Design of Specific Primer Sets for the Detection of B.1.1.7, B.1.351, P.1, B.1.617.2 and B.1.1.519 Variants of SARS-CoV-2 using Artificial Intelligence

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 continues, 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. Earlier this year, a variant labeled P.1 with 17 non-synonymous mutations was detected in Brazil. Recently the World Health Organization has raised concern for the variants B.1.617.2 mainly detected in India but now exported worldwide. It is paramount to rapidly develop specific molecular tests to uniquely identify new variants. Using a completely automated pipeline built around deep learning and evolutionary algorithms techniques, we designed primer sets specific to variants B.1.1.7, B.1.351, P.1, B.1.617.2 and B.1.1.519 respectively. We wish to raise awareness to the variant of interest, B.1.1.519, which contain several mutations of important threat to public health. Starting from sequences openly available in the GISAID repository, our pipeline was able to deliver the primer sets for each variant. In-silico tests show that the sequences in the primer sets present high accuracy and are based on 2 mutations or more. In addition, we present an analysis of key mutations for SARS-CoV-2 variants. The presented methodology can be exploited to swiftly obtain primer sets for each new variant, that can later be a part of a multiplexed approach for the initial diagnosis of COVID-19 patients.

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 globe, researchers and public health officials constantly monitor the virus for acquired mutations that may lead to a higher rate of transmissibility and/or vaccine resistance.

On December 14th, 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 adjacent to the furin cleave site of the S protein and, although there is evidence that it might play a role in

SARS-CoV-2 transmission and infection [10–12], the effect of such mutations is still under debate[13,14]. The D614G mutation has been found to enhances viral replication, viron density and infectivity[15,16]. 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[2,17]. 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[18].

In parallel, on December 18th, 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[4,7,19]. 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 receptor-binding domain (RBD) domain of S protein, L18F (C21614CT), D80A (A21801C), and D215G (A22206AG) are in the N-terminal domain, A701V (C23664T) in loop 2, and D614G (A23403G)[19]. 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[20,21].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[19]. The B.1.351 variant has now been identified in other 13 countries around the world[18].

On January 6th 2021, Japan reported a new SARS-CoV-2 variant from 4 travelers from Brazil[22]. This variant was later reported to be widely circulating in Brazil, it was named P.1, B.1.1.28.1 lineage or 20J/501Y.V3 Nexstrain clade[18,23]. 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, E484K, K417T and D614G. And although the rate of infection and mortality of the P.1 variant is still unknown, it has been identified in the Amazonia in Brazil, as in other countries[7,23,24]. It is important to notice that another B.1.1.28-derived SARS-CoV-2 variant (outside of the scope of this study) has also been circulating in other Brazil states being termed P.2 harboring similar S protein mutations to the P.1 variant including: K417T, E484K, and N501Y[23,25].

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[5,7,25–27]. A few strains of the B.1.1.7 variant has recently start acquiring the E484K mutation also present present in the in the B.1.351, P.1 and P.2 variants[28]. The E484K and K417N mutations in combination with the N501Y are have become of great global concern as they have been shown to affect neutralization by monocolonal antibodies and covalscent sera[19,29–33]. More importantly, there has also been cases of reinfection associated with the B.1.1.28-derived emerging lineages P.1[24] and P.2[34,35], 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[36,37].

On May 11th 2021, the World Health Organization announced yet another SARS-CoV-2 variant of concern from India from the B.1.617 lineage, this lineage has 3 different variants: B.1.617.1 or 20A/S.154K Nexstrain clade, B.1.617.2 or 20A/S.478K Nextstrain clade, and B.1.617.3 or 20A. The first two variants being most concerning due to the number of cases[18,38]. The B.1.617.1 harbours several mutations in the spike protein including E484Q, D614G, P681R, L452R, E154K, and Q1071H, while the B.1.617.2 contain the next mutations in the Spike protein T19R, $\triangle$156-157, L452R, T478K, D614G, P681R,R158G and D950N[18,38–41]. The L452R mutation has been shown to increase SARS-CoV-2 viral infectiousness and replication[42]. Interestingly. L452R and E484Q have been found to disrupt the interfacial interactions of the Spike RBD with neutralizing antibodies[43,44]. Furthermore, the L452R has been found to be a positive adaptive mutation driving the spread of similar variants with these mutations such like the ones in California[45]. The T478K mutation present in the B.1.617.2 variant is located at the interface of the Spike/ACE2 interaction domain[12,46], and mutations in this position has been found to confer moderate to near-complete resistance to monoclonal antibodies and convalescent serum[29,47].The variant B.1.617.1. has been found to have higher pathogenicity in hamsters[48] and increase syncytium formation in human cells[49]. Furthermore, these variants have been shown to have reduced sensitivity to Comirnaty/BNT162b2 mRNA vaccine antibodies as well as fully resistant to monoclonal antibody therapy treatment with Bamlanivimab[49,50], possibly derived by the L452R and E484Q mutations[29]. Although, more research is needed of the clinical outcome and epidemiology of the B.1.617 variants, it is believed to be playing a major role in India's recent increase in number of SARS-CoV-2 cases and deaths[41].

In addition, we wanted to raise awareness of a SARS-CoV-2 variant of interest mostly present in Mexico and USA, the variant B.1.1.519. This variant harbours the mutations in the spike protein T478K, P681H, D614G and T732A[5,51] and some strains harbour the N501Y and E484K[52]. Due to the already described potential threat of the mutations carried by this variant, it is of importance for health authorities in Mexico and the USA to continue to monitor the progression of this variant as it

could pose a higher risk for public health.

Given the rapid spread of these new variants, and their concerns to the global health, it has become of utmost importance for countries to be able to rapidly identify and detect them. 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)[53,54]. 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 notice the rise in Covid-19 positive cases in South Easter England. B.1.1.7 was detected through the increase in S-gene target failure (negative results) from the otherwise positive target genes (N, ORF1ab) in the three target gene assay in their RT-PCR diagnostic tests, and random whole genome sequencing of some of this positive samples[2,44]. In the case of South Africa, they were able to identify the rise of B.1.351 SARS-CoV-2 variant through the increase in positive Covid-19 cases and deaths in at least 3 of their provinces and by random whole genome sequencing of positive Covid-19 samples, through their national diagnosis and surveillance system[19,55,56].

In a previous work[57], 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. Then, we reduced the necessary time by half using evolutionary algortihms[58]. When compared to other primers sets suggested by GISAID, our approach proved to deliver competitive accuracy and specificity. Our results for SARS-CoV-2 detection, 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 complete 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, P.1, B.1.6173.*, and B.1.1.519 in under 10 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%. 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. While the best result corresponds to a set of 16 '21-bps' features, using only one is enough to obtain over 99% accuracy.

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 (T16176C). Using Primer3Plus we compute a primer set for each of the 10 features, using sequence EPI_ISL_601443 as the reference sequence. Only the two '21-bps' features that include mutation T16176C are suitable for a forward primer. The two features are **ACCTCAAGGTATTGGGAACCT** and **CACCTCAAGGTATTGGGAACC**: 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 **CATCACAACCTGGAGCATTG**.

For further analysis, we check the presence of the signature mutations of B.1.1.7; T1001I, A1708D, I2230T, SGF 3675-3677 deletion, HV 69-70 deletion, Y144 deletion, N501Y, A570D, P681H, T716I, S982A, D1118H, Q27stop, R52I, Y73C, D3L and S235F of variant B.1.1.7[59]. 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 **TGATATCCTTGCACGTCTTGA** in spike gene S982A as the only possible forward primer candidate, this sequence can be used for multiplex testing (Table 1).

Table 1. Frequency of appearance in B.1.1.7, B.1.351, P.1, B.1.617.*, B.1.617.2, Mexican, B.1.1.519 and other variants sequences for the main mutations in variant B.1.1.7. T16176C is the mutation found by deep learning. S982 sequence, can be used as a forward primer. The rows in bold are mutations that can be used together to make primer sets.

gene	Nucleotide	Amino acid	Used Sequence	B.1.1.7	B.1.351	P.1	B.1.617.*	B.1.617.2	Mexico*	B.1.1.519	Other
				7,257 seqs.	8,420 seqs.	6,505 seqs.	1,908 seqs.	697 seqs.	5,203 seqs.	3,035 seqs.	6,808 seqs.
ORF1ab	C3267T	T1001I	CAGACAACTATTATTCAAACA	98.97%	0.01%	0.02%	0.21%	0.00%	1.92%	2.14%	0.01%
ORF1ab	C5388A	A1708D	GGTGAAGCTGATAACTTTTGT	99.63%	0.02%	0.00%	0.00%	0.00%	1.77%	2.01%	0.00%
ORF1ab	T6954C	I2230T	ATAAATATTACAATTTGGTTT	98.72%	0.00%	0.02%	0.05%	0.00%	1.75%	2.01%	0.01%
ORF1ab	11288-11296 del	3675-3677 del	GTTTGAAGCTAAAAGACTGTG	99.41%	95.31%	98.37%	0.05%	0.00%	2.44%	2.37%	0.00%
Spike	**21765-21770 del**	**HV 69-70 del**	**TCCATGCTATCTCTGGGACCA**	**97.15%**	**0.01%**	**0.14%**	**0.37%**	**0.57%**	**2.15%**	**2.34%**	**0.57%**
Spike	**21991-21993 del**	**Y144 del**	**TTTTTGGGTGTTTACCAC**	**95.38%**	**0.84%**	**0.15%**	**0.31%**	**0.43%**	**2.19%**	**2.41%**	**0.19%**
Spike	A23063T	N501Y	CCAACCCACTTATGGTGTTGG	99.67%	98.87%	99.68%	0.21%	0.00%	2.25%	2.21%	0.04%
Spike	C23271A	A570D	AGAGACATTGATGACACTACT	99.41%	0.00%	0.09%	0.00%	0.00%	1.79%	2.04%	0.01%
Spike	C23604A	P681H	ACTAATTCTCATCGGCGGGCA	98.77%	0.07%	0.55%	0.05%	0.00%	61.27%	64.32%	0.01%
Spike	C23709T	T716I	GCCATACCCATAAATTTTACT	99.66%	0.23%	0.14%	0.00%	0.00%	1.94%	2.21%	0.01%
Spike	**T24506G**	**+S982A**	**TGATATCCTTGCACGTCTTGA**	**99.55%**	**0.00%**	**0.18%**	**0.00%**	**0.00%**	**1.83%**	**2.11%**	**0.00%**
Spike	**G24914C**	**D1118H**	**CATTACTACACACAACACATT**	**99.26%**	**0.00%**	**0.15%**	**0.00%**	**0.00%**	**1.79%**	**2.04%**	**0.00%**
Orf8	C27972T	Q27stop	GTCATGTACTTAACATCAACC	99.68%	0.48%	0.38%	0.00%	0.00%	1.98%	2.24%	0.04%
Orf9	G28048T	R52I	GTAGGAGCTATAAAATCAGCA	99.10%	0.08%	0.42%	0.00%	0.00%	1.81%	2.08%	0.00%
Orf10	A28111G	Y73C	CCCATTCAGTGCATCGATATC	98.88%	0.07%	0.32%	0.00%	0.00%	1.79%	2.04%	0.00%
N	28280 GAT->CTA	D3L	TAAAATGTCTCTAAATGGACC	2.51%	0.00%	0.00%	0.00%	0.00%	0.12%	0.16%	0.00%
N	C28977T	S235F	AGCAAAATGTTTGGTAAAGGC	99.28%	0.06%	0.03%	0.10%	0.29%	1.79%	2.08%	0.03%
ORF1ab	**C15279T**	**Synonimous**	**AAAACCCTCATCTTATGGGTT**	**99.70%**	**0.01%**	**0.03%**	**0.05%**	**0.00%**	**2.02%**	**2.27%**	**0.03%**
ORF1ab	***T16176C**	**Synonimous**	**ACCTCAAGGTATTGGGAACCT**	**99.86%**	**0.02%**	**0.06%**	**0.10%**	**0.00%**	**1.79%**	**2.04%**	**0.00%**

Variant B.1.351

Again, the first step is to run a CNN classifier on the data. This yields a classification accuracy of 99.88%. Next, we translate the CNN weights into 6,069 features, each representing a 21-bps sequence. we run the feature selection algorithm 10 times, 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. The resulting sequence is **CTCCAGGGCAAACTGGAAATA**, with a reverse primer **TGCTACCGGCCTGATAGATT**.

We check whether 21-bps sequences containing the signature mutations T265I, K1655N, K3353R, SGF 3675-3677 deletion, L18F, D80A, D215G, R246I, K417N, E484K, N501Y, A701V, 242-244 del, Q57H, S171L, P71L, and T205I of variant B.1.351[59] could function as a forward primer. From the Primer3Plus results using EPI_ISL_678597, mutations D215G (**TTAGTGCGTGGTCTCCCTCAG**) and Q57H (**CTGTTTTTCATAGCGCTTCCA**) can be used as forwards primers (Table 2).

Table 2. Frequency of appearance in B.1.1.7, B.1.351, P.1, B.1.617.*, B.1.617.2, Mexican, B.1.1.519 and other variants sequences for the main mutations in B.1.351. K417N is the mutation found by deep learning. D215G and Q57H sequences can be used as forward primers or as part of a multiplex array. The rows in bold are mutations that can be used together to make primer sets.

gene	nucleotide	amino acid	Used Sequence	B.1.1.7	B.1.351	P.1	B.1.617.*	B.1.617.2	Mexico*	B.1.1.519	Other
				7,257 seqs.	8,420 seqs.	6,505 seqs.	1,908 seqs.	697 seqs.	5,203 seqs.	3,035 seqs.	6,808 seqs.
ORF1ab	C1059T	T265I	AAATTTGACATCTTCAATGGG	0.21%	99.39%	0.31%	0.00%	0.00%	8.42%	8.60%	8.67%
ORF1ab	G5230T	K1655N	ACACTAAAAATTGGAAATACC	0.04%	98.17%	0.28%	0.05%	0.00%	0.25%	0.20%	0.24%
ORF1ab	A10323G	K3353R	CTTAAGCTTAGGGTTGATACA	1.86%	99.61%	1.08%	0.42%	0.00%	2.38%	2.14%	0.62%
ORF1ab	11288-11296 del	3675-3677 del	GTTTGAAGCTAAAAGACTGTG	99.41%	95.31%	98.37%	0.05%	0.00%	2.44%	2.37%	0.00%
Spike	C21614T	L18F	GTGTGTTAATTTTACAACCAG	0.17%	52.57%	0.22%	0.00%	0.00%	0.56%	0.66%	0.19%
Spike	**A21801C**	**D80A**	**AAGAGGTTTGCTAACCCTGTC**	**0.00%**	**99.42%**	**0.00%**	**0.05%**	**0.00%**	**0.08%**	**0.00%**	**0.04%**
Spike	**A22206G**	**+D215G**	**TTAGTGCGTGGTCTCCCTCAG**	**0.04%**	**97.01%**	**0.02%**	**0.00%**	**0.00%**	**0.10%**	**0.00%**	**0.09%**
Spike	G22299T	R246I	GCTTTACATATAAGTTATTTG	0.00%	0.06%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Spike	**G22813T**	***K417N**	**AAACTGGAAATATTGCTGATT**	**0.01%**	**97.30%**	**0.00%**	**0.00%**	**0.00%**	**0.10%**	**0.03%**	**0.04%**
Spike	G23012A	E484K	TAATGGTGTTAAAGGTTTTAA	0.30%	98.79%	98.65%	0.00%	0.00%	1.04%	1.12%	0.07%
Spike	A23063T	N501Y	CCAACCCACTTATGGTGTTGG	99.67%	98.87%	99.68%	0.21%	0.00%	2.25%	2.21%	0.04%
Spike	C23664T	A701V	TCACTTGGTGTAGAAAATTCA	0.63%	99.67%	0.06%	0.05%	0.00%	0.35%	0.20%	0.12%
Spike	22289-22298 del	242-244del	AACTTTACATAGAAGTTATTT	0.04%	90.96%	0.02%	0.05%	0.14%	0.08%	0.00%	0.00%
ORF3a	G25563T	+Q57H	CTGTTTTTCATAGCGCTTCCA	0.15%	98.99%	2.37%	0.05%	0.00%	8.71%	8.96%	21.84%
ORF3a	C25904T	S171L	GTCATTACTTTAGGTGATGGC	0.63%	97.19%	0.29%	0.00%	0.00%	0.50%	0.36%	0.19%
E	C26456T	P71L	TCTAGAGTTCTTGATCTTCTG	0.12%	99.57%	0.03%	0.00%	0.00%	0.21%	0.07%	0.06%
N	C28887T	T205I	AGTAGGGGAATTTCTCCTGCT	0.01%	99.73%	0.02%	0.00%	0.00%	4.42%	4.51%	0.47%

Variant P.1

The CNN classifier on the data yields a classification accuracy of 100%, in the 28 samples. Next, we translate the CNN weights into 727 features, each representing a 21-bps sequence. We run the feature selection algorithm 10 times, which results in only one meaningful feature achieving 100% accuracy. The 10 runs point to the same sequence, around the synonymous mutation T733C **ACTGATCCTTATGAAGACTTT**. 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 **GGCACTGATCCTTATGAAGACT**, of size 22 bps, with reverse primer **TTCGGACAAAGTGCATGAAG**.

We check whether the sequences containing signature mutations S1188L, K1795Q, E5665D, SGF 3675-3677 deletion, T20N/L18F, P26S, D138Y, R190S, K417T, EA84K, N501Y, H655Y, T1027I, E92K, ins28269-28273 AACA, and P80R of variant P.1[23] 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 using EPI_ISL_792683 , 3 of the generated sequences using mutations can be forward primers candidates; K417T (**CAAACTGGAACGATTGCTGAT**), H665Y (**AGGGGCTGAATATGTCAACAA**) and P80R (**AATAGCAGTCGAGATGACCAA**) (Table 3).

Table 3. Frequency of appearance in B.1.1.7, B.1.351, P.1, B.1.617.*, B.1.617.2, Mexican, B.1.1.519 and other variants sequences for the main mutations in P.1. T733C is the mutation found by deep learning. K417T, H665Y and P80R sequences can be used in a multiplex array as forward primers. The rows in bold are mutations that can be used together to make primer sets.

gene	nucleotide	amino acid	Used Sequence	B.1.1.7	B.1.351	P.1	B.1.617.*	B.1.617.2	Mexico*	B.1.1.519	Other
				7,257 seqs.	8,420 seqs.	6,505 seqs.	1,908 seqs.	697 seqs.	5,203 seqs.	3,035 seqs.	6,808 seqs.
ORF1ab	C3828T	S1188L	AAACTTGTTTTAAGCTTTTTG	0.26%	0.04%	98.95%	0.10%	0.29%	0.56%	0.36%	0.04%
ORF1ab	A5648C	K1795Q	ACAAGCTACACAATATCTAGT	0.00%	0.00%	99.63%	0.00%	0.00%	0.44%	0.23%	0.00%
ORF1ab	G17259T	E5665D	CTCGTGTAGATTGTTTTGATA	0.04%	0.00%	99.52%	0.00%	0.00%	0.50%	0.23%	0.13%
ORF1ab	11288-11296 del	3675-3677 del	GTTTGAAGCTAAAAGACTGTG	99.41%	95.31%	98.37%	0.05%	0.00%	2.44%	2.37%	0.00%
Spike	**C21614T/C21621A**	**T20N/L18F**	**AATTTTACAAACAGAACTCAA**	**0.00%**	**0.00%**	**99.52%**	**0.00%**	**0.00%**	**0.38%**	**0.16%**	**0.00%**
Spike	C21638T	P26S	TCAATTACCCTCTGCATACAC	0.11%	0.02%	99.51%	0.00%	0.00%	0.90%	0.59%	0.09%
Spike	**G21974T**	**D138Y**	**ATTTTGTAATTATCCATTTTT**	**0.19%**	**0.14%**	**98.65%**	**0.37%**	**0.00%**	**1.36%**	**0.92%**	**0.10%**
Spike	G22132T	R190S	AAAATCTTAGTGAATTTGTGT	0.00%	0.00%	98.54%	0.00%	0.00%	0.44%	0.23%	0.00%
Spike	A22812C	K417T	CAAACTGGAACGATTGCTGAT	0.00%	0.00%	95.07%	0.00%	0.00%	0.38%	0.16%	0.00%
Spike	G23012A	EA84K	TAATGGTGTTAAAGGTTTTAA	0.30%	98.79%	98.65%	0.00%	0.00%	1.04%	1.12%	0.07%
Spike	A23063T	N501Y	CCAACCCACTTATGGTGTTGG	99.67%	98.87%	99.68%	0.21%	0.00%	2.25%	2.21%	0.04%
Spike	C23525T	H655Y	AGGGGCTGAATATGTCAACAA	0.56%	0.08%	99.65%	0.21%	0.00%	0.50%	0.30%	0.13%
Spike	C24642T	T1027I	CTTGCTGCTATTAAAATGTCA	0.01%	0.01%	99.00%	0.00%	0.00%	0.52%	0.33%	0.09%
Orf8	G28167A	E92K	TAATTGCCAGAAACCTAAATT	0.01%	0.00%	99.00%	0.05%	0.14%	0.56%	0.43%	0.07%
Orf8	**ins28269-28273 AACA**	**-**	**AACGAACAAACAAACTAAAAT**	**0.00%**	**0.00%**	**88.59%**	**0.00%**	**0.00%**	**0.37%**	**0.16%**	**0.00%**
N	**C28512G**	**P80R**	**AATAGCAGTCGAGATGACCAA**	**0.01%**	**0.01%**	**99.54%**	**0.05%**	**0.14%**	**0.38%**	**0.16%**	**0.00%**
ORF1ab	*T733C	Synonimous	CTTATGAAGACTTTCAAGAAA	0.01%	0.00%	99.66%	0.00%	0.00%	0.38%	0.16%	0.00%

Variant B.1.617.2

Using the sequences of variant B.1.617.* downloaded from the GISAID repository we ran an evolutionary algorithm 10 times to find the most important sequences to separate the B.1.617.* from the rest of the sequences. From the 10 runs, 6 point to the mutation R203M, 2 to the mutation V82A, and 2 to the mutation L452R. From this we can use the sequence **GGCAGCAGTATGGGAACTTCT** as a forward primer using mutation R203M in the spike gene with a 98.79% accuracy. Then, we considered for an analysis with Primer3Plus in sequence EPI_ISL_1315322 (B.1.617), the characteristic mutations T1567I, T3646A, M5753I, K6711R, E154K, L452R, E484Q, P681R, S26L, V82A, R203M and D377 of B.1.617.1[60].The mutations T3646A (**TTCTCTTGCCGCTGTAGCTTA**), V82A (**GCCAGATCAGCTTCACCTAAA**) and R203M (**GGCAGCAGTATGGGAACTTCT**) can be used as primers in a multiplex approach (Table 4).

Table 4. Frequency of appearance in B.1.1.7, B.1.351, P.1, B.1.617.*, B.1.617.2, Mexican, B.1.1.519 and other variants sequences for the main mutations in B.1.617.*. R203M is the mutation found using EAs. Sequences of mutations T3646A and V82A can be used as forward primers in a multiplex approach. The rows in bold are mutations that can be used together to make primer sets.

gene	nucleotide	amino acid	Used Sequence	B.1.1.7	B.1.351	P.1	B.1.617.*	B.1.617.2	Mexico*	B.1.1.519	Other
				7,257 seqs.	8,420 seqs.	6,505 seqs.	1,908 seqs.	697 seqs.	5,203 seqs.	3,035 seqs.	6,808 seqs.
ORF1ab	C4965T	T1567I	GAAGTGAGGATTATTAAGGTG	0.21%	0.01%	0.11%	55.40%	0.00%	0.10%	0.13%	0.07%
ORF1ab	A11201G	+T3646A	TTCTCTTGCCGCTGTAGCTTA	0.00%	0.01%	0.00%	65.93%	28.69%	0.04%	0.07%	0.00%
ORF1ab	G17523T	M5753I	GTAGACTTATTAAAACTATAG	0.03%	0.00%	0.00%	55.77%	0.00%	0.10%	0.13%	0.01%
ORF1ab	A20396G	K6711R	AAACGTTTTAGGGAATCACCT	0.01%	0.01%	0.00%	24.42%	10.76%	0.00%	0.00%	0.00%
Spike	G22022A	E154K	AAGTTGGATGAAAAGTGAGTT	0.04%	0.00%	0.00%	43.76%	0.00%	0.02%	0.03%	0.00%
Spike	T22917G	L452R	TATAATTACCGGTATAGATTG	0.11%	0.00%	0.03%	99.63%	99.71%	4.02%	4.09%	0.00%
Spike	G23012C	E484Q	TAATGGTGTTCAAGGTTTTAA	0.08%	0.00%	0.00%	58.81%	0.14%	0.08%	0.10%	0.06%
Spike	C23604G	P681R	ACTAATTCTCGTCGGCGGGCA	0.21%	0.00%	0.00%	99.42%	99.86%	0.54%	0.69%	0.03%
ORF3a	C25469T	S26L	GCTACTCCTTTAGATTTTGTT	0.25%	0.10%	0.08%	96.07%	98.57%	0.21%	0.23%	0.03%
ORF7a	T27638C	+V82A	GCCAGATCAGCTTCACCTAAA	0.00%	0.00%	0.00%	98.58%	98.42%	0.02%	0.03%	0.00%
N	**G28881T**	**+*R203M**	**GGCAGCAGTATGGGAACTTCT**	**0.01%**	**0.00%**	**0.00%**	**98.85%**	**99.14%**	**0.12%**	**0.16%**	**0.04%**
N	**G29402T**	**D377Y**	**GAAGGCTTATGAAACTCAAGC**	**0.15%**	**0.17%**	**0.72%**	**97.43%**	**98.57%**	**0.98%**	**1.15%**	**0.38%**
N	**G29742T**	**Synonimous**	**TTCACCGAGGCCACTCGGAGT**	**0.14%**	**0.04%**	**0.37%**	**91.19%**	**89.53%**	**0.40%**	**0.43%**	**1.10%**

Next, we run the experiment focusing only in B.1.617.2. The results 4 out of 10 point to mutation D63G, 4 to mutation I82T and 2 to mutation T120I. The forward primer for mutation I82T is (**CTACCGCAATGGCTTGTCTT**) and for mutation D63G is (**ATGGCAAGGAAGGCCTTAAA**) using sequence EPI_ISL_1337507. Then, we make an analysis of the characteristic mutations of lineage B.1.617.2; T19R, L452R, T478K, P681R, D950N, S26L, I82T, V82A T120I, D63G, R203M and D377Y[61]. Where sequences of mutations T478K (**GCCGGTAGCAAACCTTGTAAT**), V82A (**GCCAGATCAGCTTCACCTAAA**) and R203M (**GGCAGCAGTATGGGAACTTCT**) can be used as forward primers in multiplex approach (Table 5).

Table 5. Frequency of appearance in B.1.1.7, B.1.351, P.1, B.1.617.*, B.1.617.2, Mexican, B.1.1.519 and other variants sequences for the main mutations in B.1.617.2. I82T and D63G are the mutations found using EAs. The sequences of mutations T478K, V82A and R203M can be used as forward primers in a multiplex approach. The rows in bold are mutations that can be used together in a primer set.

gene	nucleotide	amino acid	Used Sequence	B.1.1.7	B.1.351	P.1	B.1.617.*	B.1.617.2	Mexico*	B.1.1.519	Other
				7,257 seqs.	8,420 seqs.	6,505 seqs.	1,908 seqs.	697 seqs.	5,203 seqs.	3,035 seqs.	6,808 seqs.
Spike	C21618G	T19R	GTTAATCTTAGAACCAGAACT	0.00%	0.00%	0.00%	42.40%	96.84%	0.00%	0.00%	0.00%
Spike	T22917G	L452R	TATAATTACCGGTATAGATTG	0.11%	0.00%	0.03%	99.63%	99.71%	4.02%	4.09%	0.00%
Spike	C22995A	+T478K	GCCGGTAGCAAACCTTGTAAT	0.01%	0.02%	0.00%	40.20%	99.86%	57.20%	60.36%	0.03%
Spike	C23604G	P681R	ACTAATTCTCGTCGGCGGGCA	0.21%	0.00%	0.00%	99.42%	99.86%	0.54%	0.69%	0.03%
Spike	G24410A	D950N	AAAACTTCAAAATGTGGTCAA	0.01%	0.00%	0.00%	40.30%	94.69%	0.00%	0.00%	0.00%
ORF3a	C25469T	S26L	GCTACTCCTTTAGATTTTGTT	0.25%	0.10%	0.08%	96.07%	98.57%	0.21%	0.23%	0.03%
M	**T26767C**	***I82T**	**GGAATTGCTACCGCAATGGCT**	**0.19%**	**0.04%**	**0.00%**	**40.20%**	**99.57%**	**0.00%**	**0.00%**	**0.00%**
ORF7a	**T27638C**	**+V82A**	**GCCAGATCAGCTTCACCTAAA**	**0.00%**	**0.00%**	**0.00%**	**98.58%**	**98.42%**	**0.02%**	**0.03%**	**0.00%**
ORF7a	C27752T	T120I	AAAAGAAAGATAGAATGATTG	0.17%	0.00%	0.02%	40.36%	98.57%	0.08%	0.10%	0.04%
N	A28461G	*D63G	GGCAAGGAAGGCCTTAAATTC	0.00%	0.00%	0.00%	38.94%	96.56%	0.00%	0.00%	0.00%
N	G28881T	+R203M	GGCAGCAGTATGGGAACTTCT	0.01%	0.00%	0.00%	98.85%	99.14%	0.12%	0.16%	0.04%
N	G29402T	D377Y	GAAGAAGGCTTATGAAACTCA	0.15%	0.17%	0.72%	97.33%	98.42%	0.98%	1.15%	0.38%

Variant B.1.1.519

The first experiment separated the Mexican sequences from the rest, Out of 10 runs, 5 point to mutation P65S in NSP15, 3 to T732A in the spike gene and 1 to P681H and 1 to P141S. From this, we create a forward primer using mutation P65S **AATTTGGGTGTGGACATTGC** using sequence EPI_ISL_1040535. For the experiment, trying to separate B.1.1.519 sequences the EAs results 8 out of 10 point to mutation P681H and 2 to mutation C29197T. Using sequence EPI_ISL_1040535, we find the reverse primer (**CCCGCCGATGAGAATTAGT**) using mutation P618H. From our results and the information available in[62] we considered for analysis the mutations P141S, P65S, T478K, E484K, D614G, P681H and T732A for both experiments. In this case, only sequence (**GCCGGTAGCAAACCTTGTAAT**) with T478K can be considered as a forward primer (Table 6).

Table 6. Frequency of appearance in B.1.1.7, B.1.351, P.1, B.1.617.*, B.1.617.2, Mexican, B.1.1.519 and other variants sequences for the main mutations in B.1.1.519. In all the sequences available in Mexico, the solution was mutation P65S for differentiating from other variants. Specifically for variant B.1.1.519, the most important mutation is P681H. the sequence displayed for T478K can be used as a forward primer in a multiplex array. The bold rows can be used to make primers together.

gene	nucleotide	amino acid	Used Sequence	B.1.1.7	B.1.351	P.1	B.1.617.*	B.1.617.2	Mexico*	B.1.1.519	Other
				7,257 seqs.	8,420 seqs.	6,505 seqs.	1,908 seqs.	697 seqs.	5,203 seqs.	3,035 seqs.	6,808 seqs.
NSP3	C3140T	P141S	CCAAGGTAAATCTTTGGAATT	0.18%	0.00%	0.02%	0.05%	0.00%	58.47%	61.19%	0.04%
NSP15	T19839C	*P65S	AAATACTCAACAATTTGGGTG	0.11%	0.00%	0.00%	0.00%	0.00%	66.94%	68.50%	1.16%
Spike	**C22995A**	**+T478K**	**GCCGGTAGCAAACCTTGTAAT**	**0.01%**	**0.02%**	**0.00%**	**40.20%**	**99.86%**	**57.20%**	**60.36%**	**0.03%**
Spike	G23012A	E484K	TAATGGTGTTAAAGGTTTTAA	0.30%	98.79%	98.65%	0.00%	0.00%	1.04%	1.12%	0.07%
Spike	A23403G	D614G	CTTTATCAGGGTGTTAACTGC	99.66%	99.52%	99.42%	99.27%	99.71%	99.58%	99.60%	78.92%
Spike	**C23604A**	***P681H**	**ACTAATTCTCATCGGCGGGCA**	**98.77%**	**0.07%**	**0.55%**	**0.05%**	**0.00%**	**61.27%**	**64.32%**	**0.01%**
Spike	A23756G	T732A	AGTGTCTATGGCCAAGACATC	0.00%	0.01%	0.00%	0.00%	0.00%	65.12%	66.89%	0.13%

Discussion

A single-nucleotide mutation may not be enough to work as a specific primer for detecting SARS-CoV-2 variants. Thus, from the analysis of the characteristics mutations for each variant and the results of our pipeline, we created a list of primer sets based upon 2 or more mutation for each variant. For variant B.1.1.7 we created 3 primer sets, for variants B.1.351, P.1 and B.1.617.2 we created 2 options for each one, and for B.1.1.519 and B.1.617.* (B.1.617.1, B.1.617.2 and B.1617.3) we created 1 (Table 7). Using Primer3Plus for in-silico simulations, we tried to maintain acceptable levels of temperature and an 18-22 bps size. Nevertheless, it was not always possible e.g. for P.1 variant the forward primers contain sequences size 25 bps. Another thing to consider is the product size, depending of which primer is going to be used the size will vary, as each primer falls into at least one mutation. All of the selected sequences present accuracy above 95% with 2 exceptions P.1-2 forward primer with 88.99% and the generated primers for B.1.1.519 with 60.40% for the forward primer, and 64.32% for the reverse primer. This primer set in particular has problems as mutations T478K is shared with B.1.1.7 and B.1.617.2 variant and P681H with P.1 variant.

Table 7. Primer sets created considering at least 2 mutations for B.1.1.7, B.1.351, P.1, B.1.617.*, B.1.617.2 and B.1.1.519 with frequency of appearance.

Mutation	Sequence	B.1.1.7 7,257 seqs.	B.1.351 8,420 seqs.	P.1 6,505 seqs.	B.1.617.* 1,908 seqs.	B.1.617.2 697 seqs.	Mexico* 5,203 seqs.	B.1.1.519 3,035 seqs.	Other 6,808 seqs.	Size (bps)	Tm (C)
B.1.1.7-1 (244 bp)											
HV 69-70 deletion	F-CATGCTATCTCTGGGACCAAT	97.34%	0.01%	0.14%	0.37%	0.57%	2.15%	2.34%	0.57%	21	59.0
Y144 deletion	R-TGTTGTTTTTGTGGTAAACACC	95.04%	0.84%	0.15%	0.31%	0.43%	2.21%	2.44%	0.19%	22	57.5
B.1.1.7-2 (437 bp)											
S982A	F-TGATATCCTTGCACGTCTTGA	99.55%	0.00%	0.18%	0.00%	0.00%	1.83%	2.11%	0.00%	21	59.3
D1118H	R-CAGACACAAATGTGTTGTGTGTAGT	99.20%	0.00%	0.15%	0.00%	0.00%	1.79%	2.04%	0.00%	25	59.1
B.1.1.7-3 (936 bp)											
C15279T	F-TGATGTAGAAAACCCTCATCTTATG	99.59%	0.01%	0.03%	0.05%	0.00%	2.02%	2.27%	0.03%	25	58.6
T16176C	R-TCAGGTTCCCAATACCTTGAG	99.93%	0.02%	0.06%	0.10%	0.00%	1.79%	2.04%	0.00%	21	59.0
B.1.351-1 (425 bp)											
D80A*	F-AGAGGTTTGCTAACCCTGTCC	0.00%	99.36%	0.00%	0.05%	0.00%	0.08%	0.00%	0.04%	21	59.6
D215G*	R-CTGAGGGAGACCACGCACTAA	0.04%	97.01%	0.02%	0.00%	0.00%	0.10%	0.00%	0.09%	21	62.7
B.1.351-2 (619 bp)											
D215G*	F-TTAGTGCGTGGTCTCCCTCAG	0.04%	97.01%	0.02%	0.00%	0.00%	0.10%	0.00%	0.09%	21	62.7
K417N	R-AATCAGCAATATTTCCAGTTTGC	0.01%	97.23%	0.00%	0.00%	0.00%	0.10%	0.03%	0.04%	23	59.5
P.1-1 (375 bp)											
T20N/L18F	F-TTTTACAAACAGAACTCAATTACCC	0.00%	0.00%	99.43%	0.00%	0.00%	0.38%	0.16%	0.00%	25	57.9
D138Y	R-CCCAAAAATGGATAATTACAAAA	0.19%	0.14%	98.43%	0.00%	0.00%	1.35%	0.89%	0.10%	23	57.2
P.1-2 (266 bp)											
ins28269-28273 AACA	F-CGAACAAACAAACTAAAATGTCTGA	0.00%	0.00%	88.59%	0.00%	0.00%	0.37%	0.16%	0.00%	25	59.6
P80R	R-TTGGTCATCTCGACTGCTATT	0.01%	0.01%	99.54%	0.05%	0.14%	0.38%	0.16%	0.00%	21	57.4
B.1.617.* (546 bp)											
R203M	F-GGCAGCAGTATGGGAACTTCT	0.01%	0.00%	0.00%	98.85%	99.14%	0.12%	0.16%	0.04%	21	60.6
D377Y	R-GGCTTGAGTTTCATAAGCCTTC	0.15%	0.17%	0.72%	97.43%	98.57%	0.98%	1.15%	0.38%	22	59.4
B.1.617.2-1 (879 bp)											
I82T	F-CTACCGCAATGGCTTGTCTT	0.19%	0.04%	0.00%	40.20%	99.57%	0.00%	0.00%	0.00%	20	60.3
V82A	R-TGAAGCTGATCTGGCACGTA	0.00%	0.00%	0.00%	98.64%	98.71%	0.02%	0.03%	0.00%	20	60.6
B.1.617.2-2 (434 bp)											
D63G	F-ATGGCAAGGAAGGCCTTAAA	0.00%	0.00%	0.00%	38.94%	96.56%	0.00%	0.00%	0.00%	20	60.9
R203M	R-CATACTGCTGCCTGGAGTTGA	0.01%	0.00%	0.00%	98.58%	99.00%	0.12%	0.16%	0.04%	21	61.4
B.1.1.519 (626 bp)											
T478K	F-CGGTAGCAAACCTTGTAATGG	0.01%	0.02%	0.00%	40.20%	99.86%	57.20%	60.40%	0.03%	21	59.5
P681H	R-CCCGCCGATGAGAATTAGT	98.95%	0.07%	0.55%	0.05%	0.00%	61.27%	64.32%	0.01%	19	60.0

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[53]. The arising prevalence of new SARS-CoV-2 variants such as B.1.1.7, B.1.351, P.1 and B.1.617.2 have 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 △69-70 deletion in the SARS-CoV-2 spike protein, present in B.1.1.7[2]. This △69-70 deletion, which affects its N-terminal domain, has been occurring in several different SARS-CoV-2 variants around the world[7,63] 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[53,54,64]. This is consistent with other existing primer designs like CoV2R-3 in the S-gene[65], 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.[66].

Given the concern for the increase in prevalence of the new variants SARS-CoV2 B.1.1.7, B.1.351, P.1 and B.1.617.2 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, P.1 and B.1.617.2 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, P.1 or B.1.617.2 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 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.

Finally, due to the incredible rapid increase and appearance of emerging variants of concern and of interest, it is important to track important mutations that can have an adverse effect in the pandemic. Thus, we use the B.1.1.519 variant of interest mainly present in Mexico and US, given its troublesome mutations. In our analysis we found that at least 60% of the sequences in GISAID from Mexico are from this variant, this estimate is conservative given that in April 2021 it was reported that 80% of the present COVID-19 cases in Mexico are B.1.1.519[67]. Furthermore, it is important to take into account that the data we have access to, are estimated numbers from the controversial epidemiological model used in Mexico termed "sentinel", the numbers provided by this this model must be multiplied by a factor of 8 in order to get close to the correct numbers in the country territory[68,69]. Therefore, it is hard to estimate the actual number of actual COVID-19 cases in Mexico, and have an actual picture of the ongoing problem with the variants. Nevertheless, with the few access to information available at the moment, our analysis of the mutation present on the B.1.1.519 variant in Mexico, show that the virus and mutations defining this lineage are still evolving however, most sequences have present the T478K and P681H mutations which have been found to have resistance to monoclonal antibodies and convalescent serum[29,47] and potentially play a role in increased SARS-CoV-2 transmission and infection [10–12].Therefore, it is of primary importance that health authorities in countries continue to share information, monitor the rise and prevalence's of variants of concern and interest in their territory, and follow international recommendations to ramp up sequencing and testing[70]. Hopefully, this will improve target vaccination campaigns in hot-spots and change diagnostic methodologies better suited to monitor the rise/spread/containment of these variants. This will allow for better design of public policy and efforts[69] to continue the fight against the virus.

Data and Methods

Methods

For the discovery of the regions of interest in each variant we used two methodologies, that yield similar results. For the variants B.1.1.7, B.1.351 and P.1 we used a methodology based in Convolution Neural Networks (CNN) as explained in[57]. For the variants B.1.617+ and B.1.519 we used a methodology based in Evolutionary Algorithms (EAs)[58].

CNN

Following the procedure described in Lopez et al.[57], 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[71] for the features uncovered in the previous step. While in[57] 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 1, and is the same as the one previously reported in[57]. 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)[72,73] 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[71] 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[23].

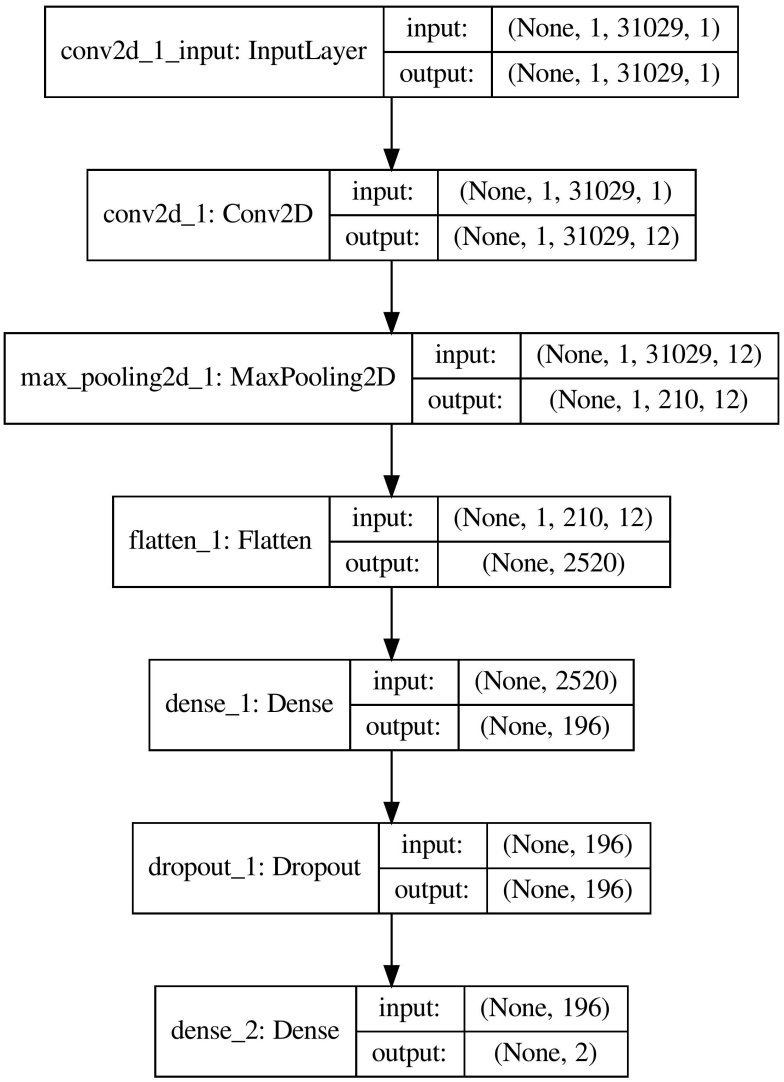

Figure 1. CNN architecture used to classify variants B.1.1.7, B.1.351, and P.1.

Evolutionary Algorithms

Another approach to generating primers is to use EAs. This approach has the advantage that we can parallelize the procedure and reduce the time required to get the regions of interest in comparison to the CNN-based method. In comparison to the 16 hours required to the CNN approach, each single run lasts around 62 minutes with 5 threads on a 64-bit Windows 10 laptop with Intel Xeon E-2186M. For the EA, We create a set of individuals of size 21-bps randomly considering the available samples of the variant of interest (e.g. B.1.1.519) and other variants. Next, we calculate what is known as cost function, which is given by the following:

$$\mathscr{F}(I) = w_p \cdot \mathscr{P}(I) + w_c \cdot \mathscr{C}(I) + w_n \cdot \mathscr{N}(I) + w_t \cdot \mathscr{T}(I) \tag{1}$$

with w_p, w_c, w_n, w_t representing the weights associated to each term.

$$\mathscr{P}(I) = \sum_{i=0}^{T} P(I, s_i), \tag{2}$$

$\mathscr{P}(I)$ is evaluating the presence of the sequence selected as candidate primer I inside training samples labeled with the variant of interest, and its absence from samples of other variants, T is the number of samples in the training set, s_i is the i-th sample in the training set. Function P is defined as:

$$P(I, s_i) = \begin{cases} 0, & \text{if } I \text{ is found inside } s_i \text{ and } L(s_i) == L(s_k) \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

where $L(s)$ returns the class label of sample s. In other words, $P(I, s_i)$ equals 1 if sequence I is found inside a sample with the same class label as sample s_k, the origin of sequence I. So, if the 21-bps sequence I is found inside a sample that does not belong to the variant of interest, or is not found in a sample that belongs to the variant of interest, the solution is penalized.

The second term of the weighted sum takes into account the GC content of the candidate primer:

$$\mathscr{C}(I) = 0.5 - \sum_{i=0}^{21} \frac{C(I(i))}{21} \quad \text{where } C(b) = \begin{cases} 1, & \text{if base } b \text{ is C or G} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

where $I(i)$ represents the base in position i inside sequence I. The following element of the weighted sum is $\mathscr{N}$, defined as:

$$\mathscr{N}(\mathscr{I}) = \sum_{i=0}^{21} N(I(i)) \quad \text{where} \quad N(b) = \begin{cases} 1, & \text{if base } b \text{ is N} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

that takes into account the presence of N symbols in the sequence, indicating an error in the read. The ideal primer candidate should only contain A, C, G, or T values.

The final term tackles the requirement of having a melting temperature T_m centered around 60°. Specialized literature[71] provides an equation to compute T_m for a sequence I:

$$T_m(I) = 81.5 + 16.6 * log_{10}([Na+]) + 41 * \mathscr{C}(I) - 600/l(I) \tag{6}$$

where $\mathscr{C}(I)$ is the content of C and G bases in sequence I, as described in Equation 4, $[Na+]$ is the molar sodium concentration, and $l(I)$ is the length of sequence I, in bps. We used the value of $[Na+] = 0.2$ as described in[71], while $l(I) = 21$ by design. The term taking into account T_m will then be:

$$\mathscr{T}(I) = |60 - T_m(I)| \tag{7}$$

The EA is set with a population of size $\mu = 200$, generating offspring of size $\lambda = 200$. The entire population is replaced by its offspring at each generation, using a (λ, μ) replacement strategy, with a tournament selection of size $\tau = 2$, a mutation acting on integer values, a one-point crossover, and a stop condition set on 100 generations. For more details, please refer to[58].

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 discovery described in detail in[57]. The 8,923 SARS-CoV-2 samples downloaded from the GISAID repository show 278 variants besides B.1.1.7. From the found combinations and known mutations we generated primer sets and test them in 7,257 non-repeated sequences of 8,269 sequences lineage B.1.1.7 submitted between 21.04.2021 and 30.04.2021 from the GISAID repository. We limited the number of sequences to a range, given that we can only download 10,000 sequences each time.

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 using CNN to find the primers. Next, we generated primer sets and test them in 8,420 non-repeated sequences of 9,473 sequences lineage B.1.351 available on May 5th from the GISAID repository.

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 using CNN to find the primers. Next, we generated primer sets and test them in 6,505 non-repeated sequences of 7,314 sequences lineage P.1 available on May 5th from the GISAID repository.

Variant B.1.617.2

From the GISAID repository, we downloaded 2,085 sequences of the lineage B.1.167+ variant on May 5th, 2021. This includes 111 B.1.617, 1,080 B.1617, 836 B.1.617.2 and 58 B.1.617.3. After removing repeated sequences, we get 1,908 non-repeated sequences. We added the 1,908 sequences to 6,819 sequences of other variants.We assigned label 1 to sequences belonging to variant B.1.617+, and 0 to the rest of the samples using EAs to find the primers. As B.1.617.2 was declared as a Variant of Concern, we made a second experiment using

Variant B.1.1.519

From the GISAID repository, we downloaded 5,265 sequences with "Mexico" as origin on May 5th, 2021. This includes 355 B.1.1.222, 401 B.1, 3,035 B.1.1.519 and 1,474 from other variants. After removing repeated sequences, we get 5,203 non-repeated sequences. We added the 5,203 sequences to 6,819 sequences of other variants.We assigned label 1 to sequences belonging to Mexico, and 0 to the rest of the samples using EAs to find the primers. Then, we made a second experiment, where we separate the 3,035 sequences B.1.1.519 from the rest and labeled them as 1, while the rest as 0.

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. Peacock, T. P. *et al.* The furin cleavage site of sars-cov-2 spike protein is a key determinant for transmission due to enhanced replication in airway cells. *bioRxiv* (2020).
11. 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* **78**, 779–784 (2020).
12. Lasek-Nesselquist, E., Pata, J., Schneider, E. & George, K. S. A tale of three sars-cov-2 variants with independently acquired p681h mutations in new york state. *medRxiv* (2021).
13. Lubinski, B., Tang, T., Daniel, S., Jaimes, J. A. & Whittaker, G. Functional evaluation of proteolytic activation for the sars-cov-2 variant b. 1.1. 7: role of the p681h mutation. *bioRxiv* (2021).
14. Brown, J. C. *et al.* Increased transmission of sars-cov-2 lineage b. 1.1. 7 (voc 2020212/01) is not accounted for by a replicative advantage in primary airway cells or antibody escape. *BioRxiv* (2021).
15. Plante, J. A. *et al.* Spike mutation d614g alters sars-cov-2 fitness. *Nature* **592**, 116–121 (2021).
16. Zhang, L. *et al.* Sars-cov-2 spike-protein d614g mutation increases virion spike density and infectivity. *Nat. communications* **11**, 1–9 (2020).
17. 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.
18. for Disease, C. & Prevention, C. Emergence of SARS-CoV-2 b.1.1.7 lineage — united states, december 29, 2020–january 12, 2021 (2021).
19. 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).
20. 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).
21. Lan, J. *et al.* Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature* **581**, 215–220 (2020).
22. Brief report: New variant strain of sars-cov-2 identified in travelers from brazil (2021).
23. Faria, N. Genomic characterisation of an emergent sars-cov-2 lineage in manaus: preliminary findings (2021).
24. Felipe Naveca, e. a. Genomic characterisation of an emergent sars-cov-2 lineage in manaus: preliminary findings. *SARS-CoV-2 coronavirus* (2021).
25. Voloch, C. M. *et al.* Genomic characterization of a novel sars-cov-2 lineage from rio de janeiro, brazil. *medRxiv* (2020).
26. 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).
27. 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).

28. England, P. H. Investigation of novel SARS-CoV-2 variant variant of concern 202012/01 (2021).

29. Liu, Z. *et al.* Identification of sars-cov-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell host & microbe* **29**, 477–488 (2021).

30. 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).

31. 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).

32. Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mrna vaccine-elicited antibodies. *Nature* 1–10 (2021).

33. Weisblum, Y. *et al.* Escape from neutralizing antibodies by sars-cov-2 spike protein variants. *Elife* **9**, e61312 (2020).

34. Nonaka, C. K. V. *et al.* Genomic evidence of a SARS-CoV-2 reinfection case with e484k spike mutation in brazil. *Preprints* (2021).

35. Resende, P. C. *et al.* Spike e484k mutation in the first SARS-CoV-2 reinfection case confirmed in brazil, 2020. *Preprints* (2021).

36. 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.

37. 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).

38. Organization, W. H. Covid-19 weekly epidemiological update: Data as received by who from national authorities, as of 25 april 2021, 10 am cet (2021).

39. Hodcroft, E. B. Covariants: SARS-CoV-2 mutations and variants of interest (2021).

40. Sars-cov-2 variants of concern and variants under investigation in england, technical briefing 9 (2021).

41. for Disease Prevention, E. C. & Control. Emergence of SARS-CoV-2 B.1.617 variants in india and situation in the eu/eea (2021).

42. Motozono, C. *et al.* An emerging sars-cov-2 mutant evading cellular immunity and increasing viral infectivity. *bioRxiv* (2021).

43. Cherian, S. *et al.* Convergent evolution of sars-cov-2 spike mutations, l452r, e484q and p681r, in the second wave of covid-19 in maharashtra, india. *bioRxiv* (2021).

44. Kumar, V., Singh, J., Hasnain, S. E. & Sundar, D. Possible link between higher transmissibility of B. 1.617 and B.1.1.7 variants of SARS-CoV-2 and increased structural stability of its spike protein and hace2 affinity. *bioRxiv* (2021).

45. Tchesnokova, V. *et al.* Acquisition of the l452r mutation in the ace2-binding interface of spike protein triggers recent massive expansion of sars-cov-2 variants. *bioRxiv* (2021).

46. Di Giacomo, S., Mercatelli, D., Rakhimov, A. & Giorgi, F. M. Preliminary report on sars-cov-2 spike mutation t478k. *J. Med. Virol.* (2021).

47. Muecksch, F. *et al.* Development of potency, breadth and resilience to viral escape mutations in sars-cov-2 neutralizing antibodies. *bioRxiv* (2021).

48. Yadav, P. D. *et al.* Sars cov-2 variant b. 1.617. 1 is highly pathogenic in hamsters than b. 1 variant. *bioRxiv* (2021).

49. Ferreira, I. *et al.* Sars-cov-2 b.1.617 emergence and sensitivity to vaccine-elicited antibodies. *bioRxiv* (2021).

50. Hoffmann, M. *et al.* Sars-cov-2 variant b. 1.617 is resistant to bamlanivimab and evades antibodies induced by infection and vaccination. *bioRxiv* (2021).

51. Sars-cov-2: Tracking variants of concern version 2.0 – updated 3.30.21 (2021).

52. Jaspe, R. C. *et al.* Importance of e484k and n501y mutations in sars-cov-2 for genomic surveillance: rapid detection by restriction enzyme analysis. *medRxiv* (2021).

53. Afzal, A. Molecular diagnostic technologies for covid-19: Limitations and challenges. *J. advanced research* (2020).

54. Organization, W. H. *et al.* Molecular assays to diagnose covid-19: summary table of available protocols (2020).

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

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

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

58. Lopez-Rincon, A. *et al.* Design of specific primer sets for sars-cov-2 variants using evolutionary algorithms. *GECCO* (2021).

59. 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.

60. Sars-cov-2 variants of concern and variants under investigation in england, technical briefing 6 (2021).

61. Sars-cov-2 variants of concern and variants under investigation in england, technical briefing 10 (2021).

62. Rambaut, A. *et al.* Addendum: A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **6**, 415–415 (2021).

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

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

65. 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).

66. 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.

67. Qué es lo que sabemos acerca de b.1.1.222, la variante detectada en méxico (2021).

68. Coronavirus en méxico: qué es el modelo centinela, el sistema de vigilancia que calcula que la epidemia de covid-19 es 8 veces mayor que lo reportado (2020).

69. Sánchez-Talanquer, M. *et al.* La respuesta de méxico al covid-19:estudio de caso. *Inst. Glob. Heal. Sci.* (2021).

70. Red regional de vigilancia genómica de covid-19 (2021).

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

72. 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).

73. 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.