Design of Specific Primer Sets for the Detection of B.1.1.7, B.1.351 and P.1 SARS-CoV-2 Variants using Deep Learning

Alejandro Lopez-Rincon[1,7,*], Carmina A. Perez-Romero[2], Alberto Tonda[3], Lucero Mendoza-Maldonado[4], Eric Claassen[5], Johan Garssen[1,6], and Aletta D. Kraneveld[1]

[1]Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, Universiteitsweg 99, 3584 CG Utrecht, the Netherlands
[2]Departamento de Investigación, Universidad Central de Queretaro (UNICEQ), Av. 5 de Febrero 1602, San Pablo, 76130 Santiago de Querétaro, Qro., Mexico
[3]UMR 518 MIA-Paris, INRAE, c/o 113 rue Nationale, 75103, Paris, France
[4]Hospital Civil de Guadalajara "Dr. Juan I. Menchaca". Salvador Quevedo y Zubieta 750, Independencia Oriente, C.P. 44340 Guadalajara, Jalisco, México
[5]Athena Institute, Vrije Universiteit, De Boelelaan 1085, 1081 HV Amsterdam, the Netherlands
[6]Department Immunology, Danone Nutricia research, Uppsalalaan 12, 3584 CT Utrecht, the Netherlands
[7]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Universiteitsweg 100, 3584 CG Utrecht
[*]a.lopezrincon@uu.nl

ABSTRACT

As the COVID-19 pandemic persists, new SARS-CoV-2 variants with potentially dangerous features have been identified by the scientific community. Variant B.1.1.7 lineage clade GR from Global Initiative on Sharing All Influenza Data (GISAID) was first detected in the UK, and it appears to possess an increased transmissibility. At the same time, South African authorities reported variant B.1.351, that shares several mutations with B.1.1.7, and might also present high transmissibility. Even more recently, a variant labeled P.1 with 17 non-synonymous mutations was detected in Brazil. In such a situation, it is paramount to rapidly develop specific molecular tests to uniquely identify, contain, and study new variants. Using a completely automated pipeline built around deep learning techniques, we design primer sets specific to variant B.1.1.7, B.1.351, and P.1, respectively. Starting from sequences openly available in the GISAID repository, our pipeline was able to deliver the primer sets in just under 16 hours for each case study. In-silico tests show that the sequences in the primer sets present high accuracy and do not appear in samples from different viruses, nor in other coronaviruses or SARS-CoV-2 variants. The presented methodology can be exploited to swiftly obtain primer sets for each independent new variant, that can later be a part of a multiplexed approach for the initial diagnosis of COVID-19 patients. Furthermore, since our approach delivers primers able to differentiate between variants, it can be used as a second step of a diagnosis in cases already positive to COVID-19, to identify individuals carrying variants with potentially threatening features.

Introduction

SARS-CoV-2 mutates with an average evolutionary rate of 10^{-4} nucleotide substitutions per site each year[1]. As the pandemic of SARS-CoV-2 continues to affects the planet, researchers and public health officials around the world constantly monitor the virus for acquired mutations that may lead to higher threat for developing COVID-19.

On December 14, 2020, Public Health authorities in England reported a new SARS-CoV-2 variant referred to as Variant Under Investigation (VUI-202012/01)[2–4], which belongs to the B.1.1.7 lineage clade GR from GISAID (Global Initiative on Sharing All Influenza Data)[2,5,6], Nextstrain clade 20I/501Y.V1[7]. This variant presents 14 non-synonymous mutations, 6 synonymous mutations, and 3 deletions. The multiple mutations present in the viral RNA encoding for the spike protein (S) are of most concern, such as the deletion △69-70 (△21765-21770), deletion △144 (△21991-21993), N501Y (A23063T), A570D (C23271A), P681H (C23604A), D614G (A23403G), T716I (C23709T), S982A (T24506G), D1118H (G24914C)[3,6]. The SARS-CoV-2 S protein mutation N501Y alters the protein interactions involved in receptor binding domain. The N501Y mutation has been shown to enhance affinity with the host cells ACE2 receptor[6,8] and to be more infectious in mice[9]. The mutation P681H is situated in key residues of the S protein needed for viral transmition and viral entry in lung cells[6,10]. Even

if the clinical outcomes and additive effects of the mutations present on the B.1.1.7 SARS-CoV-2 variant are still unknown, its rate of transmission has been estimated to be 56%-70% higher, and its reproductive number (Rt) seems to be up to 0.4 higher[2,11]. The presence of the B.1.1.7 variant has been rapidly increasing in the UK[6], and has now been identified in other 36 countries around the world[12].

In parallel, on December 18, 2020, South African authorities reported a new variant of SARS-CoV-2 that has been rapidly spreading across 3 of their provinces and displacing other variants: this variant was named 501Y.V2, also known as B.1.351 lineage, or 20H/501Y.V2 Nextrain clade[4,7,13]. The B.1.351 variant harbours 19 mutations, with 9 of them situated in the Spike protein: mutations N501Y (A23063T), E484K (G23012A), and K417N (G22813T) are at key residues in the RBD domain of S protein, L18F (*), D80A (A21801C), and D215G (*) are in the N-terminal domain, A701V (C23664T) in loop 2, and D614G (A23403G)[13]. Although the B.1.351 variant also has the N501Y mutation in the Spike protein, similarly to the B.1.1.7 variant in the UK, the B.1.351 variant arose independently in a different SARS-CoV-2 lineage, which forms part of the 20H/501Y.V2 phylogenetic clade in Nextstrain[7]. Although uncommon in SARS-CoV-2 variant strains, the E484K mutation has been shown to moderately enhance binding affinity of the ACE2 receptor[8]. Mutation K417N is located in a key contact between the S protein and the human ACE2 receptor: a previous mutation in this same residue (V404 $\rightarrow$ K417) contributed to the enhanced affinity of SARS-CoV-2 to the ACE2 receptor when compared to SARS-CoV[14,15].Although the rate of transmission of the B.1.351 variant in South Africa is still unknown, there has been a steady rise of daily cases as well as an increase in excess deaths per week[13]. The B.1.351 variant has now been identified in other 13 countries around the world[12].

Recently, a new variant SARS-CoV-2 from has been identified by scientist and clinicians in Brazil, and named P.1 variant or 20J/501Y.V3 Nextrain clade[12,16,17]. The P.1 variant harbours 17 unique amino acid changes, three deletions, four synonymous mutations, and one 4-nucleotide insertion. The P.1 Variant harbours similar mutations in the Spike protein to those find in the B.1.1.7 and B.1.351 variant like the N501Y (A23063T), E484K (G23012A), K417N (G22813T) and D614G (A23403G). An although the rate of infection and mortality of the P.1 variant is still unknown it has been identified in different areas in Brazil including the Amazonia and Rio de Janeiro, as well internationally in Japan[16,17].

Apart from B.1.1.7 in England, B.1.351 in South Africa, and P.1 in Brazil, other variants of SARS-CoV-2 carrying mutations for N501Y have also independently arisen in other parts of Australia and the USA[5,7,17–19]. The mutations K417N and E484K found in both B.1.351 and P.1 variants have become of great global concern as they have been show to slightly or significantly reduce convalescent serum neutralization[20,21]. One reinfection has been linked to P.1 (E484K mutation)[22], raising concerns about the current monoclonal/polyclonal antibodies used for treating Covid-19 disease as well as the new mRNA vaccines efficacy in combating the spread of this new SARS-CoV-2 variants[23,24]. Given the rapid spread of this new variants, and their concerns to the global health, it has become of utmost important for countries to be able to rapidly identify this new variants.

Several diagnostic kits have been proposed and developed to diagnose SARS-CoV-2 infections. Most kits rely on the amplification of one or several genes of SARS-CoV-2 by real-time reverse transcriptase-polymerase chain reaction (RT-PCR)[25,26]. Recently, Public Health England was able to identify the B.1.1.7 SARS-CoV-2 variant through their national surveillance system which allowed them to noticing the rise in Covid-19 positive cases in South Easter England and through their the increase in S-gene target failure (negative results) from the otherwise positive target genes (N, ORF1ab) in their three target gene assay in their RT-PCR diagnostic tests, and the random whole genome sequencing of some of this positive samples[2]. In the case of South Africa, they where able to identify the rise of B.1.351 SARS-CoV-2 variant through their the increase in positive Covid-19 cases and deaths reports in 3 of their provinces and random whole genome sequencing of positive Covid-19 samples, through their national diagnosis and surveillance system[13,27,28]. However, to the best of our knowledge, currently no specific test exists or has been developed to identify the B.1.1.7, B.1.351, or P.1 SARS-CoV-2 variants.

In a previous work[29], we developed a methodology based on deep learning, able to generate a primer set specific to SARS-CoV-2 in an almost fully automated way. When compared to other primers sets suggested by GISAID, our approach proved to deliver competitive accuracy and specificity. Our results, both *in-silico* and with patients, yielded 100% specificity, and sensitivity similar to widely-used diagnostic qPCR methods. One of the main advantages of the proposed methodology was its ease of adaptation to different viruses or mutations, given a sufficiently large number of whole viral RNA sequences. In this work we improved the existing semi-automated methodology, making the pipeline completely automated, and created primer sets specific for the SARS-CoV-2 variants B.1.1.7, B.1.351, or P.1 in under 16 hours for each case study. The developed primer sets, tested *in-silico*, proved not only to be specific, but also to be able to distinguish between the different variants. With this new result, we believe that our method represents a rapid and effective diagnostic tool, able to support medical experts both during the current pandemic, as new variants of SARS-CoV-2 may emerge, and possibly during future ones, as still unknown virus strains might surface.

Results

Variant B.1.1.7

As explained in the Methods section, the first step is to run a Convolution Neural Network (CNN) classifier on the data. This yields a classification accuracy of 99.66% on the test set. Secondly, from an analysis of the features constructed by the CNN, we extract 7,127 features, corresponding to 21-bps sequences. Next, we run a state-of-the-art stochastic feature selection algorithm 10 times, to uncover the most meaningful '21-bps' features for the identification of variant B.1.1.7: as shown in Fig. 1a, while the best result corresponds to a set of 16 '21-bps' features, using only one is enough to obtain over 99% accuracy, a satisfying outcome for our objective.

These features represent good candidates for forward primers. From the 10 runs, we get 10 different '21-bps' features: 5 out of the 10 point to mutation Q27stop (C27972T), 3 point to mutation I2230T (T6954C), and 2 to a synonymous mutation (C16176T). Using Primer3Plus we compute a primer set for each of the 10 features, using sequence EPI_ISL_601443 (see Table 1) as a reference. Only the two '21-bps' features that include mutation C16176T are suitable for a forward primer. The two features are **ACC TCA AGG TAT TGG GAA CCT** and **CAC CTC AAG GTA TTG GGA ACC**: it is easy to notice that the two features are actually part of the same sequence, just displaced by a bps, and therefore generate the same reverse primer **CAT CAC AAC CTG GAG CAT TG** with Tm 58.8°C and 60.6°C respectively for the forward primer, and 60.1°C for the reverse primer.

Using just feature **ACC TCA AGG TAT TGG GAA CCT** it is possible to build a simple rule-based classifier, that only checks whether the feature is present in a sample. For further validation, we test the rule-based classifier on the 893 sequences of the test set (see Fig. 4a), yielding an area under the curve (AUC) of 0.98, a result considered as an excellent diagnostic accuracy (AUC in the 0.9-1.0 range[30,31]), see Fig. 2a.

We then validate our results on 487 samples of other coronaviruses, obtained from the National Genomics Data Center (NGDC)[32]. Checking for the presence of feature **ACC TCA AGG TAT TGG GAA CCT** confirms that it is exclusive to B.1.1.7 SARS-CoV-2 samples, with no appearance in any other coronavirus sample. Further validation on 20,571 samples belonging to other non-corona viruses from the National Center for Biotechnology Information (NCBI)[33], shows no appearance of the sequence in any other virus sample.

For further analysis, we check the presence of the signature mutations T1001I, A1708D, I2230T, N501Y, A570D, P681H, T716I, S982A, D1118H, Q27stop, R52I, Y73C, and S235F of variant B.1.1.7[34]. To verify the presence of mutations, we generate 21-bps sequences, with 10 bps before and after the mutation, and search for their presence: e.g., mutation N501Y (A23063T) corresponds to sequence **CCAACCCACT T ATGGTGTTGG**.

Using the generated '21-bps' sequences for the mutations, we can also test them as forward primers using Primer3Plus, which yields **TGA TAT CCT TGC ACG TCT TGA** in spike gene (S982A) as the only possible forward primer candidate (see Table 1). The Tm of the forward primer is 59.3°C and the reverse primer sequence will be **GAG GTG CTG ACT GAG GGA AG** with Tm 60.0°C.

Table 1. Result of Primer3Plus[35] evaluation of the sequences uncovered by deep learning, and obtained through checking for the presence of mutations, when used as forward primers in reference sequence EPI_ISL_601443 for the B.1.1.7 variant.

Deep Learning Generated		
21-bps Feature	**Primer3Plus Result**	
TTCTAAACTGATAAATATTAC	Unacceptable GC content/Tm too low	
TTAACATCAACCATATGTAGT	Tm too low/High end self complementarity	
GATAAATATTACAATTTGGTT	Unacceptable GC content/Tm too low	
ATATTACAATTTGGTTTTTAC	Unacceptable GC content/Tm too low	
CTGATAAATATTACAATTTGG	Tm too low	
CTGATAAATATTACAATTTGG	Tm too low	
ACCTCAAGGTATTGGGAACCT	**CATCACAACCTGGAGCATTG**	
CACCTCAAGGTATTGGGAACC	**CATCACAACCTGGAGCATTG**	
TAACATCAACCATATGTAGTT	Tm too low/High end self complementarity	
AATATTACAATTTGGTTTTTA	Unacceptable GC content/Tm too low	
Mutation Generated		
21-bps Feature	**Primer3Plus Result**	**Mutation**
CAGACAACTATTATTCAAACA	Tm too low	T1001I
GGTGAAGCTGATAACTTTTGT	Tm too low	A1708D
ATAAATATTACAATTTGGTTT	Unacceptable GC content/Tm too low	I2230T
CCAACCCACTTATGGTGTTGG	High self complementarity/High end self complementarity	N501Y
AGAGACATTGATGACACTACT	Tm too low	A570D
ACTAATTCTCATCGGCGGGCA	Tm too high/High 3’ stability	P681H
GCCATACCCATAAATTTTACT	Tm too low/High end self complementarity	T716I
TGATATCCTTGCACGTCTTGA	**GAGGTGCTGACTGAGGGAAG**	S982A
CATTACTACACACAACACATT	Tm too low	D1118H
GTCATGTACTTAACATCAACC	Tm too low	Q27stop
GTAGGAGCTATAAAATCAGCA	Tm too low	R52I
CCCATTCAGTGCATCGATATC	High end self complementarity	Y73C
AGCAAAATGTTTGGTAAAGGC	High 3’ stability	S235F

Variant B.1.351

As shown in Figure 4b, that summarizes the experiments for this South African variant, the first step is to run a CNN classifier on the data. This yields a classification accuracy of 99.88% on the test set. Next, we translate the CNN weights into 6,069 features, each representing a 21-bps sequence. As for the B.1.1.7 variant, we run the feature selection algorithm 10 times (Fig. 1b), which results in only one meaningful feature achieving over 99% accuracy. The 10 runs point to the same mutation, K417N, with 6 different sequences but only one acceptable as a primer candidate (Table 2). The resulting sequence is **CTC CAG GGC AAA CTG GAA ATA**, with Tm 61.3°C , with a reverse primer **TGC TAC CGG CCT GAT AGA TT** with Tm 59.7°C.

Using sequence **CTC CAG GGC AAA CTG GAA ATA**, we run a rule-based classifier that simply checks for its presence in a sample. We test the rule-based classifier on the 926 sequences of the test set (see Fig. 4b), yielding an AUC of 0.99, again considered as excellent diagnostic accuracy[30,31], see Fig. 2b. The sequence does not appear in the 487 samples of other coronavirus obtained from NGDC, nor in the 20,571 samples of other non-corona viruses collected from the NCBI repository.

As we did for the B.1.1.7 variant, we check whether 21-bps sequences containing the signature mutations T265I, K1655N, K3353R, D80A, K417N, EA84K, N501Y, A701V, Q57H, S171L, P71L, and T205I of variant B.1.351[34] could function as a forward primer. From the Primer3Plus results reported in Table 2, sequence **CTG TTT TTC ATA GCG CTT CCA** is the only one that could potentially be a primer candidate (Tm 60.4 °C). This primer is in the Q57H mutation, with reverse primer **AGA GAA AAG GGG CTT CAA GG** (Tm 59.8 °C).

Table 2. Result of Primer3Plus[35] evaluation of the sequences uncovered by deep learning, and obtained through analysis of the mutation, when used as forward primers in reference sequence EPI_ISL_678597 for the B.1.351 variant.

Deep Learning Generated		
21-bps Feature	**Primer3Plus Result**	
ATATTGCTGATTATAATTATA	Unacceptable GC content/Tm too low/High self complementarity	
AACTGGAAATATTGCTGATTA	Tm too low	
AAACTGGAAATATTGCTGATT	Tm too low/High end self complementarity	
CTCCAGGGCAAACTGGAAATA	**TGCTACCGGCCTGATAGATT**	
AATATTGCTGATTATAATTAT	Unacceptable GC content/Tm too low/High end self complementarity	
TCCAGGGCAAACTGGAAATAT	High end self complementarity	
Mutation Generated		
21-bps Feature	**Primer3Plus Result**	**Mutation**
AAATTTGACATCTTCAATGGG	Tm too low/High 3' stability	T265I
ACACTAAAAATTGGAAATACC	Tm too low	K1655N
CTTAAGCTTAGGGTTGATACA	Tm too low	K3353R
AAGAGGTTTGCTAACCCTGTC	High end self complementarity	D80A
AAACTGGAAATATTGCTGATT	Tm too low/High end self complementarity	K417N
TAATGGTGTTAAAGGTTTTAA	Tm too low/High end self complementarity	EA84K
CCAACCCACTTATGGTGTTGG	High self complementarity/High end self complementarity	N501Y
TCACTTGGTGTAGAAAATTCA	Tm too low	A701V
CTGTTTTTCATAGCGCTTCCA	**AGAGAAAAGGGGCTTCAAGG**	Q57H
GTCATTACTTTAGGTGATGGC	Tm too low/High end self complementarity/High 3' stability	S171L
TCTAGAGTTCTTGATCTTCTG	Tm too low	P71L
AGTAGGGGAATTTCTCCTGCT	High self complementarity/High end self complementarity	T205I

Variant P.1

As in the other case studies, the first step is to run a CNN classifier on the data. This yields a classification accuracy of 100%, which could potentially represent overfitting on the training set, as the scarce number of samples available made it unpractical to create a separate test set. Next, we translate the CNN weights into 727 features, each representing a 21-bps sequence. We run the feature selection algorithm 10 times (Fig. 1c), which results in only one meaningful feature achieving 100% accuracy. The 10 runs point to the same sequence, around the synonymous mutation T733C **ACT GAT CCT TAT GAA GAC TTT**. The sequence cannot be used as a primer, as the Tm is too low. To compensate, we displaced the sequence two bps to the left and added a bps to rise the Tm. This procedure gave us sequence **GGC ACT GAT CCT TAT GAA GAC T**, of size 22 bps and Tm 57.4 °C, with reverse primer **TTC GGA CAA AGT GCA TGA AG**, with Tm 60.0 °C.

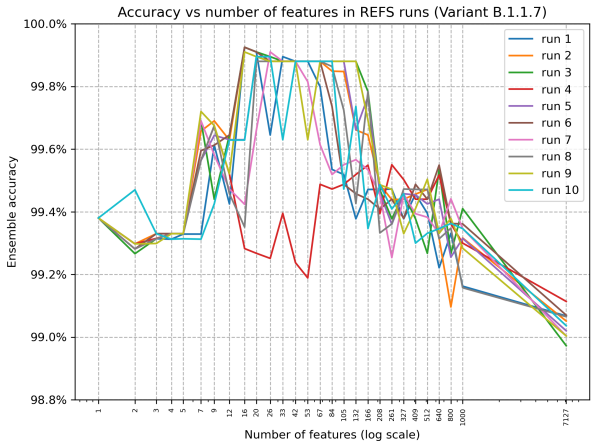
Since we miss a test set for this variant, again due to the few P.1 sequences available, we cannot compute a ROC curve like

in the other cases. However, we can validate our results by verifying that the sequence does not appear in the 487 samples of other coronavirus obtained from NGDC, nor in the 20,571 from other non-corona viruses collected from the NCBI repository.

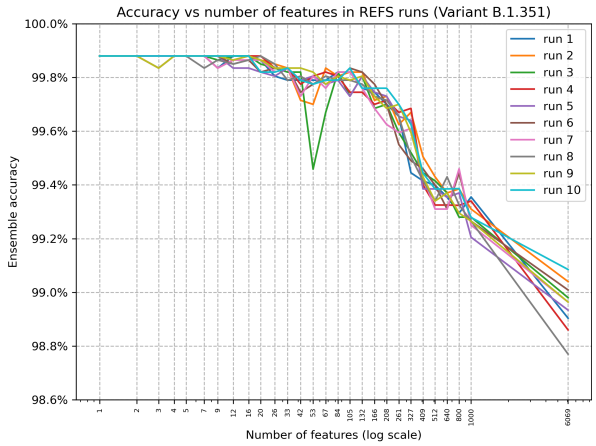
As for the other variants, we check whether '21-bps' sequences containing signature mutations S1188L, K1795Q, L18F, T20N, P26S, D138Y, R190S, K417T, EA84K, N501Y, H655Y, T1027I, E92K, and P80R of variant P.1[16] could function as a forward primers. We put mutation T20N and L18F into the same '21-bps' sequence, given their proximity. From the Primer3Plus results reported in Table 3, 3 of the generated sequences could be primers candidates.

Table 3. Result of the Primer3Plus[35] evaluation of the sequences obtained through the analysis of different mutations, when used as forward primers in reference sequence EPI_ISL_804814 for the P.1 variant.

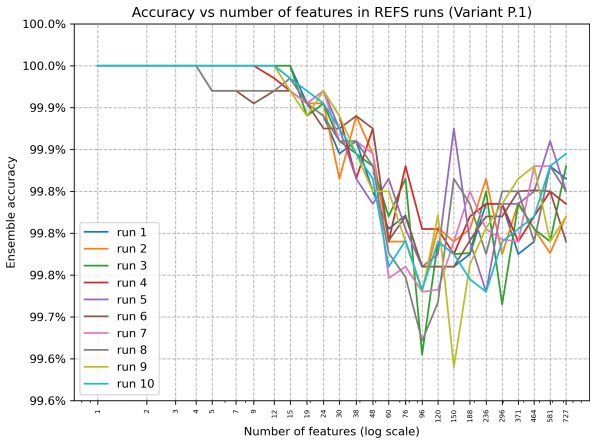
Deep Learning Generated		
21-bps Feature	**Primer3Plus Result**	
ACTGATCCTTATGAAGACTTT	Tm too low	
GGCACTGATCCTTATGAAGACT	**TTCGGACAAAGTGCATGAAG**	
Mutation Generated		
21-bps Feature	**Primer3Plus Result**	**Mutation**
AAACTTGTTTTAAGCTTTTTG	Tm too low	S1188L
ACAAGCTACACAATATCTAGT	Tm too low/High end self complementarity	K1795Q
AATTTTACAAACAGAACTCAA	Tm too low	T20N/L18F
TCAATTACCCTCTGCATACAC	Tm too low	P26S
ATTTTGTAATTATCCATTTTT	Unacceptable GC content/Tm too low	D138Y
AAAATCTTAGTGAATTTGTGT	Tm too low	R190S
CAAACTGGAACGATTGCTGAT	**TGCTACCGGCCTGATAGATT**	K417T
TAATGGTGTTAAAGGTTTTAA	Tm too low/High end self complementarity	EA84K
CCAACCCACTTATGGTGTTGG	High self complementarity/High end self complementarity	N501Y
AGGGGCTGAATATGTCAACAA	**TGAATTTTCTGCACCAAGTGA**	H655Y
CTTGCTGCTATTAAAATGTCA	Tm too low/High end self complementarity	T1027I
TAATTGCCAGAAACCTAAATT	Tm too low/High end self complementarity	E92K
AATAGCAGTCGAGATGACCAA	**GCCGTCTTTGTTAGCACCAT**	P80R



(a)

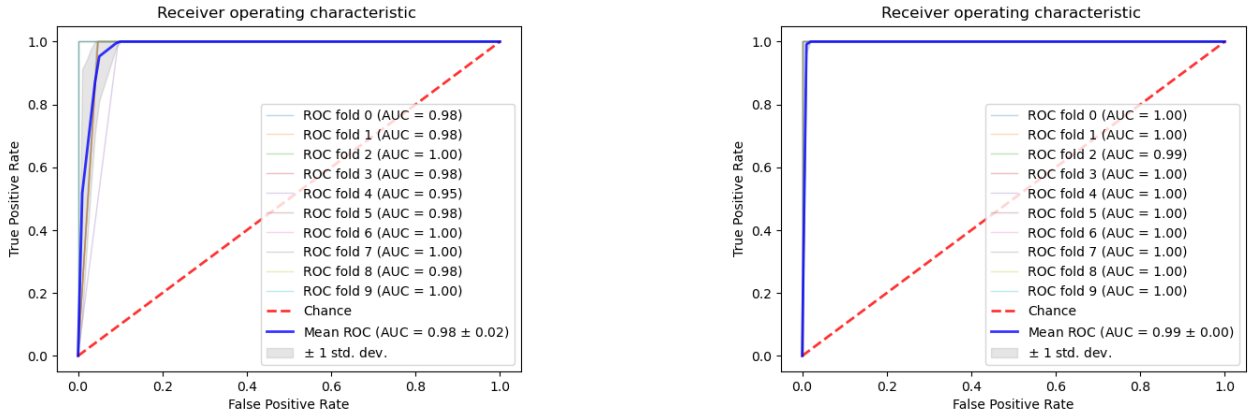


(b)



(c)

Figure 1. 1a) 10 runs of the recursive ensemble feature selection algorithm, with accuracy evaluated on 10% of the training set for the B.1.1.7 Variant. 1b) 10 runs of the recursive ensemble feature selection algorithm, with accuracy evaluated on 10% of the training set for the B.1.351 Variant. 1c) 10 runs of the recursive ensemble feature selection algorithm, with accuracy evaluated on 10% of the training set, including all available sequences labeled as P.1, for the P.1 Variant.



(a) **(b)**

Figure 2. 2a) ROC curve of a simple rule-based classifier checking the presence of feature **ACC TCA AGG TAT TGG GAA CCT**, in a 10-fold cross-validation for the 893 sequences of the test set. 2b) ROC curve of a simple rule-based classifier checking the presence of feature **CTC CAG GGC AAA CTG GAA ATA**, in a 10-fold cross-validation for the 926 sequences of the test set.

Discussion

The two primer sets generated for the B.1.1.7 variant, either by deep learning or by using sequences around known mutations, show almost the same accuracy and similar specificity: the frequency of appearance of the mutation-based primer is better by 0.05%, and both of them do not appear in other SARS-CoV-2 variants, see Table 4. In contrast, the primer set generated with deep learning for the B.1.351 variant is clearly more specific (0.04%), than the mutation-based primer set (21.84%) using the frequency of appearance in other variants. Interestingly, by analyzing the three samples representing the 0.04% error, we notice that all three of them are from South Africa, and they are labeled as B.1, B.1, B.1.302 in GISAID, which could potentially indicate a labeling error. Similarly, the generated forward primer for B.1.1.7 variant appears in 2,007 of 2,104 sequences, with a frequency of 95.39%. Nevertheless, a further analysis shows that only 2,014 of the sequences labeled as B.1.1.7 present 5 or more of the 7 studied mutations, which could point to an error in the annotation of the variant inside the GISAID dataset, or several extra mutations in the generated 21-bps sequences. If we consider as proper B.1.1.7 variants only sequences that show 5 or more of the mutations, then our primer correctly identifies 2,095 out of 2,104 samples, for a final 99.57% accuracy. In the case of the P.1 variant, the deep learning generated primer appeared in one B.1.1.7 sample. In addition, we generated 3 mutation-based primers, but given the the frequency of appearance of the *Mut-1 BR R* primer (82.14%) and that 9 sequences from other variants were detected by the *Mut-2 BR F*, the best primer set appears to be the one that uses mutation P80R *Mut-3 BR*. It is important to remark that this last case study was tackled using just the 28 P.1 sequences currently available: as more data related to this variant becomes accessible, we will be able to improve and select the best primer pair. The proposed forward and reverse primers for all the three variants are summarized in Table 5.

Table 4. Frequency of appearance of the characteristic mutations for the UK (B.1.1.7), South African (B.1.351), and Brazilian (P.1) variants in B.1.1.7 sequences (2,104), B.1.351 sequences (337), P.1 sequences (28) and sequences of other variants (6,808). The candidate primers generated with deep learning are labeled with *Deep*, while the ones based on the mutations are labeled with *Mut*.

B.1.1.7 Variant				
Mutation/ Primer	**Other samples**	**B.1.1.7 samples**	**B.1351 samples**	**P.1 samples**
T1001I	0.01%	95.53%	0.00%	0.00%
A1708D	0.00%	91.83%	0.00%	0.00%
I2230T	0.01%	94.39%	0.00%	0.00%
N501Y	0.04%	94.34%	99.70%	82.14%
A570D	0.01%	95.67%	0.00%	0.00%
P681H	0.01%	95.72%	0.30%	0.00%
T716I	0.01%	95.29%	0.00%	0.00%
S982A	0.00%	95.44%	0.00%	0.00%
D1118H	0.00%	95.58%	0.00%	0.00%
Q27stop	0.04%	90.64%	0.30%	0.00%
R52I	0.00%	90.64%	0.00%	0.00%
Y73C	0.00%	90.78%	0.00%	0.00%
S235F	0.03%	95.58%	0.00%	0.00%
Deep UK F	*0.00%*	*95.39%*	*0.00%*	*0.00%*
Deep UK R	*99.76%*	*98.62%*	*98.81%*	*100.00%*
Mut UK F	*0.00%*	*95.44%*	*0.00%*	*0.00%*
Mut UK R	*99.91%*	*99.71%*	*100.00%*	*100.00%*

B.1.351 Variant				
Mutation/ Primer	**Other samples**	**B.1.1.7 samples**	**B.1351 samples**	**P.1 samples**
T265I	8.67%	0.00%	99.11%	0.00%
K1655N	0.24%	0.00%	98.22%	0.00%
K3353R	0.62%	0.29%	100.00%	0.00%
D80A	0.04%	0.00%	97.92%	0.00%
K417N	0.04%	0.00%	99.11%	0.00%
EA84K	0.07%	0.00%	99.70%	82.14%
N501Y	0.04%	94.34%	99.70%	82.14%
A701V	0.12%	0.05%	99.11%	0.00%
Q57H	21.84%	0.00%	99.11%	0.00%
S171L	0.19%	0.48%	78.64%	0.00%
P71L	0.06%	0.00%	98.81%	0.00%
T205I	0.47%	0.00%	99.70%	0.00%
Deep SA F	*0.04%*	*0.00%*	*99.11%*	*0.00%*
Deep SA R	*98.46%*	*98.34%*	*99.41%*	*82.14%*
Mut SA F	*21.84%*	*0.00%*	*99.11%*	*0.00%*
Mut SA R	*98.75%*	*99.14%*	*96.44%*	*96.43%*

P.1 Variant				
Mutation/ Primer	**Other samples**	**B.1.1.7 samples**	**B.1351 samples**	**P.1 samples**
S1188L	0.04%	0.00%	0.00%	100.00%
K1795Q	0.00%	0.00%	0.00%	100.00%
T20N/L18F	0.00%	0.00%	0.00%	100.00%
P26S	0.09%	0.00%	0.00%	96.43%
D138Y	0.10%	0.00%	0.00%	100.00%
R190S	0.00%	0.00%	0.00%	85.71%
K417T	0.00%	0.00%	0.00%	100.00%
EA84K	0.07%	0.00%	99.70%	82.14%
N501Y	0.04%	94.34%	99.70%	82.14%
H655Y	0.13%	0.00%	0.00%	96.43%
T1027I	0.09%	0.00%	0.00%	100.00%
E92K	0.07%	0.00%	0.00%	100.00%
P80R	0.00%	0.00%	0.00%	100.00%
Deep BR F	*0.00%*	*0.05%*	*0.00%*	*100.00%*
Deep BR R	*99.81%*	*4.18%*	*100.00%*	*100.00%*
Mut-1 BR F	*0.00%*	*0.00%*	*0.00%*	*100.00%*
Mut-1 BR R	*98.46%*	*98.34%*	*99.41%*	*82.14%*
Mut-2 BR F	*0.13%*	*0.00%*	*0.00%*	*96.43%*
Mut-2 BR R	*99.34%*	*99.57%*	*0.00%*	*100.00%*
Mut-3 BR F	0.00%	0.00%	0.00%	100.00%
Mut-3 BR R	98.66%	99.57%	99.11%	100.00%

A wide variety of diagnostic tests have been used by high-throughput national testing systems around the world, to monitor the SARS-CoV-2 infection[25]. The arising prevalence of new SARS-CoV-2 variants such as B.1.1.7, B.1.351, and P.1 has become of great concern, as most RT-PCR tests to date are not be able to distinguish these new variants, not being designed for such a purpose. Therefore, public health officials most rely on their current testing systems and whole viral RNA sequencing results to draw conclusions on the prevalence of new variants in their territories. An example of such case has been seen in the UK, where the increase of the B.1.1.7 SARS-CoV-2 variant infection in their population was identified only through an increase in the S-gene target failure in their three target gene assay (N+, ORF1ab+, S-), coupled with sequencing of the virus and RT-PCR amplicons products[2]. Researchers believe that the S-gene target failure occurs due to the failure of one of the RT-PCR probes to bind, as a result of the $\triangle$69-70 deletion in the SARS-CoV-2 spike protein, present in B.1.1.7[2]. This $\triangle$69-70 deletion, which affects its N-terminal domain, has been occurring in several different SARS-CoV-2 variants around the world[7,36] and has been associated with other spike protein receptor binding domain changes[6]. Due to the likeliness of mutations in the S-gene, assays relying solely on its detection are not recommended, and a multiplex approach is required[25,26,37]. This is consistent with other existing primer designs like CoV2R-3 in the S-gene[38], that will also yield negative results for the B.1.1.7 variant, as the reverse primer sequence is in the region of mutation P681H. A more in-depth analysis of S-dropout positive results can be found in Kidd et al.[39].

Table 5. Proposed forward and reverse primers for variant B.1.1.7, B.1.351 and P.1 of SARS-CoV-2.

B.1.1.7 Variant	
UtrechtU-UKF-1	5' ACCTCAAGGTATTGGGAACCT
UtrechtU-UKR-1	5' CATCACAACCTGGAGCATTG
UtrechtU-UKF-2	5' TGATATCCTTGCACGTCTTGA
UtrechtU-UKR-2	5' GAGGTGCTGACTGAGGGAAG
B.1.351 Variant	
UtrechtU-SAF	5' CTCCAGGGCAAACTGGAAATA
UtrechtU-SAR	5' TGCTACCGGCCTGATAGATT
P.1 Variant	
UtrechtU-BRF-1	5' GGCACTGATCCTTATGAAGACT
UtrechtU-BRR-1	5' TTCGGACAAAGTGCATGAAG
UtrechtU-BRF-2	5' AATAGCAGTCGAGATGACCAA
UtrechtU-BRR-2	5' GCCGTCTTTGTTAGCACCAT

Given the concern for the increase in prevalence of the new variants SARS-CoV2 B.1.1.7, B.1.351, and P.1, and their possible clinical implication in the ongoing pandemic, diagnosing and monitoring the prevalence of such variants in the general population will be of critical importance to help fight the pandemic and develop new policies. In this work, we propose possible primer sets that can be used to specifically identify the B.1.1.7, B.1.351, and possibly P.1 SARS-CoV2 variants. We believe that all the proposed primer sets can be employed in a multiplexed approach in the initial diagnosis of Covid-19 patients, or used as a second step of diagnosis in cases already verified positive to SARS-CoV-2, to identify individuals carrying the B.1.1.7, B.1.351, or P.1 variant. In this way, health authorities could better evaluate the medical outcomes of this patients, and adapt or inform new policies that could help curve the rise of variants of interest. Although the proposed primer sets delivered by our automated methodology will still require laboratory testing to be validated, our deep learning design can enable the timely, rapid, and low-cost operations needed for the design of new primer sets to accurately diagnose new emerging SARS-CoV-2 variants and other infectious diseases.

Data and Methods

Data

Variant B.1.1.7

From the GISAID repository we downloaded 10,712 SARS-CoV-2 sequences on December 23, 2020. After removing repeated sequences, we obtained a total of 2,104 sequences labeled as B.1.1.7, and 6,819 sequences from other variants, for a total of 8,923 samples. B.1.1.7 variant samples were assigned label 1, and the rest were assigned label 0 for the CNN training. The samples were divided in a training set of 8,030 samples (90%), and a test set of 893 samples (10%), using a stratified division, so that both training and test set retain the same proportion of classes. The 8,923 SARS-CoV-2 samples downloaded from the GISAID repository show 278 variants besides B.1.1.7.

Variant B.1.351

From the GISAID repository, we downloaded 326 sequences of the B.1.351 variant on January 7, 2021. We added the 326 sequences to the 8,923-sample dataset from the B.1.1.7 experiment, obtaining a total of 9,249 sequences, where we assigned the label 1 for sequences belonging to variant B.1.351, and 0 to the rest of the samples. Next, we divided the data so that 293 B.1.351 sequences ended up in the training set (8,323), and 33 in the test set (926 sequences total), using again a stratified division.

Variant P.1

From the GISAID repository, we downloaded 28 non-repeated sequences of the P.1 variant on January 19, 2021. We added the 28 sequences to the 8,323 sequences of several other variants, including B.1.1.7 and B.1.351, for a total of 8,351 sequences. We assigned label 1 to sequences belonging to variant P.1, and 0 to the rest of the samples. For this case, given the few P.1 samples available, all the available samples are used for training, with no separate test set.

Methods

Following the procedure described in Lopez et al.[29], there are 4 steps for the automated design of a specific primer for a virus: (i) run a CNN for the classification of the target virus against other strains, (ii) translate the CNN weights into 21-bps features,

(iii) perform feature selection to identify the most promising features, and (iv) carry out a primer simulation with Primer3Plus[35] for the features uncovered in the previous step. While in[29] the proposed pipeline was only partially automatic, and still required human interventions between steps, in this work all steps have been automatized, and the whole pipeline has been run with no human interaction. The experiments, from downloading the sequences to the final *in-silico* testing of the primers, take around 16 hours of computational time on a standard end-user laptop, for each variant considered.

In a first step, we train a convolution neural network (CNN) using the training and testing sequences obtained from GISAID. The architecture of the network is shown in Fig 3, and is the same as the one previously reported in[29]. Next, if the classification accuracy of the CNN is satisfying ($> 99\%$), using a subset of the training sequences we translate the CNN weights into '21-bps' features, necessary to differentiate between the targeted variant samples and all the others. The length of the features is set as 21 bps, as a normal length for primers to be used in PCR tests is usually 18-22 bps. Then, we apply recursive ensemble feature selection (REFS)[40,41] to obtain a reduced set of the most meaningful features that separate the two classes. Finally, we simulate the outcome of treating the most promising features obtained in the previous step as primers, using Primer3Plus[35] in the canonical sequences EPI_ISL_601443 for B.1.1.7[2], EPI_ISL_678597 for B.1.351 and EPI_ISL_804814 for P.1[16].

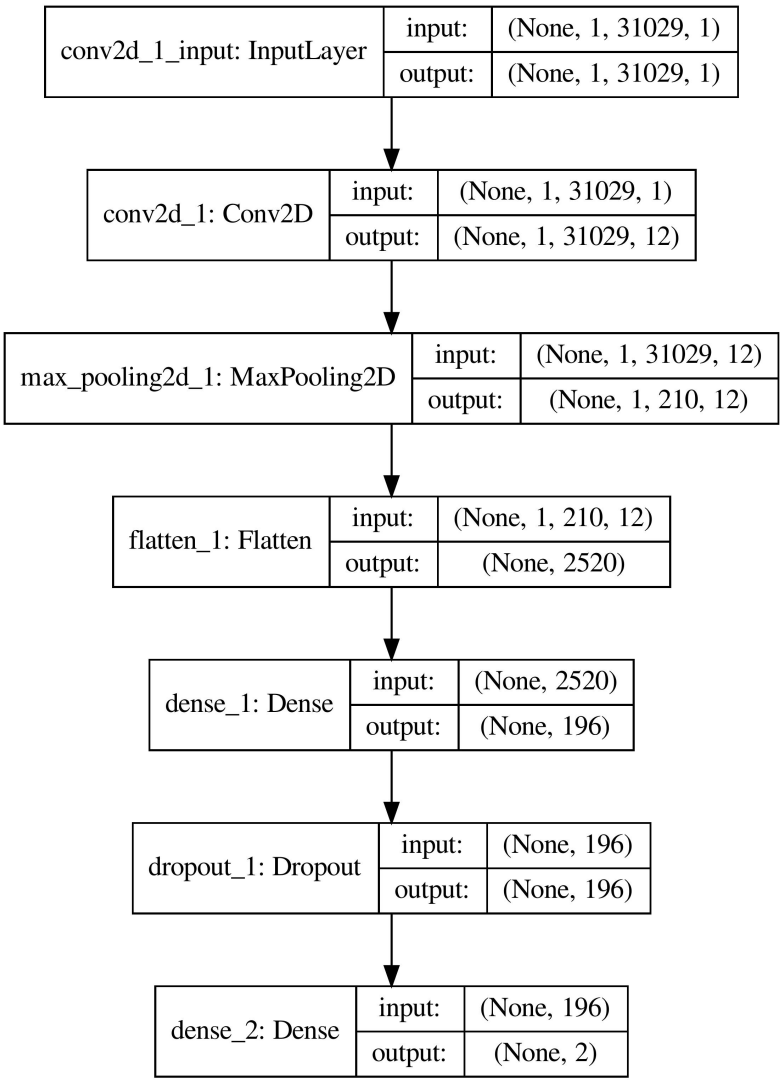


Figure 3. CNN architecture used to classify variants B.1.1.7, B.1.351, and P.1.

For further *in-silico* analysis, the resulting '21-bps' sequences are compared against 487 sequences of other coronaviruses from the NGDC repository, to check if our generated primers are specific enough to SARS-CoV-2 (Table 1 in the supplementary material). In addition, we used data from NCBI with 20,571 viral samples from other taxa, for a total of more than 584 other viruses (not considering strains and isolates). The whole procedure is summarized in Fig. 2.

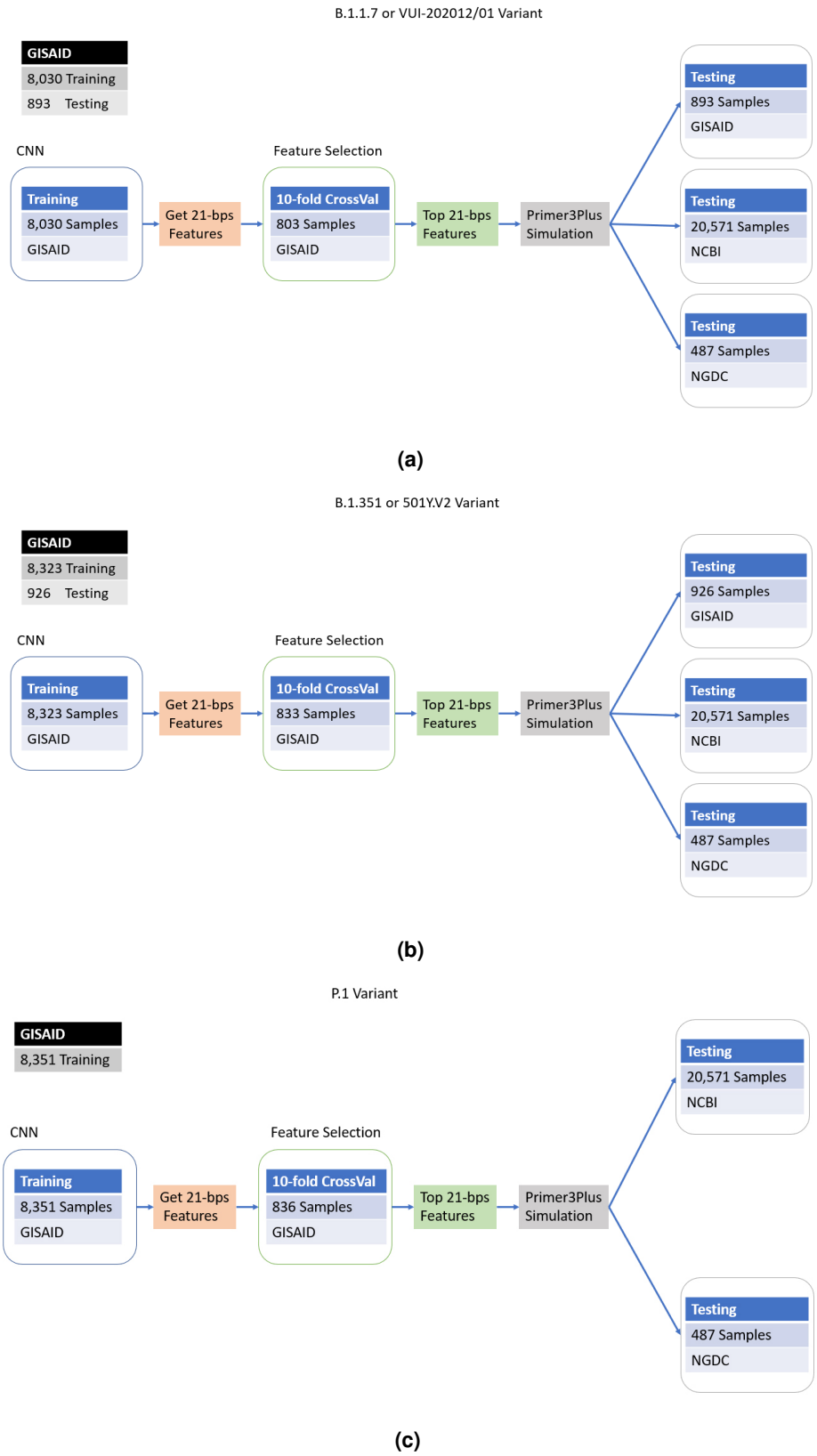


Figure 4. Scheme of the procedure carried out to find the 21-bps features to be used as candidate primers. 4a) Training and testing for the B.1.1.7 variant primer. 4b) Training and testing for the B.1.351 variant primer. 4c) Training and testing for the P.1 variant primer. There is no test set in P.1 given the number of samples (28).

References

1. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565–574 (2020).
2. Chand, M., Hopkins, S. & Dabrera, G. Investigation of novel SARS-COV-2 variant: Variant of Concern 202012/01 (2020).
3. for Disease Prevention, E. C. & Control. Rapid increase of a sars-cov-2 variant with multiple spike protein mutations observed in the united kingdom (2020).
4. Sars-cov-2 variants (2020).
5. Shu, Y. & McCauley, J. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22** (2017).
6. Arambaut, Garmstrong & Isabel. Preliminary genomic characterisation of an emergent sars-cov-2 lineage in the uk defined by a novel set of spike mutations (2020).
7. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
8. Starr, T. N. *et al.* Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310 (2020).
9. Gu, H. *et al.* Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* **369** (2020).
10. Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Mol. Cell* (2020).
11. Nicholas Davies, e. a. Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. *Available Github* (2020). Online; accessed 26 December 2020.
12. for Disease, C. & Prevention, C. Emergence of sars-cov-2 b.1.1.7 lineage — united states, december 29, 2020–january 12, 2021 (2021).
13. Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (sars-cov-2) lineage with multiple spike mutations in south africa. *medRxiv* (2020).
14. Wang, Y., Liu, M. & Gao, J. Enhanced receptor binding of sars-cov-2 through networks of hydrogen-bonding and hydrophobic interactions. *Proc. Natl. Acad. Sci.* (2020).
15. Lan, J. *et al.* Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature* **581**, 215–220 (2020).
16. Faria, N. Genomic characterisation of an emergent sars-cov-2 lineage in manaus: preliminary findings (2021).
17. Voloch, C. M. *et al.* Genomic characterization of a novel sars-cov-2 lineage from rio de janeiro, brazil. *medRxiv* (2020).
18. Alm, E. *et al.* Geographical and temporal distribution of sars-cov-2 clades in the who european region, january to june 2020. *Eurosurveillance* **25**, 2001410 (2020).
19. Elghazaly, H. *et al.* Laboratory based retrospective study to determine the start of sars-cov-2 in patients with severe acute respiratory illness in egypt at el-demerdash tertiary hospitals. *europepmc* (2020).
20. Greaney, A. J. *et al.* Complete mapping of mutations to the sars-cov-2 spike receptor-binding domain that escape antibody recognition. *Cell host & microbe* (2020).
21. Greaney, A. J. *et al.* Comprehensive mapping of mutations to the sars-cov-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. *bioRxiv* 2020–12 (2020).
22. Nonaka, C. K. V. *et al.* Genomic evidence of a sars-cov-2 reinfection case with e484k spike mutation in brazil. *Preprints* (2021).
23. Wang, Z. *et al.* mrna vaccine-elicited antibodies to sars-cov-2 and circulating variants. *bioRxiv* DOI: 10.1101/2021.01.15.426911 (2021). https://www.biorxiv.org/content/early/2021/01/19/2021.01.15.426911.full.pdf.
24. Nelson, G. *et al.* Molecular dynamic simulation reveals e484k mutation enhances spike rbd-ace2 affinity and the combination of e484k, k417n and n501y mutations (501y.v2 variant) induces conformational change greater than n501y mutant alone, potentially resulting in an esca.... *bioRxiv* DOI: 10.1101/2021.01.13.426558 (2021). https://www.biorxiv.org/content/early/2021/01/13/2021.01.13.426558.full.pdf.
25. Afzal, A. Molecular diagnostic technologies for covid-19: Limitations and challenges. *J. advanced research* (2020).
26. Organization, W. H. *et al.* Molecular assays to diagnose covid-19: summary table of available protocols (2020).

27. Giandhari, J. *et al.* Early transmission of sars-cov-2 in south africa: An epidemiological and phylogenetic report. *medRxiv* (2020).

28. Tegally, H. *et al.* Major new lineages of sars-cov-2 emerge and spread in south africa during lockdown. *medRxiv* (2020).

29. Lopez-Rincon, A. *et al.* Classification and specific primer design for accurate detection of sars-cov-2 using deep learning. *Sci. Reports* (2020).

30. Šimundić, A.-M. Measures of diagnostic accuracy: basic definitions. *Ejifcc* **19**, 203 (2009).

31. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **5**, 1315–1316 (2010).

32. Beijing Institute of Genomics, Chinese Academy of Science. China National Center for Bioinformation & National Genomics Data Center. https://bigd.big.ac.cn/ncov/?lang=en (2013). Online; accessed 27 January 2020.

33. Sherry, S. T. *et al.* dbsnp: the ncbi database of genetic variation. *Nucleic acids research* **29**, 308–311 (2001).

34. Jahn, K. *et al.* Detection of sars-cov-2 variants in switzerland by genomic analysis of wastewater samples. *medRxiv* DOI: 10.1101/2021.01.08.21249379 (2021). https://www.medrxiv.org/content/early/2021/01/09/2021.01.08.21249379.full.pdf.

35. Untergasser, A. *et al.* Primer3plus, an enhanced web interface to primer3. *Nucleic acids research* **35**, W71–W74 (2007).

36. Gupta, R. *et al.* Recurrent independent emergence and transmission of sars-cov-2 spike amino acid h69/v70 deletions. *bioArxiv* (2020).

37. Wang, R., Hozumi, Y., Yin, C. & Wei, G.-W. Mutations on covid-19 diagnostic targets. *arXiv preprint arXiv:2005.02188* (2020).

38. Kim, S. & Lee, e. a. The progression of sars coronavirus 2 (sars-cov2): Mutation in the receptor binding domain of spike gene. *Immune Netw.* **20** (2020).

39. Kidd, M. *et al.* S-variant sars-cov-2 is associated with significantly higher viral loads in samples tested by thermofisher taqpath rt-qpcr. *medRxiv* DOI: 10.1101/2020.12.24.20248834 (2020). https://www.medrxiv.org/content/early/2020/12/27/2020.12.24.20248834.full.pdf.

40. Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G. U., Schoenhuth, A. & Tonda, A. Automatic discovery of 100-mirna signature for cancer classification using ensemble feature selection. *BMC bioinformatics* **20**, 480 (2019).

41. Lopez-Rincon, A. *et al.* Machine learning-based ensemble recursive feature selection of circulating mirnas for cancer tumor classification. *Cancers* **12**, 1785 (2020).

Additional information

The authors declare no competing interests.

Author contributions statement

CAP, LMM, made the biological analysis, and primer design. ALR and AT made the programming, data collection, and experiments in silico. EC, ADK and JG made the experiment and study design. All the authors contributed to the writing.

Supplementary Material

Table 1. Organism and number of samples of other coronaviruses used to evaluate specificity in the sequences.

Organism	Number of Samples
MERS-CoV	240
HCoV-OC43	138
HCoV-229E	22
HCoV-4408	2
HCoV-NL63	58
HCoV-HKU1	17
SARS-CoV	10
Total Samples	487