



HAL
open science

Prediction of SARS-CoV-2 Variant Lineages Using the S1-Encoding Region Sequence Obtained by PacBio Single-Molecule Real-Time Sequencing

Sébastien Lhomme, Justine Latour, Nicolas Jeanne, Pauline Trémeaux, Noémie Ranger, Marion Miguères, Gérald Salin, Cécile Donnadiou, Jacques Izopet

► **To cite this version:**

Sébastien Lhomme, Justine Latour, Nicolas Jeanne, Pauline Trémeaux, Noémie Ranger, 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 (12), pp.2544. 10.3390/v13122544 . hal-03771640

HAL Id: hal-03771640

<https://ut3-toulouseinp.hal.science/hal-03771640>

Submitted on 7 Sep 2022

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

Article

Prediction of SARS-CoV-2 Variant Lineages Using the S1-Encoding Region Sequence Obtained by PacBio Single-Molecule Real-Time Sequencing

Sébastien Lhomme ^{1,2,*}, Justine Latour ², Nicolas Jeanne ², Pauline Trémeaux ², Noémie Ranger ²,
Marion Miguères ^{1,2}, Gérald Salin ³, Cécile Donnadieu ³ and Jacques Izopet ^{1,2}

¹ Infinity, Université Toulouse, CNRS, INSERM, UPS, 31300 Toulouse, France; Miguères.m@chu-toulouse.fr (M.M.); izopet.j@chu-toulouse.fr (J.I.)

² Laboratoire de Virologie, CHU Toulouse, Hôpital Purpan, 31300 Toulouse, France; latour.j@chu-toulouse.fr (J.L.); jeanne.n@chu-toulouse.fr (N.J.); tremeaux.p@chu-toulouse.fr (P.T.); ranger.no@chu-toulouse.fr (N.R.)

³ INRAE, US 1426, GeT-PlaGe, Genotoul, 31326 Castanet-Tolosan, France; gerald.salin@inrae.fr (G.S.); cecile.donnadieu@inrae.fr (C.D.)

* Correspondence: lhomme.s@chu-toulouse.fr; Tel.: +33-5-67-69-04-24

Abstract: The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is the causal agent of the COVID-19 pandemic that emerged in late 2019. The outbreak of variants with mutations in the region encoding the spike protein S1 sub-unit that can make them more resistant to neutralizing or monoclonal antibodies is the main point of the current monitoring. This study examines the feasibility of predicting the variant lineage and monitoring the appearance of reported mutations by sequencing only the region encoding the S1 domain by Pacific Bioscience Single Molecule Real-Time sequencing (PacBio SMRT). Using the PacBio SMRT system, we successfully sequenced 186 of the 200 samples previously sequenced with the Illumina COVIDSeq (whole genome) system. PacBio SMRT detected mutations in the S1 domain that were missed by the COVIDseq system in 27/186 samples (14.5%), due to amplification failure. These missing positions included mutations that are decisive for lineage assignment, such as G142D ($n = 11$), N501Y ($n = 6$), or E484K ($n = 2$). The lineage of 172/186 (92.5%) samples was accurately determined by analyzing the region encoding the S1 domain with a pipeline that uses key positions in S1. Thus, the PacBio SMRT protocol is appropriate for determining virus lineages and detecting key mutations.

Keywords: SARS-CoV-2; PacBio SMRT sequencing; Illumina sequencing; S1 domain; clade; lineage



Citation: Lhomme, S.; Latour, J.; Jeanne, N.; Trémeaux, P.; Ranger, N.; Miguères, M.; Salin, G.; Donnadieu, C.; Izopet, J. 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. <https://doi.org/10.3390/v13122544>

Academic Editor: Peng Zhou

Received: 18 November 2021

Accepted: 16 December 2021

Published: 18 December 2021

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



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a member of the *Coronaviridae* family, that caused the COVID-19 pandemic is a ~30 kb single-stranded, positive-sense RNA virus [1,2]. Its genome contains six main open reading frames (ORFs): replicase (ORF1a/ORF1b), spike (S), envelope (E), membrane (M), and nucleocapsid (N), and at least seven other putative ORFs encoding accessory proteins interspersed between the structural genes [3]. To date, a total of 16 proteins were described [4]. The 1273-amino-acid (aa) long S protein and the mature trimeric spike protein is composed of exterior S1 and transmembrane S2 subunits. The S1 subunit attaches the virus to the host receptor (angiotensin converting enzyme 2; ACE2) via its receptor binding domain (RBD, aa residues 319–529) and the S2 subunit ensures fusion of the virus and host cellular membranes [3,5–7]. The rate of evolution of SARS-CoV-2 was initially thought to be limited due to the existence of a 3′-5′ exonuclease proofreading function of nonstructural protein 14 (nsp14). However, variants with mutations in the S domain have emerged around the world. One of the first notable variants had a D614G substitution in the S1 domain that increased the affinity of the virus for ACE2 [8]. The SARS-CoV-2 genome has since

diverged to produce several clades and lineages that seem to differ in their biology and/or geographic distribution [9]. For instance, the European lineage B.1.177 (clade 20E(EU1)) differs from ancestral sequences at 6 or more positions, including a A222V mutation in the spike protein. This variant arose in Spain early in the summer of 2020 and subsequently spread across Europe, perhaps carried by infected holiday travelers [10]. Another variant belonging to lineage B.1.160 (clade 20A, previously identified 20A.EU2) that has a S477N substitution in the spike protein was common in Autumn 2020 and early 2021 in some European countries, including France [10]. The more recent Alpha variant (lineage B.1.1.7, clade 20I) has spread rapidly from the United Kingdom. It harbors three aa deletions (69del-70del and 144del) and seven mutations in the spike protein, including D614G and N501Y [11]. The emergence of the Beta variant in South Africa (lineage B.1.351, clade 20H) is of great concern as it has since spread worldwide [12]. Variants belonging to this lineage have three mutations in the RBD: K417N, E484K, and N501Y, and several others outside the RBD and are neutralized less efficiently by convalescent and vaccine sera [13,14]. The Gamma variants described in Brazil belonging to the lineage P.1 clade 20J also have the E484K mutation in addition to K417T and N501Y mutations in their RBD [15]. The latest variant of concern, variant Delta (lineage B.1.617.2, clades 21A, 21I and 21J), emerged in India in October 2020 and was later detected in France in May 2021, and has the L452R mutation that makes it more transmissible [16,17]. Since then, many B.1.617.2 sublineages were described (also called AY lineages, for which AY.1 corresponds to B.1.617.2.1) and their number is currently increasing.

The rapid identification of the lineage and related variants that are transmitted more efficiently or that escape the immune response is key for the clinical management of patients, especially those who are immunocompromised that can receive immunotherapies, and monitoring virus spread. Studies demonstrated that certain monoclonal antibodies have low activity on SARS-CoV-2 variants [18,19]. High throughput next-generation sequencing (NGS) methods have been used to study the genomic diversity of SARS-CoV-2 worldwide [20–22]. The Pacific Biosciences (PacBio) single-molecule, real-time (SMRT) sequencing system, which records the incorporation of nucleotides into a single DNA template molecule by an immobilized DNA polymerase, provides long and highly accurate sequences thanks to circular consensus reads [23]. It enables full-length virus genome sequencing and a detailed analysis of the spike protein-encoding region [24–26].

The aim of this study was to assess the performance of single molecule real time sequencing for both genotyping using the S1 encoding region sequence and determining the haplotypes of the SARS-CoV-2 RNA population thanks to long read sequencing. Compared to Sanger direct sequencing, PacBio sequencing allows high throughput sequencing and determination of the composition of the quasispecies, including low frequency variants. Thus, we compared the results obtained with PacBio SMRT sequencing to those obtained with the Illumina COVIDseq protocol, largely used to define SARS-CoV-2 lineages and clades.

2. Materials and Methods

2.1. Samples

Nasopharyngeal samples were taken from patients who tested positive for SARS-CoV-2 RNA between 19 January and 26 January 2021 ($n = 146$) and between 10 July and 23 July 2021 ($n = 54$) and stored at $-80\text{ }^{\circ}\text{C}$ in the Virology laboratory at Toulouse University Hospital. We selected 200 consecutive samples of these periods for which N gene Ct values were below 25 with the TaqPath™ COVID-19 CE-IVD RT-PCR kit (Thermo Fisher Scientific, Pleasanton, CA, USA) used on the QuantStudio™ 5 Real-Time PCR System (Applied Biosystems, Singapore, Singapore) and that were successfully sequenced with IlluminaCOVIDSeq. They were sequenced on both the PacBio SMRT (S1 region) and Illumina (whole genome) systems.

2.2. SARS-CoV-2 RNA Extraction

Virus RNA was extracted from 180 µL transport medium with the MGIEasy Nucleic Acid extraction kit on the MGI SP 960 system (Beijing Genome Institute, Shenzhen, China) according to the manufacturer's instructions.

2.3. PacBio SMRT Sequencing

Primer design: We used two primers from the Pacific Biosciences protocol for full genome sequencing (<https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Multiplexing-2.5-kb-Amplicons-for-Whole-Genome-Sequencing-of-SARS-CoV-2.pdf>, accessed on 30 June 2020). These primers target a single long amplicon named A6 (2490 nucleotides (nt)) that encodes the full length of the spike protein S1 domain.

cDNA synthesis and library preparation: we generated a single long amplicon named A6 (2490 nucleotides [nt]) that encoded the full length of the spike protein S1 domain. The 25 µL RT reaction mixture contained 10 µL of RNA, 10 µL of water, and 5 µL Superscript IV VILO (Life Technologies, Courtaboeuf, France) and random hexamer oligo(dT) primers. The cycle steps were: 10 min/25 °C, 30 min/50 °C, and 10 min/85 °C.

The A6 cDNA was amplified in 20 µL of reaction mixture: 10 µL of Platinum SuperFi Master Mix, 2 µL of water, 3 µL of forward (A6F; 5'-GTAAAACGACGGCCAGTACAAATC CAATTCAGTTGTCTTCCTATTC-3') and reverse primer (A6R 5'-CAGGAAACAGCTATGA CTGTGTACAAAACTGCCATATTGCA-3') M13 tailed primers [27], and 2 µL of cDNA. The cycle steps were: denaturation (98 °C/2 min), amplification (30 cycles of 98 °C/15 s, 65 °C/30 s, 72 °C/2 min), and a final extension (72 °C/5 min).

A second PCR was then performed using barcoding primers tailed with the universal M13 sequence. Reaction mixture: 15 µL of Kapa Ready mix 2× in 9.6 µL of nuclease-free water, 1.7 µL of each barcoded primer (final concentration: 3 µM), and 2 µL of amplified cDNA. Cycling conditions were: 95 °C/5 min, 2 cycles of 98 °C/20 s, 60 °C/15 s, and 72 °C/2 min; then 24 cycles of 98 °C/20 s, 65 °C/15 s, and 72 °C/2 min; final extension, 72 °C/5 min, samples cooled at 12 °C. Products (expected size around 3 kb) were purified with the AMPure PacBio (Pacific Bioscience, Menlo Park, CA, USA) system at ×4.2 according to the manufacturer's instructions and then quantified with the Quantifluor DSDNA system running on a Roche LC480 instrument according to the manufacturer's conditions.

Aliquots (10 ng) of each sample were pooled, purified with the AMPure PacBio system (Pacific Bioscience, Menlo Park, CA, USA) at ×0.6, quantified with the Denovix dsDNA BR kit, and the size determined with the Fragment Analyzer system using the kit NGS DNF-473 kit (Agilent, Les Ulis, France), according to the manufacturer's instructions.

SMRTbell libraries were constructed by pooling 96 barcoded samples. Barcoded amplicon libraries were prepared and sequenced with the SMRTbell Express Template Prep 2.0 kit according to the manufacturer's instructions. Libraries were quantified at each step with the Qubit HS DSDNA system (Thermo Fisher Scientific, Villebon sur Yvette, France).

Sequencing: samples were prepared with the Sequel binding and internal control kit 3.0 and sequenced with the Sequencing kit by placing 6 pM in the SMRT cell 1M v3 Tray and using 600-min movies on the Sequel Platform (Pacific Bioscience, Menlo Park, CA, USA).

Data analysis: the A6 amplicon consensus sequences were built from the PacBio reads using a home-made Snakemake pipeline. First, the HiFi reads were generated from the PacBio subreads.bam using PacBio CCS tool (v.6.0.0, <https://github.com/PacificBiosciences/ccs>, accessed on 26 October 2021) and demultiplexed with Lima (v.2.0.0, <https://github.com/PacificBiosciences/barcoding>, accessed on 26 October 2021). The resulting reads were analyzed by a PacBio open-access tool: pbAA (v.0.1.3, <https://github.com/PacificBiosciences/pbAA>, accessed on 26 October 2021) which clustered similar reads together and generated one or several haplotypes depending on the diversity of the viral population of each sample. The resulting haplotypes were then mapped on the SARS-CoV-2 reference genome (Wuhan-Hu-1 isolate, Genbank accession number NC_045512.2) with Minimap2 (v2.17) [28] to discard eventual chimeric reads.

In parallel, a consensus sequence was built from the reads and the files were generated by pbAA with another PacBio tool: CoSA (Coronavirus Sequence Analysis, v9.0.0, <https://github.com/Magdoll/CoSA>, accessed on 26 October 2021). Then, the haplotypes sequences and the consensus sequence (for each sample) were treated separately. Each haplotype/consensus sequence was aligned on the reference SARS-CoV-2 spike region (NC_045512.2) with MAFFT (v.7.475) [29] and trimmed from the beginning of the spike gene (nt 21563) to the end of the A6 amplicon (nt 23823). Aligned sequences were translated into amino acids and again aligned with MAFFT to the spike amino acids reference sequence (YP_009724390.1).

We used a variant table that included the variant strains and their key mutations identified by the French SARS-CoV-2 National Reference Center (NRC) (Table 1). A python script retrieved the mutations (SNP, indel) carried by the haplotype or consensus sequences and compared them to the NRC profiles. A profile was validated if the analyzed sequence harbored 80% of the mutations from the spike region profile covered by the A6 amplicon, as some of the mutations were not present or covered, especially with the Illumina whole genome amplification protocol. The validated profile with the most matching mutations was assigned to the consensus. Conversely, if a sequence did not match the 80% threshold for any profile, it was not assigned (NA).

2.4. Illumina COVIDSeq Sequencing

cDNA synthesis and library preparation: cDNA was synthesized from extracted RNA with random hexamers using the ARTIC primers (version 3) and the Illumina COVIDSeq protocol (Illumina, San Diego CA, USA) and amplified using a multiplex PCR protocol that produced 98 amplicons spanning the whole SARS-CoV-2 genome, according to the manufacturer's instructions. The primer pool also contained primers targeting human RNA that produced an additional 11 amplicons.

Libraries were prepared using the Illumina COVIDSeq protocol (Illumina, USA), according to the manufacturer's instructions. The PCR-amplified products were later processed for tagmentation and adapter ligation using the IDT for Illumina Nextera UD Indexes Set A, B, C, and D (384 indexes, 384 samples). Samples were further enriched and cleaned up with protocols provided by Illumina. All samples were processed as batches on 96-well plates that included one COVIDSeq positive control HT (CPC HT) and one no-template control (NTC). These 96 libraries were pooled in a microcentrifuge tube and quantified using the Qubit dsDNA HS Assay kit on a Qubit fluorometer (Invitrogen, Villebon sur Yvette, France), according to the manufacturer's instructions. Fragment sizes were determined in an Agilent Fragment analyzer 5200 (Agilent, Les Ulis, France). The pooled library was normalized to a concentration of 4 nM and 25 µL aliquots of each normalized pool containing index adapter sets A, B, C, and D were combined in a new microcentrifuge tube to produce a final pool of 384 samples at a starting concentration of 4 nM.

Sequencing: the pooled libraries were denatured and neutralized with 0.2N NaOH and 400 mM Tris-HCl (pH-8). Replicates of each 384-sample pool were placed in the S4-flow cell according to the NovaSeq-XP workflow (Illumina, USA) and subjected to dual-indexed on a NovaSeq 6000 platform. The 141 samples of January 2021 were sequenced with single-end 35 pb read length while the 45 samples collected in July 2021 were sequenced with paired-end 150 pb read length.

Data analysis: the resulting sequence data were processed using DRAGEN COVID Lineage (v3.5.1) (Illumina, CA, USA) to obtain consensus sequences, Nextstrain clades (<https://nextstrain.org/sars-cov-2>, accessed on 31 August 2021), and Pangolin lineages (<https://cov-lineages.org/pangolin.html>, accessed on 31 August 2021) when at least 5 SARS-CoV-2 probes were detected.

Table 1. Cont.

Clade	20I/ B.1.1.7	20I/ B.1.1.7	20I/ B.1.1.7	20H/ B.1.351	21A/ B.1.617.2	21A/ B.1.617.2	20J/ P.1	21H	20D/ C.36.3	20B/ B.1.1.318	20A/ B.1.620	19B/ A.27	19B/ A.28	20I/ B.1.1.7	20B/P.3	20A/ B.1.160	20E/ B.1.177	20A/ B.1.214.2	20A/ B.1.221	21C/ B.1.427	21D/ B.1.525	21F/ B.1.526	20C/ B.1.526.1	20C/ B.1.616	21B/ B.1.617.1	20A/ B.1.617.3	20A/ B.1.619	21G/ C.37	
Lineage	B.1.1.7	B.1.1.7+ E484K	B.1.1.7+ E484Q	B.1.351	B.1.617.2	AY.1/AY.2	P.1	B.1.621	C.36.3	B.1.1.318	B.1.620	A.27	A.28	B.1.1.7+ L452R	B.1.1.28.3 /P.3	B.1.160	B.1.177	B.1.214.2	B.1.221	B.1.427/ B.1.429	B.1.525	B.1.526	B.1.526.1	B.1.616	B.1.617.1	B.1.617.3	B.1.619	C.37	
247del																												del	
248del																													del
249del																													del
250del																													del
251del																													del
252del																													del
D253G/c																													del
R346K/S								K	S														G						
Q414K																			K										
K417N/T				(N)		N*	T																						
N440K																												K*	
Y449H																													
N450K																			K										
L452R/Q					R*	(R)			R*			R*		R*						R*			R*		R*	R*		Q*	
S477N											N*					N						N*							
T478K					K*	(K)																							
V483A																													A
E484K/Q		K*	Q*	K*			K*	(K)		K*	K*				K*							K*	K*			Q*	Q*	K*	
F490S																													S
N501Y/T	Y*	Y*	Y*	Y*			Y*	Y				Y*	T*	Y*	Y*														
A570D	D	D	D												D														
D614G	G	G	G	G	G	(G)	G	G		G	G			G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
A653V												V																	
H655Y							Y					Y	Y														Y*		
G669S																												S	
Q677H								H				(H)										H							
N679K																													
P681H/R	H	H	H		R	(R)		H		H*	H			H	H											R	R		
A701V				V																									V
T716I	I	I	I											I					I										

*: Essential mutation to validate the profile. ^: At least one mutation required to validate the profile. (): Possible mutation, not required to validate the profile but improves discrimination. Only mutations detected in the S1 domain were covered by the fragment amplified by PacBio SMRT sequencing.

Among them, samples covering at least 75% of the SARS-CoV-2 genome and 85% of the spike region were analyzed for alignment. Consensus sequences were aligned to the SARS-CoV-2 Wuhan-Hu-1 reference genome (NC_045512.2) and 109 GISAID sequences of Pangolin lineages detected in France using MAFFT (v.7.475). Maximum likelihood trees were then calculated from these alignments with IQTREE (v.2.0.3, evolutionary model: GTR + F + R3) [30] to confirm any clade and lineage results.

GISAID accession numbers are provided in the Supplementary Table S1. All the raw data are available on demand.

3. Results

3.1. Determination of the Clades and Lineages Based on Full-Length Genomes

Nasopharyngeal samples were taken early in the introduction of the Alpha and of the Delta variants to France (samples collected in January and July 2021, respectively). Patients were 33.7 ± 16.6 years old and most were females (105; 52.5%). Most (145/200; 72.5%) of the patients were symptomatic. Symptoms appeared the day before sampling in 55 patients, 2 to 4 days before in 67 patients, 5 to 7 days before in 15 patients, and 8 days or more before in 8 patients. Clades were identified by NextClade and lineages were determined using the PANGOLIN classification based on the analysis of full-length genome sequences (Table 2). In January 2021, three strains were more frequently detected: the Alpha variant that belonged to the clade 20I and the lineage B.1.1.7 (48/146, 32.9%), followed by 37 (25.3%) strains belonging to the clade 20A and the lineage B.1.160 and 37 variants belonging to the clade 20E(EU1) and the lineage B.1.177. Most of the strains collected in July (51/54, 95.6%) were Delta variants (clade 21A), with a majority of strains belonging to the lineage B.1.617.2 ($n = 20$) and AY.4 ($n = 19$).

Table 2. Sample lineages and clades.

Clade	Lineage	Month of Collection		Total
		January	July	
20I	B.1.1.7	48	3	51
20A	B.1.160	37		37
20E(EU1)	B.1.177	37		37
21A	B.1.617.2		20	20
21A	AY.4		19	19
20A	B.1.221	8		8
20A	B.1	4		4
21A	AY.34		6	6
21A	AY.9		4	4
20B	B.1.1.241	3		3
20B	B.1.1.317	2		2
20H	B.1.351	2		2
20C	B.1	1		1
20B	B.1.1.186	1		1
20B	B.1.1.28	1		1
20G	B.1.2	1		1
20A	B.1.214	1		1
21A	AY.23		1	1
21A	AY.7.1		1	1
	Total	146	54	200

3.2. Comparison of the Spike Sequences Obtained with the Illumina and PacBio SMRT Systems

We developed a PacBio SMRT sequencing protocol that targeted the S1 domain and assessed its suitability for accurately predicting the lineage of a variant. First, we sequenced 146 samples collected in January 2021 with Illumina COVIDseq (with the recommended single-end 35 nt read length). Among them, 141/146 (96.6%) samples were successfully

sequenced with PacBio SMRT sequencing. The first protocol (PacBio) produced a final sequence length of 2465 nt mapped on the reference genome from the positions 21358 to 23823. Therefore, 2271 nt overlapped the spike gene (the 757th first amino acids), which covered 59% of the S gene. The region encoding the S1 domain, including the RBD (22520–23186), was fully covered. The Illumina protocol generated the whole genome consensus.

We then compared the S1 domain sequences obtained with the two platforms. The PacBio SMRT protocol identified mutations in 14/141 variants (9.9%) that were missed by the Illumina protocol (Table 3). All the missed mutations were due to a lack of coverage (<10 reads) and were grouped on the same amplicon NC_045512.2:22903–23122 (missing mutations: N440K, S477N, T478K, E484K, N501Y, A522S). A second amplicon, NC_045512:22346–22516 was less represented in the data, without consequences on the resulting mutation profiles (Figure 1A).

Consequently, for the sequencing of the 54 samples from July 2021, we extended the read length to 150 nt and used a paired-end kit; among these samples, 45/54 (83.3%) were sequenced successfully with PacBio SMRT system. The obtained depth of sequencing was greatly improved (Figure 1B), but we still missed some mutations in 13/45 (28.9%) samples (Table 3). Some mutations ($n = 7$) were not detected due to a default of amplification at the beginning of the spike gene (Figure 1B) while some others were completely absent of the variant calling results ($n = 3$), but most of them ($n = 9$) correspond to the presence of the mutation at less than 50% (not present in the resulting consensus sequence). The missing mutations were not located in the same region as previously, but rather over the amplicon NC_045512.2:21743–21961 (T95I, V130F, G142D, E156G, F157del, and R158del, A222V). It was noteworthy that as the PacBio protocol amplified only one long amplicon, the resulting coverage was always close to 100% when the amplification succeeded (with a median coverage of 2137 reads per position) (Figure 1C) but sequencing failed in 14/200 samples (7%). Taken together, the most frequently missed mutations were G142D ($n = 11$), E156G, 157del, 158del ($n = 6$), N501Y ($n = 6$), S477N ($n = 5$), T95I ($n = 3$), V130F ($n = 2$), and E484K ($n = 2$).

Lastly, we checked the minority variants found by the two protocols, to assess the capacity of the PacBio SMRT system to retrieve diversity as accurately as the Illumina COVIDseq system, but with a longest PCR product. The same minority variants were detected with both protocols in three samples and the resulting percentages were quite similar (L5F: 46.4% with PacBio and 25.6% with Illumina; L176F: 22.6% with PacBio and 37.5% with Illumina; and 27.4% with PacBio and 31.2% with Illumina).

3.3. Determination of SARS-CoV-2 Lineage by S1 Domain Key Position Analysis

We developed an automated classification pipeline based on 70 S1 domain key mutations, characteristic of the major variants circulating in Europe from the beginning of the epidemic crisis to October 2021 (Table 1). The key mutations detected in the S1 domain were used to determine the SARS-CoV-2 lineage. For example, a variant harboring mutations S477N and D614G was identified as belonging to lineage B.1.160. To assess the suitability of this approach, we compared the lineages obtained with this pipeline with those obtained by analyzing full-length genomes.

Lineages were assigned for 172/186 samples (92.5%). The 14 samples whose lineage could not be determined included strains belonging to lineage B.1 ($n = 5$), B.1.241 ($n = 3$), B.1.1.317 ($n = 2$), B.1.2 ($n = 1$), B.1.1.28 ($n = 1$), B.1.1.186 ($n = 1$), and B.1.214 ($n = 1$) which were not included in the international variants monitoring (Table 4). The clade was then deduced according to the lineage (Table 1). For the variants belonging to the clade 21A, all were assigned to the lineage B.1.617.2, but we were unable to determine the AY sub-lineages.

Table 3. List of mutations detected with the Illumina and PacBio SMRT systems in 14 patient samples giving discrepant results.

Sample	Clade	Lineage	Mutations Detected with Illumina	Mutations Detected with Pacbio
20	20A	B.1.214	T95I, D614G, T716I	T95I, T478K , D614G, T716I
33	20A	B.1.160	D614G	S477N , D614G
62	20A	B.1.160	D614G	S477N , D614G
79	20A	B.1.160	D614G	S477N , A522S , D614G
110	20A	B.1.160	D614G	S477N , D614G
111	20A	B.1.160	D614G	S477N , D614G
136	20B	B.1.1.186	L18F, D614G	L18F, N440K , D614G
30	20H (Beta)	B.1.351	L18F, D80A, D215G, 241del, 242del, 243del, K417N, N501Y, D614G, A701V	L18F, D80A, D215G, 241del, 242del, 243del, K417N, E484K , N501Y, D614G, A701V
134	20H (Beta)	B.1.351	D80A, D215G, 241del, 242del, 243del, K417N, D614G, A701V	D80A, D215G, 241del, 242del, 243del, K417N, E484K , N501Y , D614G, A701V
15	20I (Alpha)	B.1.1.7	69del, 70del, 144del, A570D, D614G, P681H, T716I	69del, 70del, 144del, N501Y , A570D, D614G, P681H, T716I
27	20I (Alpha)	B.1.1.7	69del, 70del, 144del, A570D, D614G, P681H, T716I	69del, 70del, 144del, N501Y , A570D, D614G, P681H, T716I
56	20I (Alpha)	B.1.1.7	69del, 70del, 144del, A570D, D614G, P681H, T716I	69del, 70del, 144del, N501Y , A570D, D614G, P681H, T716I
73	20I (Alpha)	B.1.1.7	69del, 70del, 144del, A570D, D614G, P681H, T716I	69del, 70del, 144del, N501Y , A570D, D614G, P681H, T716I
116	20I (Alpha)	B.1.1.7	69del, 70del, 144del, A570D, D614G, P681H, T716I	69del, 70del, 144del, N501Y , A570D, D614G, P681H, T716I
143	21A (Delta)	AY.4	T19R, E156G, 157del, 158del, L452R, T478K, D614G, P681R	T19R, T95I , G142D , E156G, 157del, 158del, L452R, T478K, D614G, P681R
153	21A (Delta)	AY.4	T19R, E156G, 157del, 158del, L452R, T478K, D614G, P681R, D950N	T19R, G142D , E156G, 157del, 158del, L452R, T478K, D614G, P681R
163	21A (Delta)	AY.4	T19R, L452R, T478K, D614G, P681R, D950N	T19R, G142D , E156G , 157del , 158del , L452R, T478K, D614G, P681R
167	21A (Delta)	AY.4	T19R, L452R, T478K, D614G, P681R, D950N	T19R, G142D , E156G , 157del , 158del , L452R, T478K, D614G, P681R
169	21A (Delta)	AY.4	T19R, E156G, 157del, 158del, L452R, T478K, D614G, P681R	T19R, G142D , E156G, 157del, 158del, L452R, T478K, D614G, P681R
170	21A (Delta)	AY.4	T19R, L452R, T478K, D614G, P681R	T19R, G142D , E156G , 157del , 158del , L452R, T478K, D614G, P681R
189	21A (Delta)	AY.4	T19R, E156G, 157del, 158del, L452R, T478K, D614G, P681R	T19R, T95I , G142D , E156G, 157del, 158del, L452R, T478K, D614G, P681R
186	21A (Delta)	AY.7.1	T19R, L452R, T478K, D614G, P681R	T19R, E156G , 157del , 158del , L452R, T478K, D614G, P681R
142	21A (Delta)	AY.9	T19R, E156G, 157del, 158del, A222V, L452R, T478K, D614G, P681R	T19R, V130F , E156G, 157del, 158del, A222V, L452R, T478K, D614G, P681R

Table 3. Cont.

Sample	Clade	Lineage	Mutations Detected with Illumina	Mutations Detected with Pacbio
183	21A (Delta)	AY.9	L452R, T478K, D614G, P681R, D950N	T19R, V130F, G142D, E156G, 157del, 158del, A222V , L452R, T478K, D614G, P681R
185	21A (Delta)	AY.9	T19R, E156G, 157del, 158del, A222V, L452R, T478K, D614G, P681R	T19R, G142D , E156G, 157del, 158del, A222V, L452R, T478K, D614G, P681R
157	21A (Delta)	AY.23	T19R, L452R, T478K, D614G, P681R, D950N, V1264L	T19R, G142D, E156G, 157del, 158del , L452R, T478K, D614G, P681R
154	21A (Delta)	AY.34	T19R, E156G, 157del, 158del, L452R, T478K, D614G, Q677H, P681R, D950N	T19R, T95I, G142D , E156G, 157del, 158del, L452R, T478K, D614G, Q677H, P681R

The mutations listed above were detected using the Wuhan-Hu-1genome sequence (NC_045512.2) as reference. In **bold**: mutations detected by the PacBio SMRT platform.

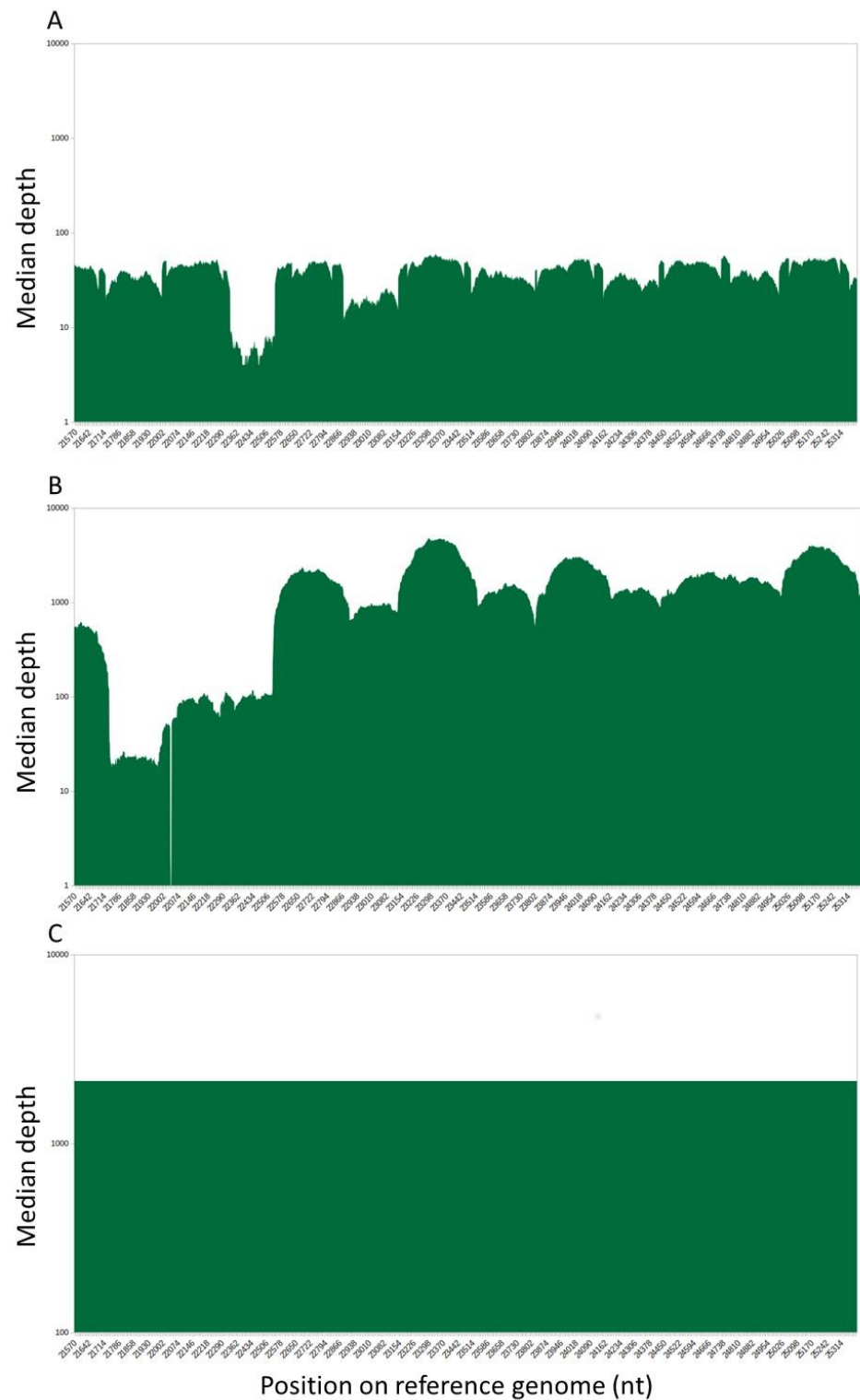


Figure 1. Sequencing depth over the spike gene with Illumina COVIDseq protocol. Logarithmic representation of the number of reads mapped on the reference genome, zoomed on the spike region. The x-axis was labeled from 21563 and 25384, corresponding to the positions on the complete genome. (A) Sequencing depth over the spike gene with Illumina COVIDseq protocol with singled-end 35 nt read-length sequencing and (B) paired-end 150 nt read-length sequencing. (C) Sequencing depth over the spike gene with PacBio protocol.

Table 4. Comparison of the lineages determined with full-length genomes and the S1-encoding region based on key positions.

Full Length Genome (Illumina)			Spike S1 Region (PacBio)	
Clade	Lineage	<i>n</i>	Lineage	<i>n</i>
20A	B.1.221	8	B.1.221	8
20A	B.1	4	NA	4
20A	B.1.214	1	NA	1
20A	B.1.160	37	B.1.160	37
20B	B.1.1.186	1	NA	1
20B	B.1.1.241	3	NA	3
20B	B.1.1.28	1	NA	1
20B	B.1.1.317	2	NA	2
20C	B.1	1	NA	1
20E (EU1)	B.1.177	37	B.1.177	37
20G	B.1.2	1	NA	1
20H	B.1.351	2	B.1.351	2
20I	B.1.1.7	45	B.1.1.7	45
21A	B.1.617.2	17		
21A	AY.4	16		
21A	AY.34	4		
21A	AY.9	4	B.1.617.2	43
21A	AY.23	1		
21A	AY.7.1	1		

NA: not assigned.

4. Discussion

The rapid emerging new SARS-CoV-2 variants must be accurately identified in order to follow-up their spread, particularly of any immune-escaping variants that might interfere with vaccination or monoclonal antibody (mAb) treatment. We developed a simple protocol for sequencing the region encoding the S1 domain using PacBio SMRT technology, because the emergence of such long-read technology has greatly facilitated the assembly of viral sequences. We also developed an automated pipeline for determining the SARS-CoV-2 variant lineages based on key amino acid positions. Analysis of the S1 domain, including the RBD, is well suited to monitor the appearance of immune-escaping mutations [31].

The sequences of the S1 domain of 186 samples obtained with the Sequel Platform (PacBio) were compared to those obtained by NovaSeq 6000 deep sequencing (Illumina). We found that the S1 domain was inadequately covered in the Illumina sequences, likely due to a less effective amplification for some of the designed ARTIC amplicons and the higher amount of material that can affect the read depths. This problem might be reduced or even suppressed with the use of the currently last available version (v4). In addition, the missing amplicons seems variant-dependent, with a difference of coverage between Delta and Alpha/B.1.160/B.1.177 variants. The full coverage of this region is critical since mutations contributing to virus escape from vaccine-induced neutralization are located in the S1 domain. For instance, the E484K mutation in the RBD notably described in the Beta and Gamma variants seems to confer resistance to the neutralizing capacity of fully vaccinated individuals [32,33]. The variant belonging to the lineage P.1 that emerged in Brazil has been responsible for reinfections, perhaps because of the three mutations in the RBD [34]. The variant belonging to lineage B.1.351 also resists neutralization due to mutations in the RBD and non-RBD domains [14,35]. Thus, the S1 domain must be fully

covered in order to detect already known mutation conferring resistance to neutralizing antibodies and to identify new mutations (SNP, deletions, or insertions) in the RBD in new variants that are less sensitive to neutralizing antibodies. Sequencing the whole of the S1 domain is also of importance as this is the region targeted by monoclonal antibody therapies. Two neutralizing antibodies cocktails (bamlanivimab/etesevimab and casirivimab/imdevimab) have been authorized by the US Food and Drug Administration and the European Medicines Agency for treating patients at high risk of severe disease. Detailed knowledge of the region encoding the S1 domain is needed in order to identify the substitutions associated with reduced neutralizing antibody activity if treatment fails, since resistant mutants may emerge following the use of these neutralizing antibodies [24]. The variant belonging to the lineage B.1.617.2 was shown to be resistant to the mAb bamlanivimab due to the escape mutation L452R [18] and was also less sensitive to sera from convalescent patients [36]. In addition, long reads data offer the opportunity to get the haplotypes and, thus, analyze whether the mutations detected are located on the same SARS-CoV-2 genome or on distinct genomes.

Lastly, we wanted to determine whether sequencing the S1 domain would enable us to determine accurately the virus lineage. The automated pipeline based on key mutations that are characteristic of some variants enabled us to assign clades to 172/186 (92.5%) sequences with a perfect agreement with the Illumina pipeline. The remaining unclassified variants belonged to less frequent variants circulating in Europe, which were not followed by the French monitoring system. The determination of the lineage is important because some are naturally less sensitive to neutralizing antibodies than are others. The lineage determination is also helpful to assess the current and future circulation of the different variants, like that of the strains belonging to the lineage B.1.1.7 and more recently that of the B.1.617.2 [37,38].

Our study has some limitations. Our pipeline does not allow the determination of the precise sub-lineage of Delta variant, because the differences detected in the S1 domain are not discriminant enough, with an exception for the AY.1 and AY.2 lineages, characterized by the K417N mutation [39]. Interestingly, K417N, L452R, E484K, and E484Q are the mutations known to disrupt receptor-binding domain (RBD) binding capacity, making them more infectious by immune escape against the current vaccines [40]. Another limitation is the design that was not a head-to-head comparison on consecutive clinical samples. Samples were selected based on successful sequencing by using the IlluminaCOVidSeq approach. A total of 14/200 (7%) samples failed to sequence using PacBio SMRT, indicating the need of optimization, particularly for samples with low viral load. Further studies are also needed to compare the two sequencing protocols.

5. Conclusions

We conclude that the PacBio SMRT technology is a suitable system for sequencing the SARS-CoV-2 spike protein S1 domain. It covers the whole domain, which is essential for virological follow-up, especially for monitoring the acquisition of mutations conferring full or partial resistance to neutralizing antibodies. This approach also offers the possibility to determine the lineage of the circulating variants by a partial sequencing.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/v13122544/s1>, Table S1: GISAID Accession Numbers.

Author Contributions: Conceptualization, S.L., P.T., and J.I.; methodology, N.J., N.R., J.L., G.S. and C.D.; formal analysis, S.L., J.L., N.J., P.T., J.L. and J.I.; data curation, N.J., N.R., J.L., G.S. and C.D.; writing—original draft preparation, S.L.; writing—review and editing, S.L., N.J., P.T., M.M., J.L. and J.I.; Supervision, S.L. and J.I.; funding acquisition, J.I. All authors have read and agreed to the published version of the manuscript.

Funding: ANRS-MIE0152 Emergen/Quasicov.

Institutional Review Board Statement: These analyses were part of the national SARS-CoV-2 surveillance. French law (CSP Art.L1121-1.1) does not require institutional review board approval for anonymous non-interventional studies.

Informed Consent Statement: Data were analysed using an anonymised database. Such a protocol does not require written informed consent according to French Public Health law (CSP Art L 1121-1.1).

Data Availability Statement: Raw data are available on demand to the corresponding author: lhomme.s@chu-toulouse.fr.

Acknowledgments: We thank Owen Parkes for editing the English text.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhou, P.; Yang, X.L.; Wang, X.G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.R.; Zhu, Y.; Li, B.; Huang, C.L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579*, 270–273. [[CrossRef](#)] [[PubMed](#)]
2. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.M.; Wang, W.; Song, Z.G.; Hu, Y.; Tao, Z.W.; Tian, J.H.; Pei, Y.Y.; et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265–269. [[CrossRef](#)] [[PubMed](#)]
3. Hu, B.; Guo, H.; Zhou, P.; Shi, Z.L. Characteristics of SARS-CoV-2 and COVID-19. *Nat. Rev. Microbiol.* **2021**, *19*, 141–154. [[CrossRef](#)] [[PubMed](#)]
4. Redondo, N.; Zaldivar-Lopez, S.; Garrido, J.J.; Montoya, M. SARS-CoV-2 Accessory Proteins in Viral Pathogenesis: Knowns and Unknowns. *Front. Immunol.* **2021**, *12*, 708264. [[CrossRef](#)]
5. Wan, Y.; Shang, J.; Graham, R.; Baric, R.S.; 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*, e00127-20. [[CrossRef](#)]
6. Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Kruger, N.; Herrler, T.; Erichsen, S.; Schiergens, T.S.; Herrler, G.; Wu, N.H.; Nitsche, A.; et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **2020**, *181*, 271–280.e278. [[CrossRef](#)]
7. Walls, A.C.; Park, Y.J.; Tortorici, M.A.; Wall, A.; McGuire, A.T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181*, 281–292.e286. [[CrossRef](#)] [[PubMed](#)]
8. Hou, Y.J.; Chiba, S.; Halfmann, P.; Ehre, C.; Kuroda, M.; Dinno, K.H., III; Leist, S.R.; Schäfer, A.; Nakajima, N.; Takahashi, K.; et al. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **2020**, *370*, 1464–1468. [[CrossRef](#)]
9. Rotondo, J.C.; Martini, F.; Maritati, M.; Mazziotta, C.; Di Mauro, G.; Lanzillotti, C.; Barp, N.; Gallerani, A.; Tognon, M.; Contini, C. SARS-CoV-2 Infection: New Molecular, Phylogenetic, and Pathogenetic Insights. Efficacy of Current Vaccines and the Potential Risk of Variants. *Viruses* **2021**, *13*, 1687. [[CrossRef](#)] [[PubMed](#)]
10. Hodcroft, E.B.; Zuber, M.; Nadeau, S.; Vaughan, T.G.; Crawford, K.H.; Althaus, C.L.; Reichmuth, M.L.; Bowen, J.E.; Walls, A.C.; Corti, D.; et al. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* **2021**, *595*, 707–712. [[CrossRef](#)] [[PubMed](#)]
11. Davies, N.G.; Abbott, S.; Barnard, R.C.; Jarvis, C.I.; Kucharski, A.J.; Munday, J.D.; Pearson, C.A.B.; Russell, T.W.; Tully, D.C.; Washburne, A.D.; et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **2021**, *372*, eabg3055. [[CrossRef](#)]
12. Tegally, H.; Wilkinson, E.; Lessells, R.J.; Giandhari, J.; Pillay, S.; Msomi, N.; Mlisana, K.; Bhiman, J.N.; von Gottberg, A.; Walaza, S.; et al. Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat. Med.* **2021**, *27*, 440–446. [[CrossRef](#)] [[PubMed](#)]
13. Wang, P.; Nair, M.S.; Liu, L.; Iketani, S.; Luo, Y.; Guo, Y.; Wang, M.; Yu, J.; Zhang, B.; Kwong, P.D.; et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **2021**, *593*, 130–135. [[CrossRef](#)] [[PubMed](#)]
14. Garcia-Beltran, W.F.; Lam, E.C.; St Denis, K.; Nitido, A.D.; Garcia, Z.H.; Hauser, B.M.; Feldman, J.; Pavlovic, M.N.; Gregory, D.J.; Poznansky, M.C.; et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* **2021**, *184*, 2372–2383.e2379. [[CrossRef](#)]
15. Voloch, C.M.; da Silva Francisco, R., Jr.; de Almeida, L.G.P.; Cardoso, C.C.; Brustolini, O.J.; Gerber, A.L.; Guimarães, A.P.D.C.; Mariani, D.; da Costa, R.M.; Ferreira, O.C., Jr.; et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J. Virol.* **2021**, *95*. [[CrossRef](#)]
16. Peng, J.; Liu, J.; Mann, S.A.; Mitchell, A.M.; Laurie, M.T.; Sunshine, S.; Pilarowski, G.; Ayscue, P.; Kistler, A.; Vanaerschot, M.; 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**, *31*, ciab283. [[CrossRef](#)] [[PubMed](#)]
17. Motozono, C.; Toyoda, M.; Zahradnik, J.; Saito, A.; Nasser, H.; Tan, T.S.; Ngare, I.; Kimura, I.; Uriu, K.; Kosugi, Y.; et al. SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. *Cell Host Microbe* **2021**, *29*, 1124–1136. [[CrossRef](#)]
18. Corti, D.; Purcell, L.A.; Snell, G.; Veesler, D. Tackling COVID-19 with neutralizing monoclonal antibodies. *Cell* **2021**, *184*, 3086–3108. [[CrossRef](#)]

19. Hoffmann, M.; Hofmann-Winkler, H.; Kruger, N.; Kempf, A.; Nehlmeier, I.; Graichen, L.; Arora, P.; Sidarovich, A.; Moldenhauer, A.S.; Winkler, M.S.; et al. SARS-CoV-2 variant B.1.617 is resistant to bamlanivimab and evades antibodies induced by infection and vaccination. *Cell Rep.* **2021**, *36*, 109415. [[CrossRef](#)]
20. Bhojar, R.C.; Jain, A.; Sehgal, P.; Divakar, M.K.; Sharma, D.; Imran, M.; Jolly, B.; Ranjan, G.; Rophina, M.; Sharma, S.; et al. High throughput detection and genetic epidemiology of SARS-CoV-2 using COVIDSeq next-generation sequencing. *PLoS ONE* **2021**, *16*, e0247115. [[CrossRef](#)] [[PubMed](#)]
21. Harilal, D.; Ramaswamy, S.; Loney, T.; Suwaidi, H.A.; Khansaheb, H.; Alkhaja, A.; Varghese, R.; Deesi, Z.; Nowotny, N.; Alsheikh-Ali, A.; 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. [[CrossRef](#)] [[PubMed](#)]
22. Bal, A.; Destras, G.; Gaymard, A.; Bouscambert-Duchamp, M.; Valette, M.; Escuret, V.; Frobert, E.; Billaud, G.; Trouillet-Assant, S.; Cheynet, V.; 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. [[CrossRef](#)] [[PubMed](#)]
23. Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B.; et al. Real-time DNA sequencing from single polymerase molecules. *Science* **2009**, *323*, 133–138. [[CrossRef](#)] [[PubMed](#)]
24. Vellas, C.; Del Bello, A.; Debard, A.; Steinmeyer, Z.; Tribaudeau, L.; Ranger, N.; Jeanne, N.; Martin-Blondel, G.; Delobel, P.; Kamar, N.; et al. Influence of treatment with neutralizing monoclonal antibodies on the SARS-CoV-2 nasopharyngeal load and quasispecies. *Clin. Microbiol. Infect.* **2021**. [[CrossRef](#)]
25. Ko, S.H.; Bayat Mokhtari, E.; Mudvari, P.; Stein, S.; Stringham, C.D.; Wagner, D.; Ramelli, S.; Ramos-Benitez, M.J.; Strich, J.R.; Davey, R.T., Jr.; 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. [[CrossRef](#)]
26. Sun, F.; Wang, X.; Tan, S.; Dan, Y.; Lu, Y.; Zhang, J.; Xu, J.; Tan, Z.; Xiang, X.; Zhou, Y.; et al. SARS-CoV-2 Quasispecies Provides an Advantage Mutation Pool for the Epidemic Variants. *Microbiol. Spectr.* **2021**, *9*, e0026121. [[CrossRef](#)] [[PubMed](#)]
27. Eden, J.S.; Rockett, R.; Carter, I.; Rahman, H.; de Ligt, J.; Hadfield, J.; Storey, M.; Ren, X.; Tulloch, R.; Basile, K.; et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* **2020**, *6*, veaa027. [[CrossRef](#)] [[PubMed](#)]
28. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [[CrossRef](#)]
29. Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acid. Res.* **2002**, *30*, 3059–3066. [[CrossRef](#)]
30. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534. [[CrossRef](#)]
31. Harvey, W.T.; Carabelli, A.M.; Jackson, B.; Gupta, R.K.; Thomson, E.C.; Harrison, E.M.; Ludden, C.; Reeve, R.; Rambaut, A.; Consortium, C.-G.U.; et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **2021**, *19*, 409–424. [[CrossRef](#)] [[PubMed](#)]
32. Greaney, A.J.; Starr, T.N.; Gilchuk, P.; Zost, S.J.; Binshtein, E.; Loes, A.N.; Hilton, S.K.; Huddleston, J.; Eguia, R.; Crawford, K.H.D.; et al. Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host Microbe* **2021**, *29*, 44–57.e49. [[CrossRef](#)] [[PubMed](#)]
33. Jangra, S.; Ye, C.; Rathnasinghe, R.; Stadlbauer, D.; Krammer, F.; Simon, V.; Martinez-Sobrido, L.; Garcia-Sastre, A.; Schotsaert, M. SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *Lancet Microbe* **2021**, *2*, e283–e284. [[CrossRef](#)]
34. Paiva, M.H.S.; Guedes, D.R.D.; Docena, C.; Bezerra, M.F.; Dezordi, F.Z.; Machado, L.C.; Krokovsky, L.; Helvecio, E.; da Silva, A.F.; Vasconcelos, L.R.S.; et al. Multiple Introductions Followed by Ongoing Community Spread of SARS-CoV-2 at One of the Largest Metropolitan Areas of Northeast Brazil. *Viruses* **2020**, *12*, 1414. [[CrossRef](#)]
35. Planas, D.; Bruel, T.; Grzelak, L.; Guivel-Benhassine, F.; Staropoli, I.; Porrot, F.; Planchais, C.; Buchrieser, J.; Rajah, M.M.; Bishop, E.; et al. Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. *Nat. Med.* **2021**, *27*, 917–924. [[CrossRef](#)] [[PubMed](#)]
36. Planas, D.; Veyer, D.; Baidaliuk, A.; Staropoli, I.; Guivel-Benhassine, F.; Rajah, M.M.; Planchais, C.; Porrot, F.; Robillard, N.; Puech, J.; et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* **2021**, *596*, 276–280. [[CrossRef](#)]
37. Gaymard, A.; Bosetti, P.; Feri, A.; Destras, G.; Enouf, V.; Andronico, A.; Burrell, S.; Behillil, S.; Sauvage, C.; Bal, A.; et al. Early assessment of diffusion and possible expansion of SARS-CoV-2 Lineage 20I/501Y.V1 (B.1.1.7, variant of concern 202012/01) in France, January to March 2021. *Euro Surveill* **2021**, *26*, 2100133. [[CrossRef](#)] [[PubMed](#)]
38. Alizon, S.; Haim-Boukobza, S.; Foulongne, V.; Verdurme, L.; Trombert-Paolantoni, S.; Lecorche, E.; Roquebert, B.; Sofonea, M.T. Rapid spread of the SARS-CoV-2 Delta variant in some French regions, June 2021. *Eurosurveillance* **2021**, *26*, 2100573. [[CrossRef](#)]
39. Ramesh, S.; Govindarajulu, M.; Parise, R.S.; Neel, L.; Shankar, T.; Patel, S.; Lowery, P.; Smith, F.; Dhanasekaran, M.; Moore, T. Emerging SARS-CoV-2 Variants: A Review of Its Mutations, Its Implications and Vaccine Efficacy. *Vaccines* **2021**, *9*, 1195. [[CrossRef](#)]
40. Wang, R.; Chen, J.; Gao, K.; Wei, G.W. Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries. *Genomics* **2021**, *113*, 2158–2170. [[CrossRef](#)]