



# Whole-genome sequencing of SARS-CoV-2: Comparison of target capture and amplicon single molecule real-time sequencing protocols

Florence Nicot, Pauline Trémeaux, Justine Latour, Nicolas Jeanne, Noémie Ranger, Stéphanie Raymond, Chloé Dimeglio, Gérald Salin, Cécile Donnadieu, Jacques Izopet

## ► To cite this version:

Florence Nicot, Pauline Trémeaux, Justine Latour, Nicolas Jeanne, Noémie Ranger, et al.. Whole-genome sequencing of SARS-CoV-2: Comparison of target capture and amplicon single molecule real-time sequencing protocols. *Journal of Medical Virology*, 2023, 95 (1), pp.e28123. 10.1002/jmv.28123 . hal-03988344

**HAL Id: hal-03988344**

**<https://hal.science/hal-03988344>**

Submitted on 14 Feb 2023



**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Whole-genome sequencing of SARS-CoV-2: Comparison of target capture and amplicon single molecule real-time sequencing protocols

Florence Nicot<sup>1</sup>  | Pauline Trémeaux<sup>1</sup> | Justine Latour<sup>1</sup> | Nicolas Jeanne<sup>1</sup> |  
Noémie Ranger<sup>1</sup> | Stéphanie Raymond<sup>1,2</sup> | Chloé Dimeglio<sup>1,2</sup>  | Gérald Salin<sup>3</sup> |  
Cécile Donnadieu<sup>3</sup> | Jacques Izopet<sup>1,2</sup>

<sup>1</sup>Virology Laboratory, Toulouse University Hospital, Toulouse, France

<sup>2</sup>Toulouse Institute for Infectious and Inflammatory Diseases (INFINITY), INSERM UMR 1291 – CNRS UMR 5051, Toulouse, France

<sup>3</sup>Genotoul-Genome & Transcriptome—Plateforme Génomique (GeT-PlaGe), US INRAE 1426, Castanet-Tolosan, France

## Correspondence

Florence Nicot, Laboratoire de Virologie, IFB, Hôpital Purpan, 330, Ave de Grande-Bretagne, Cedex 9, 31059 Toulouse, France.  
Email: nicot.f@chu-toulouse.fr

## Abstract

Fast, accurate sequencing methods are needed to identify new variants and genetic mutations of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome. Single-molecule real-time (SMRT) Pacific Biosciences (PacBio) provides long, highly accurate sequences by circular consensus reads. This study compares the performance of a target capture SMRT PacBio protocol for whole-genome sequencing (WGS) of SARS-CoV-2 to that of an amplicon PacBio SMRT sequencing protocol. The median genome coverage was higher ( $p < 0.05$ ) with the target capture protocol (99.3% [interquartile range, IQR: 96.3–99.5]) than with the amplicon protocol (99.3% [IQR: 69.9–99.3]). The clades of 65 samples determined with both protocols were 100% concordant. After adjusting for  $C_t$  values, S gene coverage was higher with the target capture protocol than with the amplicon protocol. After stratification on  $C_t$  values, higher S gene coverage with the target capture protocol was observed only for samples with  $C_t > 17$  ( $p < 0.01$ ). PacBio SMRT sequencing protocols appear to be suitable for WGS, genotyping, and detecting mutations of SARS-CoV-2.

## KEYWORDS

long read sequencing, SARS-CoV-2 genotyping, SMRT sequencing, whole-genome sequencing

## 1 | INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a member of the *Coronaviridae* family that caused the COVID-19 pandemic,<sup>1,2</sup> has a single-stranded positive-sense RNA genome of approximately 29.9 kb that encodes several proteins, including the spike (S) structural protein.<sup>3</sup> The 1273 amino-acid long S protein

attaches the virus to the host cell receptor via its receptor binding domain (RBD, residues 319–529).<sup>4</sup> The virus genome accumulates mutations that are associated with transmissibility, escape to neutralizing monoclonal antibodies (mAbS), and virulence.<sup>5</sup> The SARS-CoV-2 genome has diverged during the pandemic to produce several variants (clades and lineages) that differ in their biology and/or geographical distribution. Five of these variants have been

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Journal of Medical Virology* published by Wiley Periodicals LLC.



classified as variants of concern<sup>6</sup>: Alpha (B.1.1.7),<sup>7</sup> Beta (B.1.351),<sup>8</sup> Gamma (P.1),<sup>9</sup> and more recently Delta (B.1.617.2), and Omicron (B.1.1.529).<sup>10,11</sup>

The rapid identification of variants that are transmitted most efficiently or that escape the host immune response is essential for genomic surveillance and clinical management. Next-generation sequencing (NGS) protocols, mostly based on Illumina and Oxford Nanopore Technologies (ONT) platforms,<sup>12–14</sup> have been developed to study the genomic diversity of SARS-CoV-2 worldwide. Metagenomic NGS was used to sequence the complete SARS-CoV-2 genome early in the pandemic.<sup>15</sup> Since then, multiplex amplicon or hybrid capture-based methods have been developed to run on the Illumina<sup>16–19</sup> and ONT platforms.<sup>16,20</sup> Illumina sequencing devices are commonly used; they generate accurate, high-quality sequences. ONT is based on long-read sequencing and offers real-time sequencing, but the sequences may contain substantial errors.<sup>21</sup> The Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing system provides long, highly accurate sequences by using circular consensus sequencing reads; this feature could provide accurate whole-genome sequences (WGS). PacBio SMRT sequencing has already been used to sequence the S gene<sup>22–25</sup> but no data are available for the whole SARS-CoV-2 genome.

This study assesses the performance of a newly-available complete end-to-end kit of a target capture SMRT sequencing protocol for genotyping of SARS-CoV-2 and detecting mutations. Results were compared with those obtained with a 1.2 kb amplicon SMRT sequencing protocol.

## 2 | MATERIALS AND METHODS

### 2.1 | Samples

We sequenced 84 nasopharyngeal samples from patients who were SARS-CoV-2 RNA positive (N gene cycle thresholds,  $C_t$  8–28) and stored them at  $-80^{\circ}\text{C}$  in the Virology Laboratory at Toulouse University Hospital. The  $C_t$  values were obtained with the QuantStudio™ 5 Real-Time PCR system (Applied Biosystems). A total of 19 were taken between January 19 and March 19, 2021, during the Alpha wave in France, and 65 were taken between July 18 and August 8, 2021, at the beginning of the Delta wave in France. Two synthetic positive RNA controls (Twist Bioscience Control 14 [TB14] England/205041766/2020 [lineage B.1.1.7] and Twist Bioscience Control 17 [TB17] Japan/IC-0564/2021 [lineage P.1]), one strain of SARS-CoV-2 B.1.1.254/20B (EPI\_ISL\_804374 from a culture supernatant), a negative template control (NTC), and one SARS-CoV-2 negative sample were also analyzed.

### 2.2 | SARS-CoV-2 RNA extraction

Virus RNA was extracted from a 180  $\mu\text{l}$  transport medium with the MGIEasy Nucleic Acid Extraction Kit on the MGI SP 960

system (Beijing Genome Institute) according to the manufacturer's instructions.

### 2.3 | Target capture PacBio SMRT sequencing

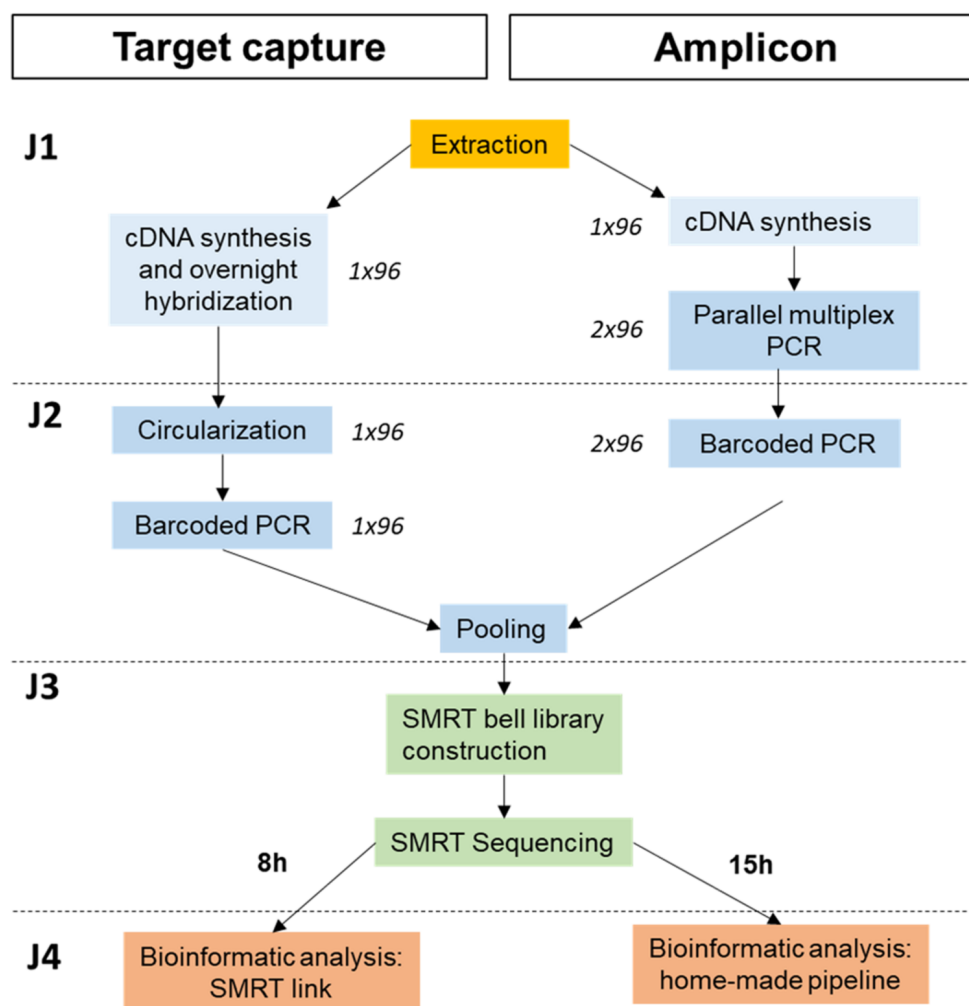
We used the SARS-CoV-2 Enrichment Early Access (PacBio) Kit following the PacBio HiFiViral for SARS-CoV-2 workflow, High-Throughput Multiplexing, 800 bp Target Capture for Full-Viral Genome Sequencing of SARS-CoV-2 (hereafter referred to as « target capture »). This protocol uses molecular inversion probes with a 675 bp target insert. The first step was simultaneous complementary DNA (cDNA) synthesis and probe hybridization. The 8  $\mu\text{l}$  RT-hybridization reaction mixture contained 6  $\mu\text{l}$  RNA, 1.6  $\mu\text{l}$  RT mix, and 0.4  $\mu\text{l}$  probe mix. The cycle steps were: 10 min/ $25^{\circ}\text{C}$ , 50 min/ $50^{\circ}\text{C}$ , 1 min/ $95^{\circ}\text{C}$ , and 16 h/ $55^{\circ}\text{C}$ . Fill-in mix (2  $\mu\text{l}$ ) was then added to each sample ( $55^{\circ}\text{C}/60$  min), followed by clean-up mix (2  $\mu\text{l}$ ) for 60 min/ $45^{\circ}\text{C}$ , 3 min/ $95^{\circ}\text{C}$ , and hold  $4^{\circ}\text{C}$ . The second step was circularization. The cDNA amplification mixture, 24  $\mu\text{l}$  containing 9.6  $\mu\text{l}$  clean-up reaction, 12  $\mu\text{l}$  polymerase chain reaction (PCR) mix, and 2.4  $\mu\text{l}$  asymmetric barcoded M13 Primer mix were subjected to 3 min/ $95^{\circ}\text{C}$ , 26 cycles of  $98^{\circ}\text{C}/15$  s,  $55^{\circ}\text{C}/15$  s, and  $72^{\circ}\text{C}/90$  s. Samples were pooled for library construction: aliquots (5  $\mu\text{l}$ ) of each sample were pooled in DNA LoBind tubes and purified with 1.3X AMPure PacBio beads (Pacific Bioscience), and the cDNA was quantified with the Quantifluor DSDNA system running on a Roche LC480 instrument. A SMRT bell library was prepared and sequenced with the SMRTbell Express Template Prep 2.0 Kit according to the manufacturer's instructions running on a Sequel IIe platform (Genotoul platform; GeTPlaGe) loading 200 pM.

Bioinformatic analysis was done with SMRT Link software using the HiFiViral SARS-CoV-2 Analysis Application with the default parameters ([https://www.pacb.com/wp-content/uploads/SMRT\\_Link\\_User\\_Guide\\_v10.2.pdf](https://www.pacb.com/wp-content/uploads/SMRT_Link_User_Guide_v10.2.pdf)), except for the minimum base coverage, which was changed to 10 reads (ECDC recommendations for SARS-CoV-2 sequencing<sup>26</sup>). Consensus sequences were then analyzed with Pangolin lineages (v3.1.20, <https://cov-lineages.org/resources/pangolin.html>) and Nextstrain clades (v1.11.0, <https://nextstrain.org/sars-cov-2>). Finally, a customized python script (v3.8.8) analysis was used to generate a user-friendly report, including the number of mapped reads, median coverage, list of S gene mutations (substitutions, insertions, and deletions), S gene missing positions, and previously found clades and lineages. Target capture protocol steps are summarized in Figure 1.

### 2.4 | 1.2 kb amplicons PacBio SMRT sequencing

We used the PacBio HiFiViral for SARS-CoV-2 workflow, High-Throughput Multiplexing 1.2 kb Amplicons for Full-Viral Genome Sequencing of SARS-CoV-2, based on 29 1.2 kb amplicons across the SARS-CoV-2 genome (hereafter referred to as « amplicon »). We first generated cDNAs. Aliquots (20  $\mu\text{l}$ ) of RT reaction mixture containing





**FIGURE 1** Target capture and amplicon protocol workflows for SARS-CoV-2 sequencing. cDNA, complementary DNA; PCR, polymerase chain reaction; SMRT, single-molecule real-time.

10 µl RNA, 1.9 µl nuclease-free water, 2 µl Superscript IV VILO (Life Technologies), and 0.1 µl of random hexamer oligo(dT) primers (100 µM) were incubated. Cycle steps: 10 min/23°C, 60 min/50°C, and 10 min/80°C. The 29 amplicons were amplified by two multiplex PCR (with two primer pools, Pool 1 and Pool 2) in 25 µl of reaction mixture each: 5 µl cDNA, 0.5 µl Q5 Hot Start High Fidelity DNA polymerase, 12 µl nuclease-free water, 1.5 µl of Pool 1 or 2 M13 tailed primers, 1 µl 10 nM deoxynucleoside triphosphates and 5 µl Q5 Reaction Buffer. The cycle steps were denaturation (98°C/30 s), amplification, and extension (35 cycles of 98°C/15 s, 65°C/5 min). A second PCR was then performed using barcoding primers tailed with the universal M13 sequence. Reaction mixture: 12.5 µl Kapa HiFi HotStart ReadyMix, 4.5 µl nuclease-free water, 5 µl of M13 forward and reverse barcoded primer mixture, and 3 µl aliquots of first-round PCR products for Pool 1 or Pool 2. Cycling conditions: 98°C/3 min, 3 cycles of 98°C/30 s, 60°C/15 s, and 72°C/1 min, then 21 cycles of 98°C/20 s, 65°C/15 s, 72°C/1 min, final extension 72°C/5 min. Samples were pooled for library construction: 1 µl of each sample from Pool 1 and Pool 2 was transferred in DNA LoBind tubes, purified with AMPure PacBio beads (Pacific Bioscience) at 0.60X, and

quantified with the Quantifluor DSDNA system running on a Roche LC480 instrument. SMRT bell libraries were constructed by pooling the barcoded samples. Barcoded amplicon libraries were prepared and sequenced with SMRTbell Express Template Prep 2.0 Kits according to the manufacturer's instructions on a Sequel IIe platform (Toulouse University Hospital).

Bioinformatic analysis was done with the Snakemake pipeline for complete genome construction. This starts by demultiplexing HiFi reads with lima (v.2.2.0, <https://github.com/PacificBiosciences/barcoding>), then creates the VCF file from pbAA clusters (PacBio tool, v.0.1.3 <https://github.com/PacificBiosciences/pbAA>), which is used, along with the samtools (v1.12) depth file, to build the consensus sequence (CoSA, coronavirus Sequence Analysis, v9.0.0, <https://github.com/Magdoll/CoSA>). A minimum base coverage of 10 reads was defined to report a position. The consensus sequences were then analyzed with Pangolin (v3.1.20, <https://cov-lineages.org/resources/pangolin.html>) for lineages and Nextstrain (v1.11.0, <https://nextstrain.org/sars-cov-2>) for clades. Our python script (v3.8.8) in-house analysis was used to generate a user-friendly report including the number of mapped reads, the median coverage, a list of





S gene mutations (substitutions, insertions, and deletions), S gene missing positions, and the previously found clades and lineages. Amplicon protocol steps are summarized in Figure 1.

See Supporting Information: Table 1 for GISAID accession numbers.

## 2.5 | Statistical analysis

Categorical variables were tested by a  $\chi^2$  test. Continuous variables were tested by a Wilcoxon sign-rank sum test. Multivariable logistic regression models were used to assess the impact of the protocol and  $C_t$  values on genome coverage.  $p$ -values of  $<0.05$  were considered to be significant.

## 3 | RESULTS

### 3.1 | Clades and lineages from full-length genomes determined by PacBio SMRT target capture

The complete genomes of positive controls (TB14, TB17, and EPI\_ISL\_804374) were covered 86.4%, 92.8%, and 99.6%, respectively, with a median read depth of 33, 33, and 1137. Fewreads ( $n = 16$ ) were detected in the NTC and the SARS-CoV-2 negative sample ( $n = 11$ ) without genome coverage. Samples 26 and 82 were tested in triplicate with N gene  $C_t$  of 14.4 and 14.6. Read numbers (sample 26: 16 993–32 327, sample 82: 48 382–67 576), median read depth (Sample 26: 355–662, Sample 82: 945–1413), and genome coverage (Sample 26:  $99.6 \pm 0.1\%$ , Sample 82: 99.6%) were similar for each triplicate. Each of the triplicate analyses found the same clades and lineages (Sample 26: 20I (Alpha, V1)/B.1.1.7; sample 82: 21J (Delta)/AY125). S gene mutation profiles were also identical. We sequenced 84 samples (median  $C_t$ : 17.6 [interquartile range, IQR: 14.2–22.5]). Sequencing failed for five samples (6%). Sequence data were obtained for 79 samples with a median read number per sample of 13 693 [IQR: 1066–27 338], and a median read depth of 231 [IQR: 19–568] (Figure 2A and Supporting Information: Table S1). The median genome coverage was 98.9% [IQR: 85.9–99.5], and  $>95\%$  for 53 (63%) samples. Most of the strains were clade 21J (Delta) (36; 45%) and clade 20I (Alpha, V1) (29; 37%). Eleven samples (14%) were determined only to the clade level but their complete genome coverage was too low to determine the lineage. Sample 16 (genome coverage of 30.9%) was flagged by SMRT link analysis as having multiple strains (probability  $>0.95$ ), indicating possible sample crossover.

### 3.2 | Clades and lineages from full-length genomes determined by PacBio SMRT amplicon

We also sequenced the 84 samples with the amplicon protocol (Supporting Information: Table S1). Sequencing failed for 18 samples

(21.4%). Sequence data were obtained for 66 samples, with a median read number per sample of 8036 [IQR: 2131–20 597] and a median read depth of 881 [IQR: 411–1289] (Figure 2B and Supporting Information: Table S1). The median genome coverage was 99.3% [IQR: 69.9–99.3], and  $>95\%$  for 41 (62%) samples. The strains were clade 21J (Delta) (30; 45%) and clade 20I (Alpha, V1) (26; 39%). Sixteen samples (24%) were determined only to the clade level but their complete genome coverage was too low to determine the lineage. No sample crossover was observed.

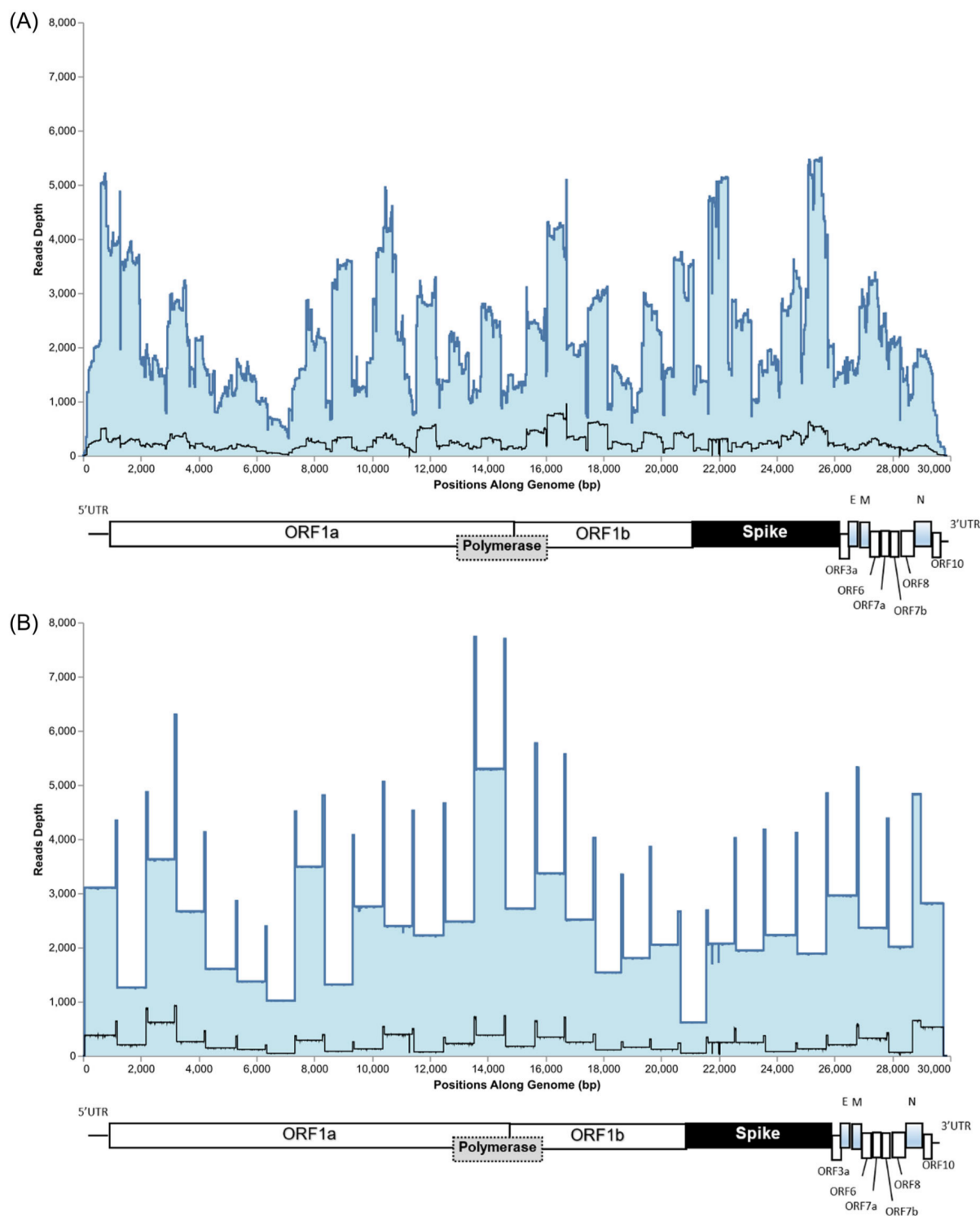
### 3.3 | Comparison of the complete genome sequences obtained with the two methods

Extraction and library preparation required 3 days for both protocols (Figure 1). The target capture protocol used a single plate for 96 samples throughout the whole process, whereas the amplicon protocol needed two plates for two steps. The target capture protocol sequencing run is shorter (8 h) than that of the amplicon protocol (15 h). Sequencing failed for more samples ( $p < 0.01$ ) with the amplicon ( $n = 18$ ) than with the target capture protocol ( $n = 5$ ). Four samples failed to sequence with both protocols (2/4 with  $C_t > 25$ ), 14 samples with the amplicon protocol only (5/14 with  $C_t > 25$ ), and sample 32 with the target capture protocol only ( $C_t = 27$ ). Among the 14 that failed to sequence with the amplicon protocol, 10 (71%) were sequenced with a genome coverage  $>74\%$  with the target capture protocol, allowing clade and lineage assignment. Sample 32 was sequenced with a genome coverage of 63% with the amplicon protocol allowing only clade assignment. The median genome coverage compared of 65 samples was significantly higher ( $p < 0.05$ ) for the target capture protocol (99.3 [IQR: 96.3–99.5]) than for the amplicon protocol (99.3 [IQR: 69.9–99.3]). Genome coverage was  $>95\%$  for 52 (80%) samples analyzed by the target capture protocol and for 41 (63%) samples by the amplicon protocol ( $p < 0.05$ ). Genome coverage for samples with high N gene  $C_t$  decreased for both protocols (Figure 3A). The clades of 65 samples were determined with both protocols, with 100% concordant results. The lineage of 49 samples was determined with both protocols with 98% concordant results. Sample 44 were determined AY.43 with the amplicon protocol (genome coverage = 69.7%) and AY.4 with the target capture protocol (genome coverage = 90.4%).

### 3.4 | Comparison of the spike sequences obtained with the two methods

The target capture protocol failed to sequence the S gene in 7/84 (8%) samples and the amplicon protocol in 19/84 (23%) samples ( $p < 0.01$ ). The median S gene coverage compared to 64 samples was significantly higher ( $p < 0.05$ ) for the target capture protocol (100 [IQR: 100–100]) than for the amplicon protocol (100 [IQR: 52–100]). S gene coverage was  $>95\%$  for 62 samples (81%) analyzed with the target capture protocol and for 44 samples (68%) with the amplicon



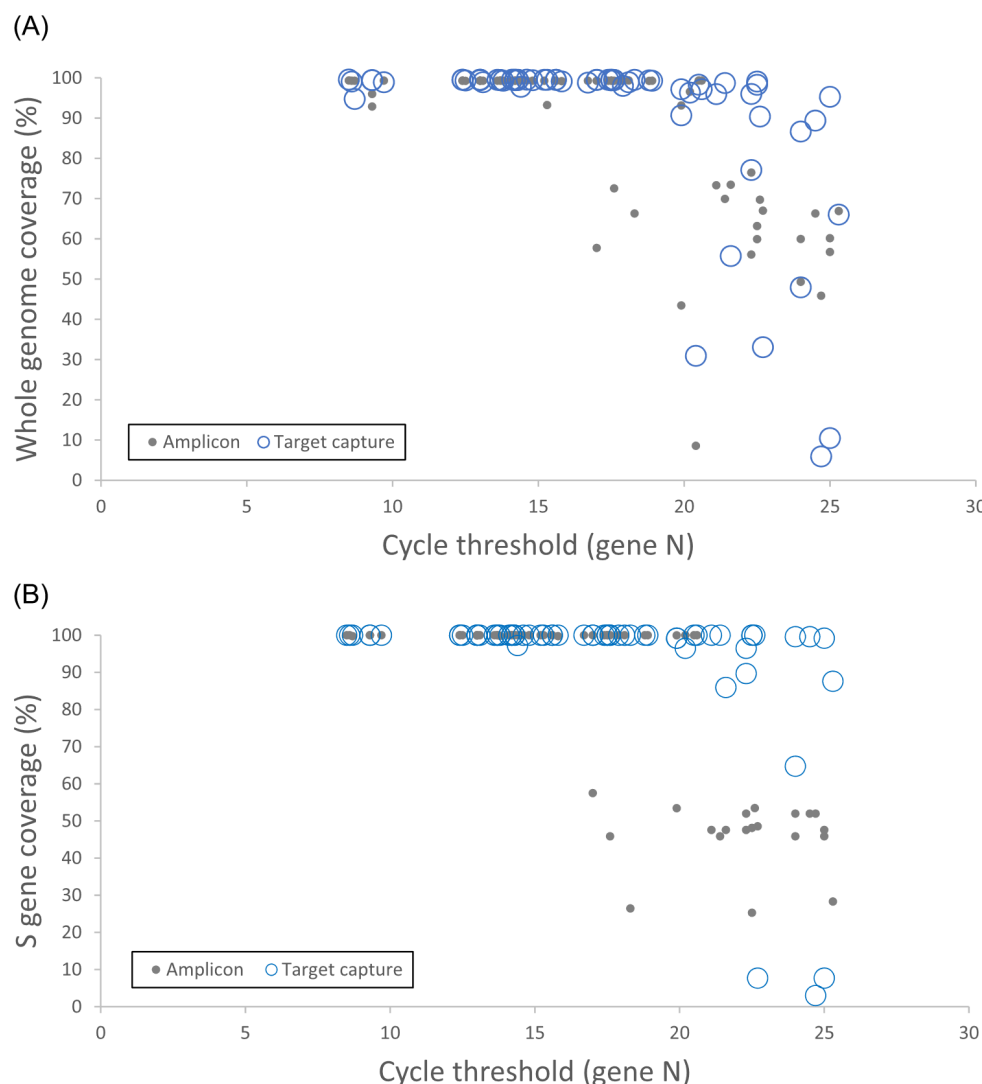


**FIGURE 2** Median read depth (blue line) and maximum coverage (blue area) along the entire SARS-CoV-2 genome (reference NC-045512.2) (A) target capture or (B) amplicon. 3'-UTR, 3'-untranslated region; ORF, open reading frame.

protocol ( $p < 0.08$ ). S gene coverage was higher for the target capture than for the amplicon protocol (Figure 3B). After stratification on  $C_t$  values (median 17 IQR: 13.8–20.6), the target capture protocol gave higher S gene coverage than the amplicon protocol only for  $C_t > 17$  ( $p < 0.01$ ). The mutations detected in 43/64 (67%) samples with the target capture and amplicon protocols were 100%

concordant. The target capture protocol detected mutations that were missed by the amplicon protocol in 17/21 (81%) samples and the amplicon protocol detected mutations in four (19%) samples that were not detected by target capture (Table 1). Mutations missed by the amplicon protocol were due to failure to amplify 2/4 amplicons covering the S gene in 18 samples and 3/4 amplicons in





**FIGURE 3** Whole genome (A) and S gene (B) coverage depending on the N gene cycle threshold with the two protocols.

three samples. Mutations missed by the target capture protocol were due to failure to sequence the almost whole S gene in 3/4 samples.

## 4 | DISCUSSION

Large-scale, high-throughput WGS of SARS-CoV-2 is essential for rapid surveillance and efficient follow-up of the spread of new variants, particularly immune-escaping variants that might interfere with vaccination or treatment. Our work focuses on SARS-CoV-2 WGS with Pacbio SMRT sequencing, while most published data are based on Illumina and ONT sequencing. We find that Pacbio SMRT sequencing is suitable for WGS of SARS-CoV-2, identifying variants and detecting mutations.

Both the target capture and amplicon protocols provided good genome coverage (median > 99%). Clades and lineages were concordant with both methods except for the lineage of one sample due to

the low genome coverage obtained with amplicon protocol, leading to misclassification. Sample 16 was flagged as having multiple strains with the target capture protocol, indicating possible contamination whereas no contamination was observed with the amplicon protocol. Further investigations were not possible due to the low genome coverage and low read depth obtained with both protocols. As sample crossover has already been reported with sequencing protocols on the Illumina platform,<sup>27</sup> decontamination strategies are clearly important.<sup>28</sup>

The target capture protocol was more sensitive than the amplicon protocol. In fact, a higher number of samples were successfully sequenced on the whole genome, a higher number of sequences had a genome coverage >95% (the minimum coverage value recommended by ECDC<sup>26</sup>) and a higher S gene coverage was observed for samples with  $C_t > 17$ . The target capture protocol amplifies a short 675 bp fragment of the SARS-CoV-2 genome and each base is covered by around 20 probes, while the amplicon protocol amplifies a 1.2 kb fragment. This is probably why target



**TABLE 1** Spike gene mutations detected by target capture and amplicon protocols in samples gave discrepant results (n = 21)

Sample	Clade <sup>a</sup>	Target capture		Amplicon	
		C <sub>t</sub> value	S gene coverage (%)	Missing position in spike (AA)	Mutation detected
1	21J (Delta)	25	99.6	522-526	T19R, G142D, E156-, F157-, R158G, L452R, T478K, D614G, P681R, D950N
2	21J (Delta)	21	100	-	T19R, G142D, E156-, F157-, R158G, L452R, T478K, D614G, P681R, D950N
3	21J (Delta)	24.7	3	1-1236	T19R, G142D, E156-, F157-, R158G, L452R, T478K, D614G
4	21J (Delta)	25	7.7	1-1176	T19R, G142D, E156-, F157-, R158G, L452R, T478K, D614G, P681R, D950N
9	21J (Delta)	17.6	100	-	T19R, G142D, E156-, F157-, R158G, L452R, T478K, D614G, P681R, D950N
10	21J (Delta)	22.3	89.7	255-296, 522-607	T19R, G142D, E156-, F157-, R158G, P251L, L452R, T478K, D614G, P681R, D950N
13	20I (Alpha, V1)	18.3	100	-	H69-, V70-, Y144-, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H
14	21J (Delta)	21.6	85.9	22-26, 252-296, 512-577, 803-866	T19R, T95I, G142D, E156-, F157-, R158G, L452R, T478K, D614G, Q677H, P681R, D950N
19	21J (Delta)	24	99.6	22-26	T19R, T95I, G142D, E156-, F157-, R158G, L452R, T478K, D614G, Q677H, P681R, D950N
23	21J (Delta)	19.9	99.2	252-256, 292-296	T19R, G142D, E156-, F157-, R158G, L452R, T478K, D614G, P681R, D950N
24	21J (Delta)	22.5	100	-	T19R, T29A, G142D, E156-, F157-, R158G, T250I, T299I, L452R, I468V, T478K, Q613H, D614G, P681R, D950N
29	21J (Delta)	21.1	100	-	E156-, F157-, T19R, G142D, R158G, L452R, T478K, D614G, P681R, D950N

(Continues)



Sample	Clade <sup>a</sup>	Target capture			Amplicon		
		C <sub>t</sub> value	S gene coverage (%)	Missing position in spike (AA)	Mutation detected	S gene coverage (%)	Missing position in spike (AA) <sup>b</sup>
32	19A	27.0	NA	1-1273		51.98	683-1274
35	21J (Delta)	25.0	94.2	22-26, 522-526	T19R, T95I, G142D, E156-, F157-, R158G, L452R, T478K, D614G, P681R, D950N	47.58	343-661, 683-1032
37	21J (Delta)	22.3	97.6	262-286, 292-296	T19R, G142D, E156-, F157-, R158G, L452R, T478K, D614G, P681R, D950N	47.58	343-661, 683-1032
44	21J (Delta)	22.6	100	-	T19R, T95I, G142D, E156-, F157-, R158G, L452R, T478K, D614G, P681R, D950N	53.47	683-1032, 1051-1274
46	21J (Delta)	24	64.8	1-26, 252-296, 522-866, 1093-1126	G142D, E156-, F157-, R158G, L452R, T478K, D950N	45.88	343-1032
47	21J (Delta)	25	87.6	1-26, 252-296, 522-526, 752-786, 799, 801, 803-816, 833-866	G142D, E156-, F157-, R158G, L452R, L455F, T478K, D614G, P681R, D950N	28.32	343-661, 681-1032, 1051-1274
63	20A	22.7	7.7	1-1176		48.6	20-325, 683-1032
71	20E (EU1)	17	100	-	A222V, D614G	57.5	343-661, 1051-1274
77	20I (Alpha, V1)	22.5	100	-	H69-, V70-, Y144-, N501Y, A570D, D614G, P681H, T716I, S982A, A1020S, D1118H	48.1	1-661

Amplicons covering S gene with 1.2 kb protocol: NC\_045512.2: 21 563–22 612 (spike AA mutations: 1–350), NC\_045512.2: 22 538–23 631 (spike AA mutations, covering the receptor binding domain (RBD): 325–690), NC\_045512.2: 23 545–24 736 (spike AA mutations: 661–1058) and NC\_045512.2: 24 659–25 790 (spike AA mutations: 1032–1409).

capture performed better and provide better genome coverage for samples with a low viral load.

For quasispecies studies where haplotyping is necessary, protocols amplifying long fragments should be preferred. With the amplicon protocol, the S gene can be sequenced with only four amplicons, one covering the RBD. This allows haplotyping of the main mutations in the S gene. PacBio SMRT sequencing of the S gene with amplicon lengths of 2.5–6.1 kb has been used to study spike gene quasispecies.<sup>22–25</sup> Sun et al.<sup>24</sup> showed that the virus population may consist of one predominant haplotype combined with numerous minor haplotypes and that different quasispecies complexity is observed depending on the tissue suggesting independent replication. SARS-CoV-2 spike protein evolution in an infected patient treated with mAbs indicated that key activity-reducing mutations can appear in patients treated with mAbs.<sup>25</sup> Long amplicon sequencing can provide accurate monitoring of SARS-CoV-2 quasispecies for compartmentalization studies or surveillance of mutations escaping variants for patients treated with mAbs.

The two laboratory workflows take 3 days from extraction to sequencing. The target capture protocol workflow is simpler, using only one plate throughout the process, which reduces the risks of technical errors and sample contamination. Automation is possible, but only with specific small-volume liquid handlers, whereas the amplicon protocol can be automated on the usual liquid handlers more readily available in the laboratory. The target capture sequencing time is shorter, providing results sooner. Target capture sequencing is analyzed with an automatic Pacbio analysis application on an SMRT link and can be run by a biologist on a computer. The amplicon protocol, in contrast, requires the development of a bioinformatic pipeline by a bioinformatician and specific computing resources. The target capture protocol is a complete end-to-end solution that is easier and faster to implement than the amplicon protocol.

The target capture protocol for SARS-CoV-2 WGS provided data that were similar to those obtained in previous studies with Illumina or ONT sequencing. SARS-CoV-2 WGS with amplicons on the Illumina platform generated sequences with >95% genome coverage for 67% of Delta variant samples.<sup>29</sup> We obtained similar results, with >95% genome coverage for 63% of samples. Some (19/84) of the samples sequenced by the target capture Pacbio SMRT system had been previously sequenced using the CovidSeq Illumina protocol with a similar median genome coverage (data not shown). Other studies obtained 90%–100% genome coverage with ONT sequencing.<sup>30–32</sup> Genome coverage was >99% for all high viral load samples ( $C_t < 20$ ) of SARS-CoV-2 sequenced using different protocols (mNGS, hybrid-capture-based enrichment, or amplicon-based protocols with Illumina sequencing and amplicon-based protocols with ONT sequencing).<sup>16</sup> The genome coverage was >98% for 94% of our samples with a  $C_t < 20$ . The small differences between studies could be due to the protocol used or to the variants sequenced, depending on the pandemic wave during which the samples were collected.

Newly-emerging strains with increased numbers of mutations can be a challenge for sequencing, as mutated primer-binding sites

may cause amplicon dropout or uneven sequencing coverage, resulting in lost or inaccurate data. The mutations in the Omicron variant introduced at the end of 2021 can decrease the enrichment efficiency of PCR amplification.<sup>33</sup> No Omicron variant samples were sequenced in this study, but the amplification could be influenced if the primers hybridize to regions with mutations. The SARS-CoV-2 target capture probes and 1.2 kb amplicon primers will probably have to be optimized to ensure good amplification with Omicron or other new variants. For example, ARTIC Network V4 primers were proposed to optimize the sequencing of Delta variants<sup>29</sup> and ARTIC Network V4.1 primers were recently proposed for Omicron sequencing (<https://community.artic.network/t/sars-cov-2-v4-1-update-for-omicron-variant>). A recent study demonstrated that combining short and long sequencing data, obtained by combining Illumina and ONT sequencing, improved genome coverage and provided uniform, maximum genome coverage.<sup>30</sup> Short and long-read sequencing could be done on a single sequencing platform with the target capture and amplicon Pacbio SMRT sequencing protocols.

To conclude, we find the PacBio SMRT technology suitable for the WGS of SARS-CoV-2, the rapid identification of circulating SARS-CoV-2 variants, and mutation detection. The genome coverage of the target capture protocol was similar to that of Illumina or ONT sequencing and the accurate long reads produced by the amplicon SMRT sequencing protocol can be used to study quasispecies. Further studies are now needed to compare the performance of PacBio SMRT technology with those of other platforms for sequencing Omicron variants.

## AUTHOR CONTRIBUTIONS

Pauline Trémeaux, Stéphanie Raymond, and Jacques Izopet designed the project. Noémie Ranger and Cécile Donnadieu carried out the experiments. Gérald Salin, Justine Latour, and Nicolas Jeanne performed bioinformatics analyses. Florence Nicot, Pauline Trémeaux, and Jacques Izopet analyzed the study data. Chloé Dimeglio carried out statistical analysis. Florence Nicot and Jacques Izopet wrote the manuscript. All the authors have approved the manuscript.

## ACKNOWLEDGMENT

The authors would like to thank Dr Owen Parkes for editing the English text. Target capture reagents were provided by Pacific Biosciences.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The raw data are available on demand. Sequences are available on GISAID (accession numbers in Supporting Information: Table 1).

## ORCID

Florence Nicot  <http://orcid.org/0000-0003-2722-2159>

Chloé Dimeglio  <http://orcid.org/0000-0002-3713-4913>



## REFERENCES

- Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579:265-269.
- Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579:270-273.
- Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. *Gene Rep*. 2020;19:100682.
- Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor recognition by the novel coronavirus from Wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *J Virol*. 2020;94:e000127-20.
- Rotondo JC, Martini F, Maritati M, et al. SARS-CoV-2 infection: new molecular, phylogenetic, and pathogenetic insights. efficacy of current vaccines and the potential risk of variants. *Viruses*. 2021;13:1687.
- CDC. December 1, 2021. Centers for Disease Control and Prevention: SARS-CoV-2 variant classifications and definitions. Accessed December 1, 2021. <https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html>
- Davies NG, Jarvis CI, Group CC-W, et al. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. *Nature*. 2021;593:270-274.
- Tegally H, Wilkinson E, Giovanetti M, et al. detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*. 2021;592:438-443.
- Voloch CM, da Silva Francisco R Jr, de Almeida LGP, et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J Virol*. 2021;95:e00119-e00121. doi:10.1128/JVI.00119-21
- Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet*. 2021;398:2126-2128.
- Peng J, Liu J, Mann SA, et al. Estimation of secondary household attack rates for emergent spike L452R SARS-CoV-2 variants detected by genomic surveillance at a community-based testing site in San Francisco. *Clin Infect Dis*. 2021;74:32-39. doi:10.1093/cid/ciab283
- Bal A, Destras G, Gaymard A, et al. Molecular characterization of SARS-CoV-2 in the first COVID-19 cluster in France reveals an amino acid deletion in nsp2 (Asp268del). *Clin Microbiol Infect*. 2020;26:960-962.
- Bhoyar RC, Jain A, Sehgal P, et al. High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next-generation sequencing. *PLoS One*. 2021;16:e0247115.
- Harilal D, Ramaswamy S, Loney T, et al. SARS-CoV-2 whole genome amplification and sequencing for effective population-based surveillance and control of viral transmission. *Clin Chem*. 2020;66:1450-1458.
- Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382:727-733.
- Charre C, Ginevra C, Sabatier M, et al. Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol*. 2020;6:veaa075.
- Nasir JA, Kozak RA, Aftanas P, et al. A comparison of whole genome sequencing of SARS-CoV-2 using amplicon-based sequencing, random hexamers, and bait capture. *Viruses*. 2020;12:895.
- Xiao M, Liu X, Ji J, et al. Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. *Genome Med*. 2020;12:57.
- Goswami C, Sheldon M, Bixby C, et al. Identification of SARS-CoV-2 variants using viral sequencing for the centers for disease control and prevention genomic surveillance program. *BMC Infect Dis*. 2022;22:404.
- Bull RA, Adikari TN, Ferguson JM, et al. Analytical validity of nanopore sequencing for rapid SARS-CoV-2 genome analysis. *Nat Commun*. 2020;11:6272.
- Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. 2020;21:597-614.
- Ko SH, Bayat Mokhtari E, Mudvari P, et al. High-throughput, single-copy sequencing reveals SARS-CoV-2 spike variants coincident with mounting humoral immunity during acute COVID-19. *PLoS Pathog*. 2021;17:e1009431.
- Lhomme S, Latour J, Jeanne N, et al. Prediction of SARS-CoV-2 variant lineages using the S1-Encoding region sequence obtained by PacBio single-molecule real-time sequencing. *Viruses*. 2021;13:2544.
- Sun F, Wang X, Tan S, et al. SARS-CoV-2 quasispecies provides an advantage mutation pool for the epidemic variants. *Microbiol Spectr*. 2021;9:e0026121.
- Vellas C, Del Bello A, Debarb A, et al. Influence of treatment with neutralizing monoclonal antibodies on the SARS-CoV-2 nasopharyngeal load and quasispecies. *Clin Microbiol Infect*. 2022;28(139):139.
- ECDC. Sequencing of SARS-CoV-2: first update. Control ECfDPa; 2021.
- Rosenthal SH, Gerasimova A, Ruiz-Vega R, et al. Development and validation of a high throughput SARS-CoV-2 whole genome sequencing workflow in a clinical laboratory. *Sci Rep*. 2022;12:2054.
- Mwangi P, Mogotsi M, Ogunbayo A, et al. A decontamination strategy for resolving SARS-CoV-2 amplicon contamination in a next-generation sequencing laboratory. *Arch Virol*. 2022;167:1175-1179. doi:10.1007/s00705-022-05411-z
- Lambisia AW, Mohammed KS, Makori TO, et al. Optimization of the SARS-CoV-2 ARTIC network V4 primers and whole genome sequencing protocol. *Front Med*. 2022;9:836728.
- Arana C, Liang C, Brock M, et al. A short plus long-amplicon based sequencing approach improves genomic coverage and variant detection in the SARS-CoV-2 genome. *PLoS One*. 2022;17:e0261014.
- Brinkmann A, Ulm SL, Uddin S, et al. AmpliCoV: rapid whole-genome sequencing using multiplex PCR amplification and real-time Oxford Nanopore MinION sequencing enables rapid variant identification of SARS-CoV-2. *Front Microbiol*. 2021;12:651151.
- Li J, Wang H, Mao L, et al. Rapid genomic characterization of SARS-CoV-2 viruses from clinical specimens using nanopore sequencing. *Sci Rep*. 2020;10:17492.
- Ma W, Yang J, Fu H, et al. Genomic perspectives on the emerging SARS-CoV-2 omicron variant. *Genomics Proteomics Bioinformatics*. 2022. doi:10.1016/j.gpb.2022.01.001

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Nicot F, Trémeaux P, Latour J, et al. Whole-genome sequencing of SARS-CoV-2: comparison of target capture and amplicon single molecule real-time sequencing protocols. *J Med Virol*. 2022;95:e28123. doi:10.1002/jmv.28123

