaluation and cannot be discarded. In our experiments, we noticed that, for corrected quantification of libraries, qPCR is indispensable for an accurate dilution for 4 nM libraries and must be included in the protocol as a required step (critical point of the protocol). In addition, the quantification of diluted 4 nM libraries results in well-balanced and high-quality sequencing.

Since our untreated libraries showed positive amplification patterns and accurate quantifications, we chose those to perform three rounds of sequencing. We diluted the libraries to 4 nM and pooled four indexed libraries for sequence runs 1 and 2 and five indexed libraries for run 3. The pooled libraries were diluted to 20 pM or 18 pM, and denatured by sodium hydroxide 0.2 N and by heat, in order to obtain better clustering in the flow cell. Clustering and sequencing throughput with AmpliSeq libraries is increased by increasing the input concentration in MiSeq. We sequenced using MiSeq Reagent v2 Nano for up to five samples in each run. The time for sequencing on the Nano kit is described in Figure 1. A variation in sequencing reagent kit procedures can be implemented by increasing the number of samples for each run with an increase in time up to 36 hours using MiSeq Reagent v3.

The first sequenced SARS-CoV-2 genomes were determined using a metagenomic approach (e.g. Chen et al., 2020; Holshue et al., 2020; Wu et al., 2020; Yadav et al., 2020).

Faster methodologies have since been developed for quicker whole-genome sequencing using amplicons. Various groups have used amplicon methodologies with commercial panels or ARTIC primers pools (Seemann et al., 2020; Alteri et al., 2021; Giandhari et al., 2021; Thielen et al., 2021).

The sequences of the 13 libraries resulted in an average of 712,940 reads per sample (ranging from 445,160 to 964,182 reads) (Table 2). The recommendation for a good assembly of the SARS-CoV-2 genome is of at least 250,000 reads per sample. We were able to generate more than 400,000 reads per sample, maintaining good sequence quality (average 95.23% and sequencing quality score >Q30).

Table 2. Sequencing summary results of the 13 SARS-CoV-2 samples obtained for optimization of the Illumina AmpliSeq Library Plus Protocol. Results shown are: number of reads of each sample, genome coverage (average, median, minimum, and maximum) and identified lineage.

Samples [a]	Number of reads	Genome coverage (%) [b]	Average coverage [c]	Median coverage [d]	Minimum coverage [e]	Maximum coverage [f]	Identified Lineage [g]
Cov2-24	806250	94.75	246.9	198	0	1886	P.1
Cov2-25	731602	98.40	258.1	175	0	12618	B.1.1.7
Cov2-26	859530	97.43	315.3	248	0	10226	P.1
Cov2-29	688890	97.61	228.3	175	0	6931	B.1.1.7
Cov2-46	796564	96.20	263.4	200	0	8672	P.1
Cov2-47	709598	98.46	241.4	184	0	3150	P.1
Cov2-49	687486	97.18	240.4	198	0	4218	P.2
Cov2-51	680606	95.72	237.5	205	0	1871	P.1
Cov2-32	964182	98.71	454	340	0	18594	P.2
Cov2-02	445160	90.06	96.85	75	0	3485	B.1.1
Cov2-44	517718	99.23	226.3	179	0	3389	P.1
Cov2-48	643410	98.59	181.5	145	0	3560	B.1.1.7
Cov2-50	737218	98.94	300.8	248	0	3512	P.2

a) Sample registration code. b) Percentage coverage of the SARS-CoV-2 genome. c) Average coverage per base sequenced. d) Median coverage per base sequenced. e) Minimum coverage per base sequenced. f) Maximum coverage per base sequenced. g) SARS-CoV-2 lineage identified.

Coverage analysis showed that, even with more reads than recommended, we still had regions of the SARS-CoV-2 genome with very low reads (Figure 2). We noticed that the libraries had different fragment patterns of amplification on the Bioanalyzer but it was not related to the sequence coverage depth of the genome (Figure 2). Although missing regions with poor coverage depth are present in all assembled genomes, amplicons generated by the AmpliSeq SARS-CoV-2 panel were sufficient to discriminate between lineages.

In the first 13 samples, we were able to identify four SARS-COV-2 variants: P.1 (Gamma), P.2, B.1.1.7 (Alpha) and B.1.1, with the latter identified on the sample collected in November (Table 3). We identified specific mutations for each of the lineages characterized by amino acid substitutions or deletions (Table 3 and Table S2 for discrimination of all mutations). Samples from lineage P.1 showed more than 10 amino acid substitutions on gene S followed by samples from lineage B.1.1.7 with eight substitutions in this gene, including important mutations N501Y, D614G and E484K (the last one found only in the P.1 variant) (Naveca et al., 2021). Even though the samples of B.1.1.7 had fewer substitutions than P.1, they were characterized as one of the variants with the most deletions, including the characteristic del69-70 on gene S (Rambaut et al., 2020).

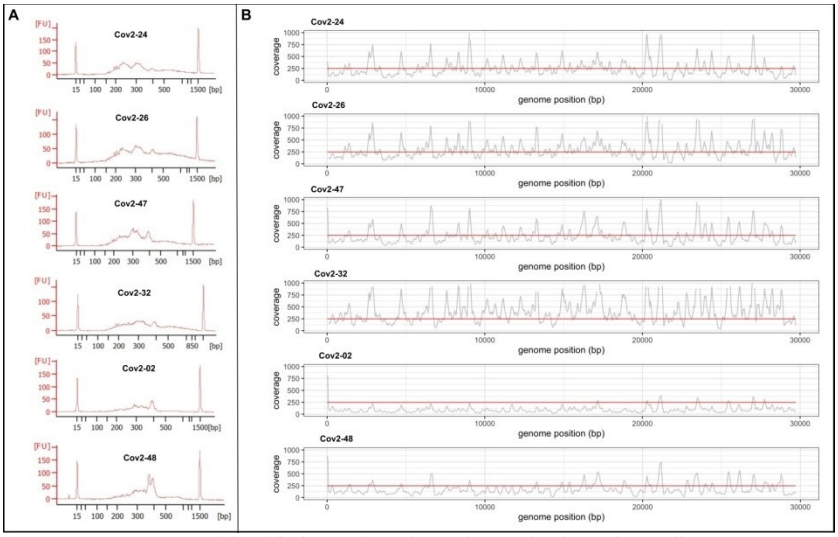

Figure 2. Most common patterns identified on Bioanalyzer for evaluation of AmpliSeq SARS-CoV-2 libraries (A). Respective coverage analysis for each of the libraries found in the left panel (B).

Table 3. Mutations identified in each of the samples and variants of the SARS-CoV-2 genome. The number of amino acid substitutions (left side of the bar) and the number of deletions (right side of the bar) are described for each ORF and gene region.

		Number of amino acid substitutions/deletions per ORF										
Sample	**Lineage**	**ORF1a**	**ORF1b**	**S**	**ORF3a**	**M**	**ORF6**	**ORF7a**	**ORF8**	**N**	**ORF9b**	**Total**
Cov2-24	P.1	1/3	1/0	10/0	2/0	-	-	-	1/0	2/0	-	17/3
Cov2-25	B.1.1.7	5/3	1/0	8/2	-	-	-	1/0	4/0	4/0	-	23/5
Cov2-26	P.1	1/3	1/0	11/0	2/0	-	-	-	1/0	4/0	1/0	21/3
Cov2-29	B.1.1.7	4/3	-	8/3	-	-	-	-	4/0	4/0	-	20/6
Cov2-46	P.1	1/3	1/0	10/0	2/0	-	-	-	1/0	4/0	1/0	20/3
Cov2-47	P.1	2/3	1/0	12/0	1/0	-	-	-	1/0	3/0	1/0	21/3
Cov2-49	P.2	3/0	1/0	3/0	1/0	-	-	-	-	5/0	1/0	14/0
Cov2-51	P.1	2/3	1/0	10/0	2/0	-	-	-	1/0	2/0	-	18/3
Cov2-32	P.2	2/0	3/0	3/0	1/0	-	-	-	1/0	5/0	-	15/0
Cov2-02	B.1.1	2/0	-	2/0	-	-	-	-	-	4/0	-	8/0
Cov2-44	P.1	2/3	3/0	12/0	1/0	1/0	-	-	1/0	4/0	1/0	25/3
Cov2-48	B.1.1.7	6/3	-	8/2	-	-	-	-	4/0	4/0	-	22/5
Cov2-50	P.2	2/0	2/0	3/0	-	-	1/0	-	-	4/0	-	14/0

Six samples were identified containing the variant P.1 (Gamma) and three the variant P.2 (Table 3 and Table S2). These variants are known to have risen independently from lineage B.1.1.28 and are recognized as being originally from the cities of Manaus (northern Brazil) and Rio de Janeiro (southeast Brazil), respectively. Later in 2020, cases of COVID-19 started to grow at a faster rate in Brazil. The second wave of the disease increased the number of cases and deaths first in Amazonas State and in early 2021, it spread all over the country. Genomic analysis revealed that a variant (first described in Japan in a traveler from Amazonas) was prevalent in the population of this state in January 2021, identified as variant P.1 (Naveca et al., 2021). This variant spread through several other countries (Di Giallonardo et al., 2021; Firestone et al., 2021; Funk et al., 2021). In the

state of Goiás, according to GISAID, the earliest collected sample of variant P.1 (Gamma) known until now is from October 26, 2020.

Lineage B.1.1 is known as the ancestor of the lineage that independently gave rise to B.1.1.28, with this latter giving rise to the variants P.1 and P.2 (Naveca et al., 2021). This ancestral lineage was identified only in one sample from November 2020. The SARS-CoV-2 genome of this sample had the fewest number of amino acid substitutions among the analyzed samples.

Three samples were recognized as the B.1.1.7 (Alpha) variant. This variant was first described in the United Kingdom. According to GISAID, the first sample of this variant in Goiás was collected January 2, 2021. Together with P.1 (Gamma), B.1.351 (Beta), B.1.1.617.2 (Delta) and B1.1.529 (Ômicron), these variants are described as VOC (Variant of Concern) as the transmission risks and disease gravity are supposed to be higher when patients are infected with them (Altmann et al., 2021).

Distribution of SARS-CoV-2 circulating variants in the state of Goiás

Access and the use of high-throughput sequencing methodologies is different within a country. For the state of Goiás, the first sequence deposited in GISAID was EPI_ISL_1181569 from a patient COVID-19 positive in May 2020. Since then, scientific and genomic surveillance efforts have advanced to assess the distribution of SARS-CoV-2 variants. From the standardization of the Illumina AmpliSeq SARS-CoV-2 protocol, it was possible to expand the genomic investigation within the state of Goiás, evaluating the main circulating variants, considering sampling carried out at the beginning of the second wave in Brazil (end of 2020 and beginning of 2021) that entailed a great impact on the number of cases and hospitalizations in Goiás.

Considering the 305 samples from December 2020 to July 2021, 13 variants were found, including three VOCs, Alpha (B.1.1.7), Gamma (P.1) e Delta (B.1.617.2) (Figure 3). The months of January (seven variants) and February (six variants) had the greatest number of variants with the presence of Alpha and Gamma. However, from February to July the Gamma variant (P.1) showed to be predominant in the state of Goiás, overlapping with the second wave of the COVID-19 pandemic in Brazil (Bastos et al., 2021).

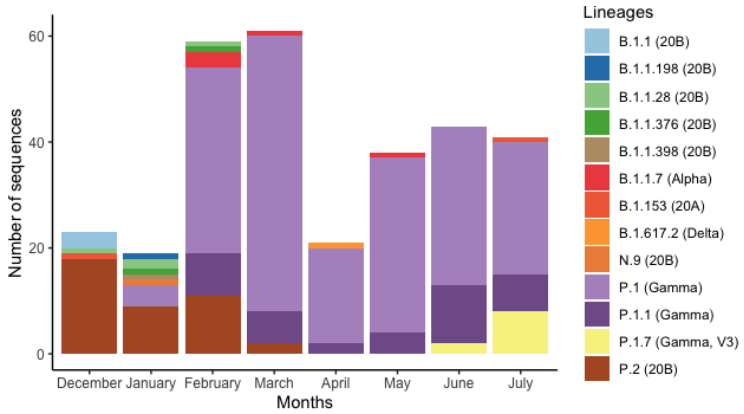

Figure 3. Number of SARS-CoV-2 genome sequences corresponding to each of the 13 lineages found by Amplicon Sequencing from December 2020 to July 2021 in patients from the state of Goiás, Brazil. All genome sequences identified as "P.1 like" during genome assembly were considered here as P.1.

Lineage P.1.7, descendent from P.1, was found only in samples from June and July, with an increase in numbers in July. The same might have happened in other Brazilian states as we can notice at the platform for Genomic surveillance epidemiologic hosted by Fiocruz/Rede Genômica (http://www.genomahcov.fiocruz.br/dashboard/). De Souza et al. (2021) also demonstrated the prevalence of P.1 lineage in Central-West Brazil until June 30, 2021. In this work, the authors accessed all sequenced samples of SARS-CoV-2 from Brazil until the first semester of 2021 on GISAID and identified higher rates of mutations on those samples. The increase in new genomic mutations is linked to higher viral transmission rates. Therefore, genomic surveillance is important to detect and monitor possible new variants circulating.

In this context, within the seasonal effort of local sampling, we detected the Delta (B.1.617.2) variant in April 2021 for the first time in Goiás. This highlights that the primers used in AmpliSeq allowed recovery of the genome of this worldwide VOC circulating in early 2021. The Delta variant appeared in late 2020 in India and is characterized by additional mutations on the protein spike at 417N and 484K.

The spread of the Delta variant in Brazil was proposed to have occurred multiple times with the first cases detected in April 2021 (Lamarca et al., 2022). The Delta sample we detected was collected on April 28, 2021 from a female who had contact with her husband who arrived from Mozambique, and also tested positive and was confirmed for the Delta variant days after, considered though an isolated case. Those cases plus a third one detected also on April 28, 2021 (GISAID) are the first cases from Delta detected in Goiás.

The detection of these variants can have a big impact on epidemiologic surveillance and the faster the detection occurs, the faster efficient decisions can be proposed to restrict virus transmission throughout the population. Therefore, we described our modified workflow using the amplicon methodology of the SARS-COV-2 panel (Illumina) that recovers high coverage genomes to access variants circulating in Brazil. We indicated modifications and check-up points that were assertive for the implementation of genomic surveillance in Goiás, Brazil (central Brazil), which brought the possibility to optimize the use and enlarge the analysis. In this way, the standardized protocol in this work can also be used for genomic surveillance in other Brazilian states, as well as in other countries.

CONCLUSIONS

We developed useful protocol modifications for amplicon-based sequencing of the SARS-CoV-2 genome for Illumina Platforms and indicated critical points for a balanced and high-quality sequencing. The amplicon methodology is recommended for surveillance of virus transmission as it can be quickly performed. We demonstrated its use for sequencing samples from Goiás, Brazil, and applied the protocol to detect variants circulating until July 2021. We were able to detect principal VOCs as Gamma (P.1), Delta (B.1.617.2) and Alpha (B.1.1.7).

ACKNOWLEDGMENTS

We thank Laboratório de Análises Clínicas e Ensino em Saúde - Universidade Federal de Goiás (LACES), Laboratório Biovida from Goiânia, Goiás, Brazil, Laboratório Estadual de Saúde Pública Dr. Giovanni Cysneiros (LACEN - GO) along with Secretaria

Estadual de Saúde do Estado de Goiás (SES-GO) and Secretaria Municipal de Saúde de Goiânia (SMS Goiânia). We thank Post-graduate programs in Biotecnologia e Biodiversidade - Rede Pró-Centro-Oeste and in Genética e Biologia Molecular – UFG and also Programa de Cooperação Acadêmica em Defesa Nacional (PROCAD) – CAPES. RN were supported by PDCTR post-doctoral fellowship from CNPq/FAPEG (Proc. CNPq: 317717/2021-9; Proc. FAPEG: 202110267000863). This work was supported by: FAPEG (Agência Financiadora Pesquisa do Estado de Goiás- 202010267000278 – 06/2020); National Institutes for Science and Technology (INCT) in Ecology, Evolution and Biodiversity Conservation (EECBio), supported by MCTIC/CNPq (proc. 465610/2014-5) and FAPEG (proc. 201810267000023); CONIF (Conselho Nacional das Instituições da Rede Federal de Educação Profissional, Científica e Tecnológica - 01/2020). The research was also supported by Instituto Federal de Goiás (IFG), Pontifícia Universidade Católica de Goiás (PUC-GO) and Universidade Federal de Goiás (UFG).

AUTHORS CONTRIBUTIONS

CPT: Conceptualization, Methodology, Investigation, Resources, Writing - Original Draft, Project administration. RSBF: Conceptualization, Methodology, Investigation, Resources, Writing - Original Draft, Project administration. AAM: Conceptualization, Methodology, Investigation, Writing - Original Draft. JSC: Methodology, Investigation, Data Curation, Writing - Original Draft. RN: Conceptualization, Software, Formal analysis, Investigation, Data Curation. ROD: Conceptualization, Software, Formal analysis, Investigation, Data Curation. FMA: Conceptualization, Resources, Supervision, Funding acquisition. DMS: Conceptualization, Supervision. EPSL: Conceptualization, Supervision. TGC: Resources. TMAP: Resources. LAP: Resources, Sample screening. AFM: Resources, Sample screening. RMA: Resources, Sample screening. VLS: Resources, Sample screening. MPCT: Conceptualization, Methodology, Investigation, Resources, Supervision, Project administration, Funding acquisition.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Alteri C, Cento V, Piralla A, Costabile V, et al. (2021). Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. *Nat. Commun.* 12: 434. DOI: http://doi.org/10.1038/s41467-020-20688-x.

Altmann DM, Boyton RJ and Beale R (2021). Immunity to SARS-CoV-2 variants of concern. *Science*. 371: 1103-1104. DOI: https://doi.org/10.1126/science.abg7404.

Andrews S (2010). FastQC: a quality control tool for high throughput sequence data. Available at [http://www.bioinformatics.babraham.ac.uk/projects/fastqc].

Bastos LSL, Ranzani OT, Souza TML, Hamacher S, et al. (2021). COVID-19 hospital admissions: Brazil's first and second waves compared. *Lancet Respir. Med.* 9: e82-e83. DOI: https://doi.org/10.1016/S2213-2600(21)00287-3.

Bolger AM, Lohse M and Usadel B (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 30: 2114-2120. DOI: https://doi.org/10.1093/bioinformatics/btu170.

Botelho-Souza LF, Nogueira-Lima FS, Roca TP, Naveca FG, et al. (2021). SARS-CoV-2 genomic surveillance in Rondônia, Brazilian Western Amazon. *Sci. Rep.*11: 3770. DOI: https://doi.org/10.1038/s41598-021-83203-2.

Candido DS, Claro IM, de Jesus JG, de Souza WM, et al. (2020). Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*. 369: 1255-1260. DOI: https://doi.org/DOI: 10.1126/science.abd2161.

Cascella M, Rajnik M, Aleem A, Dulebohn SC, et al. (2021). Features, Evaluation, and Treatment of Coronavirus (COVID-19). In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; PMID: 32150360.

Charre C, Ginevra C, Sabatier M, Regue H, et al. (2020). Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterization. *Virus Evol.* 6: veaa075. DOI: https://doi.org/10.1093/ve/veaa075.

Chen J, Hilt EE, Li F, Wu H, et al. (2020). Epidemiological and genomic analysis of SARS-CoV-2 in 10 patients from a Mid-Sized City outside of Hubei, China in the early phase of the COVID-19 outbreak. *Frontiers in Public Health* 8: 567621. DOI: http://doi.org/10.3389/fpubh.2020.567621.

Chiara M, D'Erchia AM, Gissi C, Manzari C, et al. (2021). Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. *Brief. Bioinform.* 22(2): 616-630. DOI: https://doi.org/10.1093/bib/bbaa297.

de Souza UJB, dos Santos RN, Campos FS, Lourenço KL, et al. (2021). High Rate of Mutational Events in SARS-CoV-2 Genomes across Brazilian Geographical Regions, February 2020 to June 2021. *Viruses*. 13 (9): 1806. DOI: 10.3390/v13091806.

Di Giallonardo F, Puglia I, Curini V, Cammà C, et al. (2021). Emergence and Spread of SARS-CoV-2 Lineagens B.1.1.7 and P.1 in Italy. *Viruses*. 13: 794. DOI: http://doi.org/10.3390/v13050794.

Faria NR, Claro IM, Candido D, Franco LAM, et al. (2021). Genomic characterization of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological.* https://virological.org/t/genomic-characterisation-of- an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586, (accessed 23 January 2021).

Firestone MJ, Lorentz AJ, Meyer S, Wang X, et al. (2021). First identified cases of SARS-CoV-2 variant P.1 in the United States - Minnesota, January 2021. *Morb. Mortal. Wkly. Rep.* 70: 346-347.

Funk T, Pharris A, Spiteri G, Bundle N, et al. (2021). Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. *Eurosurveillance*. 26(16): 2100348. DOI: http://doi.org/10.2807/1560-7917.ES.2021.26.16.2100348.

Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO.

Giandhari J, Pillay S, Wilkinson E, Tegally H, et al. (2021). Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. *Int. J. Infect. Dis.* 103: 234-241.

Holshue ML, DeBolt C, Lindquist S, Lofy KH, et al. (2020). First case of 2019 novel coronavirus in the United States. *The New England J. Med*. 382: 929-936. DOI: http://doi.org/10.1056/NEJMoa2001191.

Lamarca AP, de Almeida LGP, Francisco-Junior RS, Cavalcante L, et al. (2022). Genomic surveillance tracks the first community outbreak of the SARS-CoV-2 Delta (B.1.617.2) variant in Brazil. *J. Virol.* 96: e0122821. DOI: https://doi.org/10.1128/JVI.01228-21.

Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-60. DOI: https://doi.org/10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25: 2078-2079. DOI: https://doi.org/10.1093/bioinformatics/btp352.

Li H (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics*. 27: 2987–2993. DOI: https://doi.org/10.1093/bioinformatics/btr509.

Naveca F, da Costa C, Nascimento V, Souza V, et al. (2021). Three SARS-CoV-2 reinfection cases by the new variant pf concern (VOC) P.1/501Y.V3. *Preprint from Research Square*. DOI: http://doi.org/10.21203/rs.3.rs-318392/v1.

Okada P, Buathong R, Phuygun S, Thanadachakul T, et al. (2020). Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from Wuhan to Thailand, January 2020. *Euro Surveill.* 25: 2000097. DOI: https://doi.org/10.2807/1560-7917.ES.2020.25.8.2000097.

Pastorino B, Touret F, Gilles M, Lamballerie X, et al. (2020). Heat inactivation of different types of SARS-CoV-2 samples: what protocols for biosafety, molecular detection and serological diagnostics. *Viruses*. 12: 735. DOI: doi:10.3390/v12070735.

Quinlan AR and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26: 841-842. DOI: https://doi.org/10.1093/bioinformatics/btq033.

Rambaut A, Holmes EC, O'Toole Á, Hill V, et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5: 1403-1407. DOI: https://doi.org/10.1038/s41564-020-0770-5.

Seemann T, Lane CR, Sherry NL, Duchene S, et al. (2020). Tracking the COVID-19 pandemic in Australia using genomics. *Nat. Commun.* 11: 4376. DOI: http://doi.org/10.1038/s41467-020-18314-x.

Thielen PM, Wohl S, Mehoke T, Ramakrishnan S, et al. (2021). Genomic diversity of SARS-CoV-2 during early introduction into the Baltimore-Washington metropolitan area. *JCI Insight.* 6: e144350, DOI: http://doi.org/10.1172/jci.insight.144350.

Wolf JM, Kipper D, Borges GR, Streck AF, et al. (2021). Temporal spread and evolution of SARS-CoV-2 in the second pandemic wave in Brazil. *J. Med. Virol.* 2021: 10.1002/jmv.27371. DOI: 10.1002/jmv.27371.

Wu F, Zhao S, Yu B, Chen Y, et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature.* 579: 265-285. DOI: http://doi.org/10.1038/s41586-020-2008-3.

Xavier J, Giovanetti M, Adelino T, Fonseca V, et al. (2020). The ongoing COVID-19 epidemic in Minas Gerais, Brazil: insights from epidemiological data and SARS-CoV-2 whole genome sequencing. *Emerg. Microbes Infect.* 9: 1824-1834. DOI: https://doi.org/10.1080/22221751.2020.1803146.

Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, et al. (2020). Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J. Med. Res.* 151: 200-209.

Zhou P, Lou Yang X, Wang XG, Hu B, et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature.* 579: 270-273. DOI: https://doi.org/10.1038/s41586-020-2012-7.