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

Carmina A. Perez-Romero[2], Alberto Tonda[3], Lucero Mendoza-Maldonado[4], Etienne Coz[8], Patrick Tabeling[9], Jessica Vanhomwegen[9], Eric Claassen[5], Johan Garssen[1,6], Aletta D. Kraneveld[1], and Alejandro Lopez-Rincon[1,7,*]

[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
[8]CBI, ESPCI Paris, Université PSL, CNRS, 75005 Paris, France
[9]CIBU, Institut Pasteur, 25-28 rue du Dr Roux, Paris 15
[*]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 labelled 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 and B.1.617.2 respectively. 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. Finally, we tested the designed primers for B.1.1.7 using RT-PCR. 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 variants of concern (VOC) with acquired mutations that may pose a treat to global health such as: a higher rate of transmissibility, change in epidemiology, virulence, clinical presentation, mortality, vaccine/therapeutics resistance, or decrease in effectiveness of public health measures[2].

On December 14th, 2020, Public Health authorities in England reported a new SARS-CoV-2 variant, [3–5], which belongs to the B.1.1.7 (which include all Q. lineages used for fine geographical localization of the variant) Pango lineage[6,7], GRY clade from GISAID (Global Initiative on Sharing All Influenza Data)[3,8,9], Nextstrain clade 20I(V1)[10]. This was the first VOC raised by the World Health Organization (WHO), and was recently termed as the Alpha variant[2]. 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)[4,9]. Additional amino acid changes in the Spike domain, found only in a subset of the population, which are being monitored by the WHO, include those in the 484K and 452R positions[2]. 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[9,11] and to be more infectious in mice[12]. 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 [13–15], the effect of such mutations is still under debate[16,17]. The D614G mutation has been found to enhances viral replication, viron density and infectivity[18,19]. The alpha variant has 43 to 90% higher reproduction number than pre-existing variants increasing its transmissibility [3,20,21]. The jury is still out on its effect on clinical severity and outcome, although there is evidence of increased clinical severity[22] and increased mortality risk when compared with previous variants[23]. The Alpha variant remains susceptible to most monoclonal antibody therapy treatment targeting Spike protein for Covid-19 like: Bamlanivimab-etesevimab, Casirivimab-imdevimab, and others [24–28]. Plus it has minimal impact on neutralization by convalescent and post-vaccination sera[29–32]. The presence of the Alpha variant rapidly increasing in the UK earlier this year, spreading across the globe becoming the first major global circulating VOC [8,9,21,33,34], before the arrival of the most recent VOC, the Delta variant which is now displacing other variants[8,35–37].

In parallel, on December 18th, 2020, the WHO declared a new VOC, the Beta variant, first identified by the South African authorities that was rapidly spreading across 3 of their provinces and displacing other variants[5,38]. This variant is also knows know as B.1.351 Pango lineage, GH/501Y.V2 GISAID clade, and 20H (V2) Nextrain clade. 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)[38]. Additional amino acid changes being monitored by the WHO are those on L18F position[2]. 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[10]. Although uncommon in SARS-CoV-2 variant strains, the E484K mutation has been shown to moderately enhance binding affinity of the ACE2 receptor[11]. Mutation K417N is located in a key contact between the S protein and the human ACE2 receptor: a previous mutation in this residue has been shown to contribute to enhanced affinity of SARS-CoV-2 to the ACE2 receptor when compared to SARS-CoV[39,40]. The E484K and K417N mutations in combination with the N501Y have been shown to affect neutralization by monoclonal antibodies and convalescent sera[38,41–46]. As such the Beta variant has reduced susceptibility to monoclonal antibody therapy like bamlanivimab and etesevimab[25,28,47,48] and it also shows reduced neutralization to convalescent and post-vaccination sera[27,31,32,49], however it remains susceptible to casirivimab and imdevimab monoclonal antibody therapies[25,28,50]. The Beta variant also has a 50% increase in transmission rate[51], and is associated with an increase in excess deaths per week[38]. The Beta variant spread across Africa becoming the main VOC in this region earlier this year[6,8,34], before being displaced by the Delta variant.

On early January, Japan reported a new SARS-CoV-2 variant from 4 travellers from Brazil[52]. This variant was later reported to be widely circulating in Brazil[33,53]. On the 11th of January the WHO assigned it as VOC, the Gamma variant[2], it is also know as P.1 Pango lineage, GR/501Y.V3 GISAID clade or 20J (V3) Nexstrain clade. 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. Additional amino acid changes being monitored by the WHO are those on 681H position [2]. Although the rate of infection and mortality of the Gamma variant are still being investigated, studies have shown a change in pattern of infection, as well as a significant change in case fatality rates associated with this variant suggesting a change in pathogenicity and virulence profile of the Gamma variant[54,55]. The Gamma variant significantly reduces the susceptibility to bamlanivimab and etesevimab monoclonal antibodies[28,47,48], but remains susceptible to casirivimab and imdevimab monoclonal antibody therapies[28,50]. The Gamma variant is more resistant to neutralization by convalescent and post-vaccination sera[56], plus an increase in reinfection cases has been associated to this variant[57–59], raising concern about the need to test monoclonal antibodies in clinical against new VOC, and plausible updates to mRNA vaccines to avoid a potential loss of clinical efficacy[60]. The P.1 variant is effects still requires further study, it has been identified in the Amazonia in Brazil, as in other countries[10,53,57]. The Gamma variant spread across South America becoming the main VOC in this region until this day[8].

On May 11^{th} 2021, the WHO announced a VOC, the Delta variant, which was first identified in India and linked to the rapid and deathly second wave of the country[33,61]. The Delta variant is also knows as B.1.617.2 Pango lineage (which include all AY. lineages used for fine geographical localization of the variant and do not imply any functional biological differences)[37], G/478K.V1 GISAID clade or 21A Nexstrain clade. The B.1.617.2 harbours the next mutations in the Spike protein T19R, △156-157, L452R, T478K, D614G, P681R,R158G and D950N[33,61–64]. Additional amino acid changes being monitored by the WHO are those on 417N position [2]. The L452R mutation has been shown to increase SARS-CoV-2 viral infectiousness and replication[65]. Interestingly. L452R and E484Q have been found to disrupt the interfacial interactions of the Spike RBD with neutralizing antibodies[66,67]. 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[68]. The T478K is located at the interface of the Spike/ACE2 interaction domain[15,69], and enhance stabilization of the Spike RBD with the ACE2 complex[70]. The Delta variant has been show to lead to a 64% higher household transmission rate[71], higher viral loads[72,73] and increase in hospital and ICU admissions when compared with the Alpha variant[74,75]. It has been associated to an increase in infections and excess deaths across the globe[64,72,76,77]. The Delta variant has been shown to be resistant to monoclonal antibody therapy treatment with bamlanivimab[25,28,47,78], but remains susceptible to casirivimab, imdevimab and etesevimab[25,28,47,50]. Furthermore, the Delta variant has reduce neutralization by convalescent[25] and modestly reduced sensitivity to Comirnaty/BNT162b2 mRNA vaccine[75,78–80] and ChadOx-1 vaccine sera[36,75,80] requiring two-doses for full protection. Although several cases of reinfection from the Delta variant have been documented in partially and fully vaccinated individuals and previously Covid-19 infected patients[81], the severity of infection and risk of hospitalization remains lower in vaccinated individuals[74,82]. Therefore, the main concern remains in the greater burden seeing so far on health-care services due to outbreaks on unvaccinated population[74]. The Delta variant has spread like wildfire across the globe rapidly since its explosive rise in India on May, it has now become the major VOC globally rapidly displacing other variants[8,35,71,73]. Given its high prevalence and clinical implications many countries have started implemented stronger vaccine policies[83], considering the need of additional doses for vulnerable population[84] and the need for vaccine updates to combat waning immunity[85].

Given the rapid spread of these VOC, and their impact on 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)[86,87]. Recently, Public Health England was able to identify the Alpha variant through their national surveillance system which allowed them to notice the rise in Covid-19 positive cases in South Easter England. The Alpha variant 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[3,67]. In the case of the Beta[38,88,89] and Gamma[53,90] variants were identify through the increase in positive Covid-19 cases and deaths and by random whole genome sequencing of positive Covid-19 samples, through their national and international diagnosis and surveillance system efforts. In the case of the Delta variant, most countries rely on the clinical differences between the variants and if the RT-PCR has a S fall out (indicating Alpha variant) or not indicating a plausible Delta variant[75]. Although several PCR methods and kits for the identification of this variants have been proposed[91,92], they usually rely on the well know Spike protein mutations, which often are shared with other VOIs and variants of SARS-CoV-2, and rapidly get obsolete due to the virus evolution[93]. Making sequencing the only reliable way to diagnose it, however this is cost prohibiting and many times not available in lower income countries.

In a previous work[94], 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 algorithms[95]. 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. In addition, we have validated our B.1.1.7 primers by RT-PCR, finding them to be specific to B.1.1.7 and sensitive enough for detection by this method. 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[96]. 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 (Fig. 1).

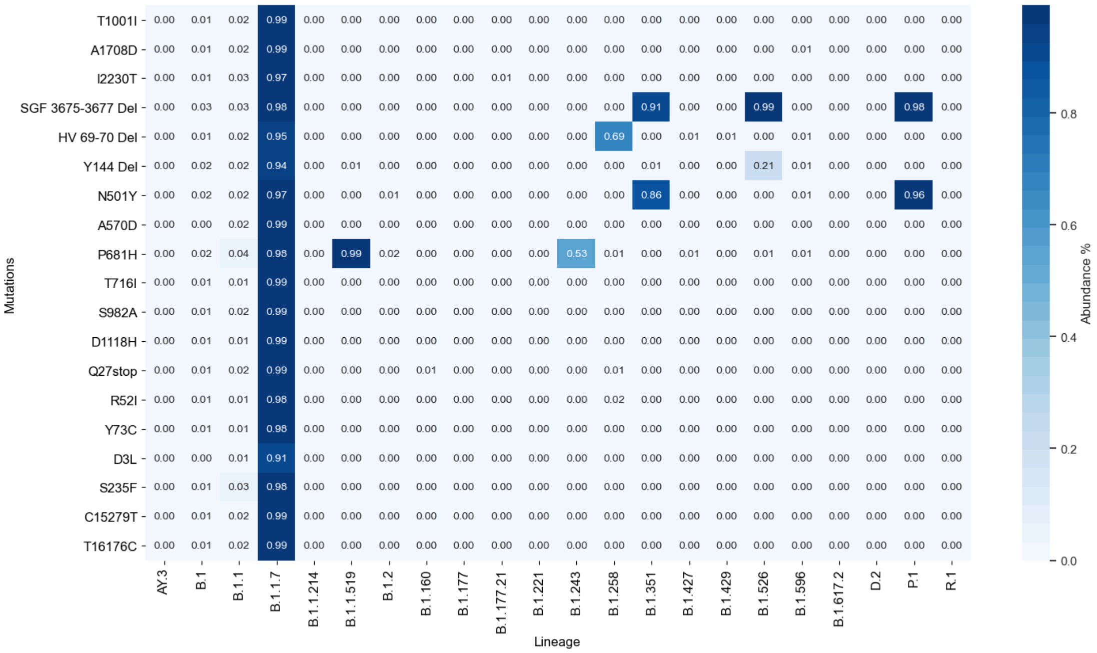


Figure 1. Frequency of appearance of characteristic mutations of B.1.1.7 Variant in 2,096,390 SARS-CoV-2 sequences downloaded from GISAID in August 11th, 2021.

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[96] 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 (Fig. 2).

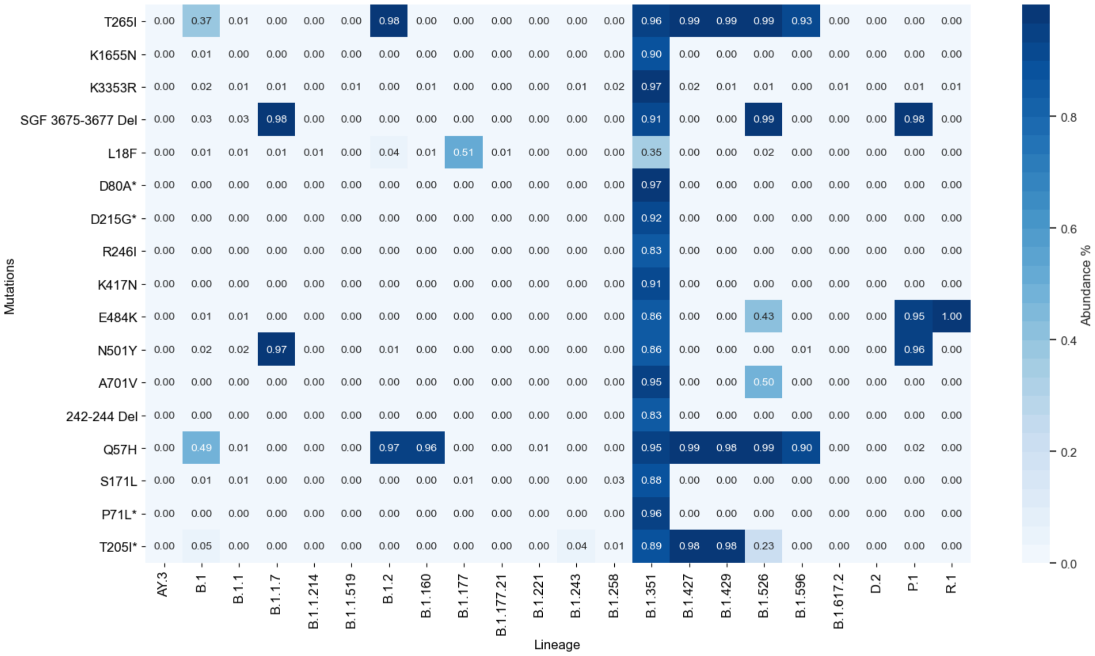
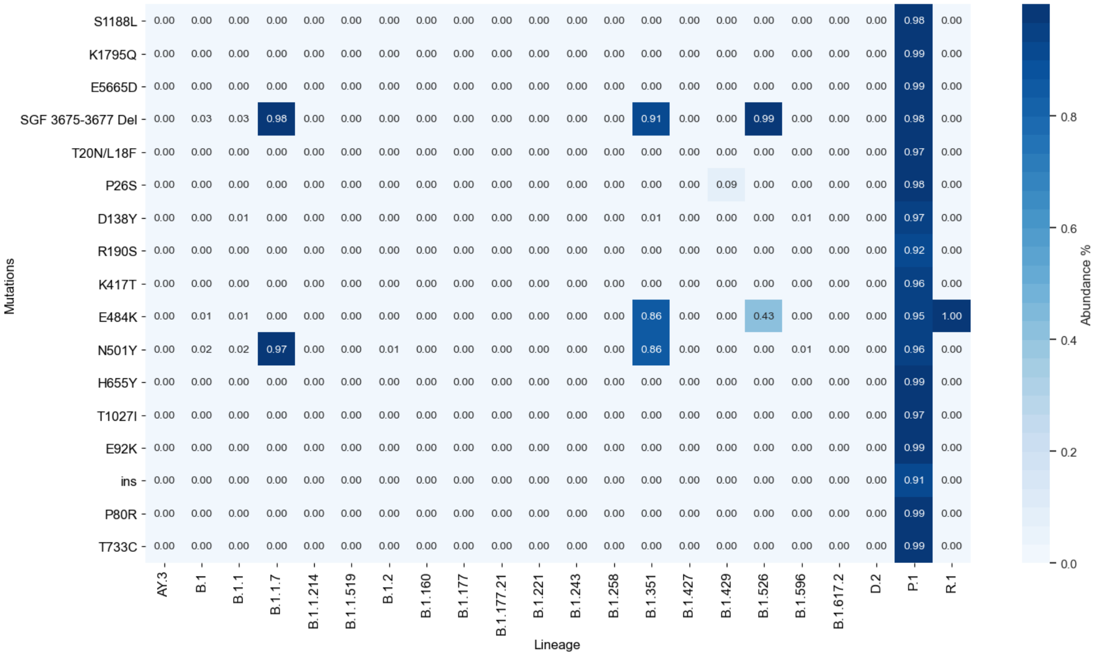


Figure 2. Frequency of appearance of characteristic mutations of B.1.351 Variant in 2,096,390 SARS-CoV-2 sequences downloaded from GISAID in August 11th, 2021.

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[53] 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**) (Fig. 3).



	AY.3	B.1	B.1.1	B.1.1.7	B.1.1.214	B.1.1.519	B.1.2	B.1.160	B.1.177	B.1.177.21	B.1.221	B.1.243	B.1.258	B.1.351	B.1.427	B.1.429	B.1.526	B.1.596	B.1.617.2	D.2	P.1	R.1
S1188L	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.98	0.00
K1795Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00
E5665D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00
SGF 3675-3677 Del	0.00	0.03	0.03	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.99	0.00	0.00	0.00	0.98	0.00
T20N/L18F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00
P26S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.98	0.00
D138Y	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.97	0.00
R190S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.00
K417T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00
E484K	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.00	0.43	0.00	0.00	0.00	0.95	1.00
N501Y	0.00	0.02	0.02	0.97	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.01	0.00	0.00	0.96	0.00
H655Y	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00
T1027I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00
E92K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00
ins	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.91	0.00
P80R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00
T733C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00

Figure 3. Frequency of appearance of characteristic mutations of P.1 Variant in 2,096,390 SARS-CoV-2 sequences downloaded from GISAID in August 11^{th}, 2021.

Variant B.1.617.2

Using the sequences of variant B.1.617.2 downloaded from the GISAID repository we ran an evolutionary algorithm 10 times to find the most important sequences to separate the B.1.617.2 from the rest of the sequences. From 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, D377Y and Del 156-157[97]. Where sequences of mutations T478K (**GCCGGTAGCAAACCTTGTAAT**), V82A (**GCCAGATCAGCTTCACCTAAA**) and R203M (**GGCAGCAGTATGGGAACTTCT**) can be used as forward primers in multiplex approach (Fig. 4). It is important to consider that lineage AY.3 with B.1.617.2 are part of the *Delta* variant.

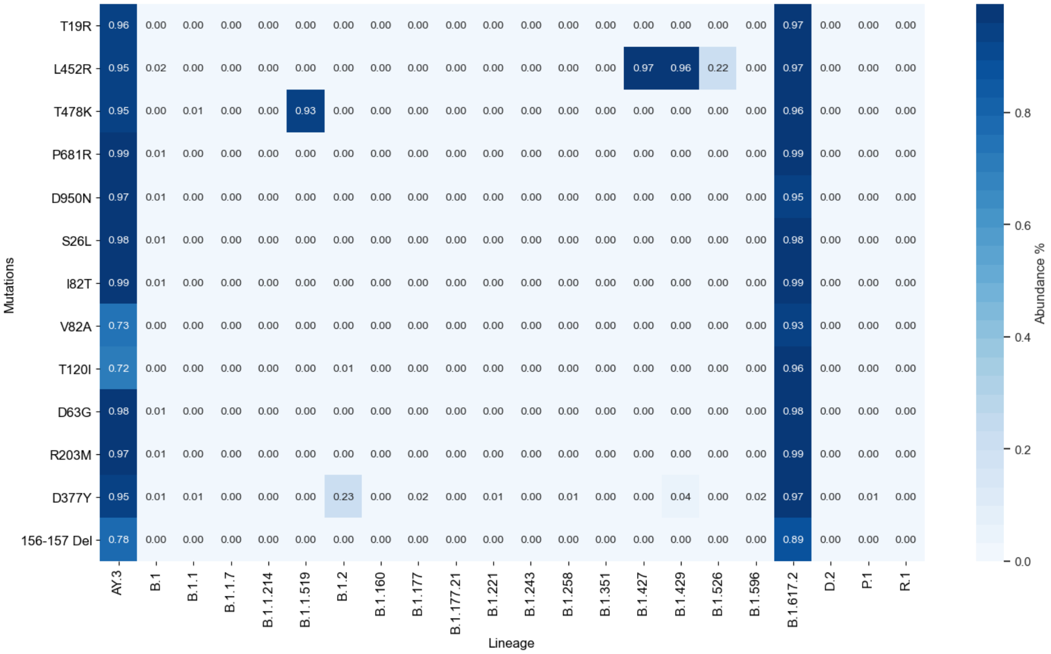


Figure 4. Frequency of appearance of characteristic mutations of B.1.617.2 Variant in 2,096,390 SARS-CoV-2 sequences downloaded from GISAID in August 11^{th}, 2021.

Experimental evaluation of B.1.1.7 Primer specificity

The RT-PCR amplification curves obtained with B.1.1.7 specific primer set B.1.1.7-1 and SARS-CoV-2 generic primers IP2 and IP4 (following the Pasteur Institute Protocol)[98] on two SARS-CoV-2 strains are shown result in Fig. 5. As expected, only the B.1.1.7 strain is amplified the B.1.1.7-1 primers, while both the Wuhan reference strain and the B.1.1.7 strain are detected by the generic primers. Fig. 5. Comparison between non-specific primer set IP2 and IP4 and our designed primer set B.1.1.7-1 for B.1.1.7 variant and the original Wuhan SARS-CoV-2 strain.

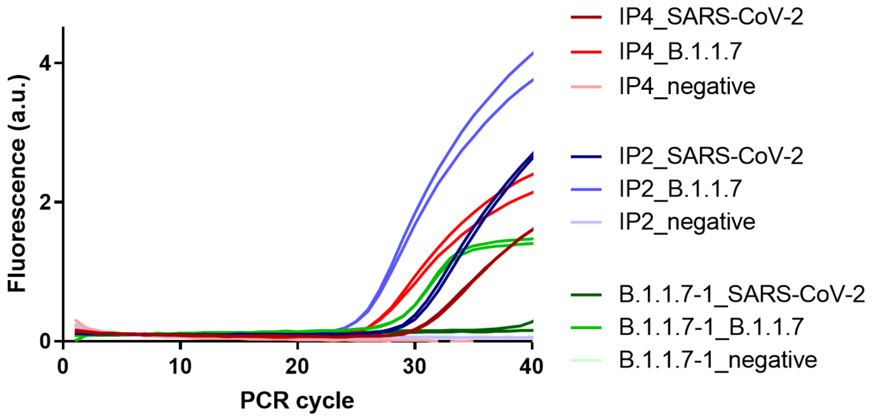


Figure 5. Comparison between non-specific primer set IP2 and IP4 and our designed primer set B.1.1.7-1 for B.1.1.7 variant and others.

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, and P.1 we created 2 options for each one, and for B.1.617.2 we created 3 different combinations (Table 1). 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. The frequency of appearance for the different sequences is in Fig. 6.

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[86]. 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[3]. 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[3]. This $\triangle$69-70 deletion, which affects its N-terminal domain, has been occurring in several different SARS-CoV-2 variants around the world[10, 99] and has been associated with other spike protein receptor binding domain changes[9]. 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[86, 87, 100]. This is consistent with other existing primer designs like CoV2R-3 in the S-gene[101], 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.[102].

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. In

Table 1. Sequences of the primer sets for B.1.1.7, B.1.351, P.1 and B.1.617.2 lineages.

	Mutation	Sequence	Target
B.1.1.7-1 (244 bp)	HV 69-70 deletion	F-CATGCTATCTCTGGGACCAAT	B.1.1.7
	Y144 deletion	R-TGTTGTTTTTGTGGTAAACACC	B.1.1.7
B.1.1.7-2 (437 bp)	S982A	F-TGATATCCTTGCACGTCTTGA	B.1.1.7
	D1118H	R-CAGACACAAATGTGTTGTGTGTAGT	B.1.1.7
B.1.1.7-3 (936 bp)	C15279T	F-TGATGTAGAAAACCCTCATCTTATG	B.1.1.7
	T16176C	R-CTCAAGGTATTGGGAACCTGA	B.1.1.7
B.1.351-1 (425 bp)	D80A*	F-AGAGGTTTGCTAACCCTGTCC	B.1.351
	D215G*	R-CTGAGGGAGACCACGCACTAA	B.1.351
B.1.351-2 (619 bp)	D215G*	F-TTAGTGCGTGGTCTCCCTCAG	B.1.351
	K417N	R-AATCAGCAATATTTCCAGTTTGC	B.1.351
P.1-1 (375 bp)	T20N/L18F	F-TTTTACAAACAGAACTCAATTACCC	P.1
	D138Y	R-CCCAAAAATGGATAATTACAAAA	P.1
P.1-2 (266 bp)	ins28269-28273 AACA	F-CGAACAAACAAACTAAAATGTCTGA	P.1
	P80R	R-TTGGTCATCTCGACTGCTATT	P.1
B.1.617.2-1 (879 bp)	I82T	F-CTACCGCAATGGCTTGTCTT	B.1.617.2
	V82A	R-TGAAGCTGATCTGGCACGTA	B.1.617.2
B.1.617.2-2 (434 bp)	D63G	F-ATGGCAAGGAAGGCCTTAAA	B.1.617.2
	R203M	R-CATACTGCTGCCTGGAGTTGA	B.1.617.2
B.1.617.2-3 (429 bp)	T19R	F-TCTTAGAACCAGAACTCAATTACCC	B.1.617.2
	del156-157	R-CTAGAATAAACTCCACTTTCCATCC	B.1.617.2

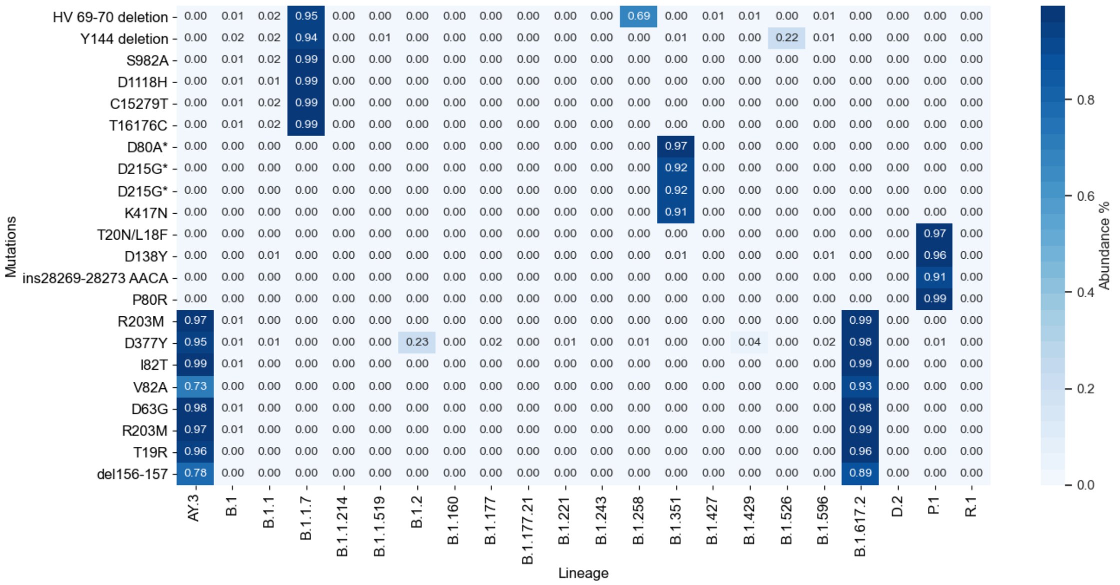


Figure 6. Frequency of appearance of the different primer sets in 2,096,390 SARS-CoV-2 sequences downloaded from GISAID in August 11th, 2021.

addition, we tested the detection and specificity for the variant B.1.1.7. 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 rest of the proposed primer sets delivered by our automated methodology will still require laboratory testing to be validated, our methodology 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

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[94]. For the variants B.1.617.2 we used a methodology based in Evolutionary Algorithms (EAs)[95].

CNN

Following the procedure described in Lopez et al.[94], 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[103] for the features uncovered in the previous step. While in[94] 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 7, and is the same as the one previously reported in[94]. 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)[104, 105] 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[103] in the canonical sequences EPI_ISL_601443 for B.1.1.7[3], EPI_ISL_678597 for B.1.351 and EPI_ISL_804814 for P.1[53].

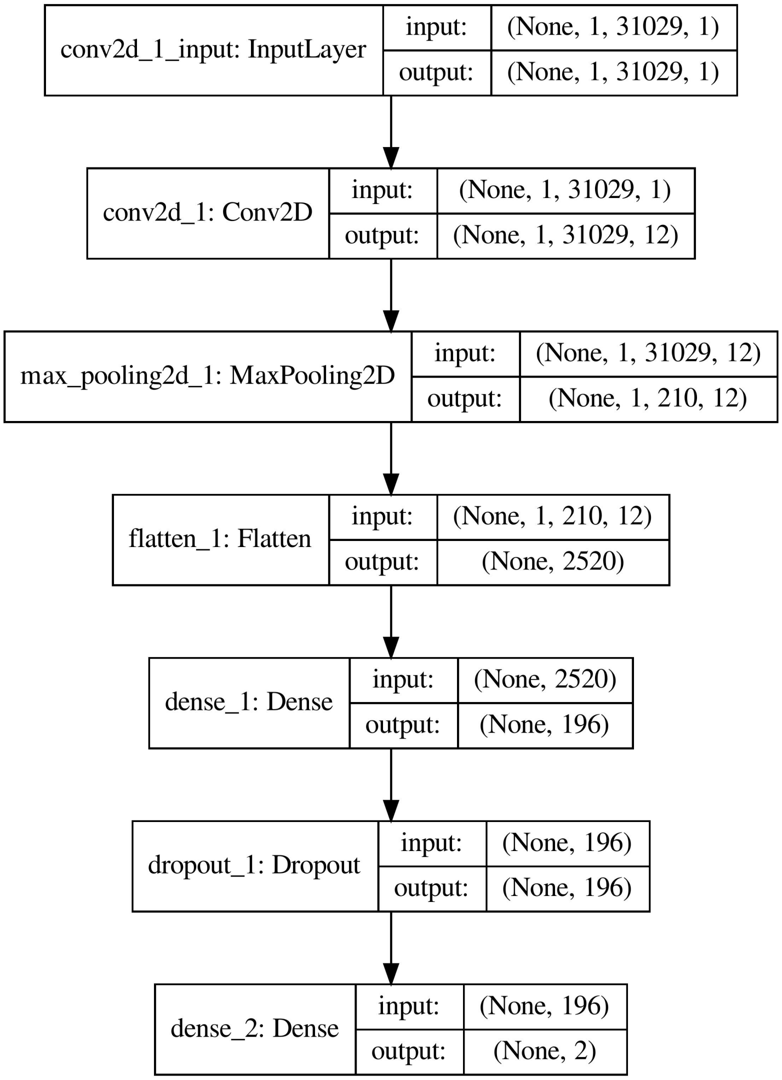

conv2d_1_input: InputLayer
input: (None, 1, 31029, 1)
output: (None, 1, 31029, 1)
conv2d_1: Conv2D
input: (None, 1, 31029, 1)
output: (None, 1, 31029, 12)
max_pooling2d_1: MaxPooling2D
input: (None, 1, 31029, 12)
output: (None, 1, 210, 12)
flatten_1: Flatten
input: (None, 1, 210, 12)
output: (None, 2520)
dense_1: Dense
input: (None, 2520)
output: (None, 196)
dropout_1: Dropout
input: (None, 196)
output: (None, 196)
dense_2: Dense
input: (None, 196)
output: (None, 2)

Figure 7. 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[103] 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[103], 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[95].

Experimental evaluation of B.1.1.7 Primers

Amplification efficiency of the designed primer sets were evaluated using viral RNA extracts from two sequenced SARS-CoV-2 strains: the original Wuhan strain 210207 (GISAID N° EPI_ISL_437689) and VOC B.1.1.7 strain (GISAID N° EPI_ISL_683466). Viral RNA were extracted from infected cell culture supernatants using the NucleoSpin Dx Virus kit (Macherey-Nagel), following the manufacturers' protocol. Viral RNA extracts (5 µL) were analyzed either using the IP2/IP4 dualplex real-time reverse-transcriptase (RT)–PCR assay, developed by following the Pasteur Institute and targeting conserved regions of the SARS-CoV-2 RdRP gene[98], or primer set B.1.1.7-1 (Table 1), using the LightCycler EvoScript RNA SYBR Green I Master kit (Roche). Both RT-PCR assays were conducted on a LightCycler® 480 System (Roche), using the thermal cycling program described in the Pasteur Institute protocol[98].

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[94]. Next, from the found combinations and known mutations we generated primer sets and test them in 2,096,390 SARS-CoV-2 sequences downloaded in August 11^{th}, 2021, where 1,051,740 sequences are B.1.1.7. The total number of sequences by lineage is in the Table 2 on the supplemenatary material.

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 2,096,390 SARS-CoV-2 sequences downloaded in August 11^{th}, 2021 where 30,650 sequences are B.1.351.

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 2,096,390 SARS-CoV-2 sequences downloaded in August 11^{th}, 2021 where 59,692 sequences are P.1.

Variant B.1.617.2

From the GISAID repository, we downloaded 836 sequences of the B.1.617.2 variant on May 5th, 2021. We added sequences to 6,819 sequences of other variants, and we assigned label 1 to sequences belonging to variant B.1.617.2, and 0 to the rest of the samples using EAs to find the primers. Finally, we generated primer sets and test them in 2,096,390 SARS-CoV-2 sequences downloaded in August 11^{th}, 2021 where 366,831 sequences are B.1.617.2.

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. Tracking sars-cov-2 variants (2021).

3. Chand, M., Hopkins, S. & Dabrera, G. Investigation of novel SARS-COV-2 variant: Variant of Concern 202012/01 (2020).

4. 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).

5. SARS-CoV-2 variants (2020).

6. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **30** (2021).

7. Rambaut, A. *et al.* A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nat. microbiology* **5**, 1403–1407 (2020).

8. Shu, Y. & McCauley, J. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22** (2017).

9. 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).

10. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

11. 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).

12. Gu, H. *et al.* Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* **369** (2020).

13. 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).

14. 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).

15. 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).

16. 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).

17. 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).

18. Plante, J. A. *et al.* Spike mutation d614g alters sars-cov-2 fitness. *Nature* **592**, 116–121 (2021).

19. Zhang, L. *et al.* Sars-cov-2 spike-protein d614g mutation increases virion spike density and infectivity. *Nat. communications* **11**, 1–9 (2020).

20. 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.

21. Davies, N. G. *et al.* Estimated transmissibility and impact of sars-cov-2 lineage b. 1.1. 7 in england. *Science* **372** (2021).

22. Giles, B., Meredith, P., Robson, S., Smith, G. & Chauhan, A. The sars-cov-2 b. 1.1. 7 variant and increased clinical severity—the jury is out. *The Lancet Infect. Dis.* (2021).

23. Davies, N. G. *et al.* Increased mortality in community-tested cases of sars-cov-2 lineage b. 1.1. 7. *Nature* **593**, 270–274 (2021).

24. Shen, X. *et al.* Sars-cov-2 variant b. 1.1. 7 is susceptible to neutralizing antibodies elicited by ancestral spike vaccines. *Cell host & microbe* **29**, 529–539 (2021).

25. Planas, D. *et al.* Reduced sensitivity of sars-cov-2 variant delta to antibody neutralization. *Nature* 1–7 (2021).

26. Widera, M. *et al.* Bamlanivimab does not neutralize two sars-cov-2 variants carrying e484k in vitro. *MedRxiv* (2021).

27. Wang, P. *et al.* Antibody resistance of sars-cov-2 variants b. 1.351 and b. 1.1. 7. *Nature* **593**, 130–135 (2021).

28. Kumar, S., Chandele, A. & Sharma, A. Current status of therapeutic monoclonal antibodies against sars-cov-2. *PLoS pathogens* **17**, e1009885 (2021).

29. Edara, V. V. *et al.* Infection and mrna-1273 vaccine antibodies neutralize sars-cov-2 uk variant. *medRxiv* (2021).

30. Collier, D. A. *et al.* Sars-cov-2 b. 1.1. 7 sensitivity to mrna vaccine-elicited, convalescent and monoclonal antibodies. *medRxiv* (2021).

31. Wu, K. *et al.* mrna-1273 vaccine induces neutralizing antibodies against spike mutants from global sars-cov-2 variants. *BioRxiv* (2021).

32. Emary, K. R. *et al.* Efficacy of chadox1 ncov-19 (azd1222) vaccine against sars-cov-2 variant of concern 202012/01 (b. 1.1. 7): an exploratory analysis of a randomised controlled trial. *The Lancet* **397**, 1351–1362 (2021).

33. for Disease, C. & Prevention, C. Emergence of SARS-CoV-2 b.1.1.7 lineage — united states, december 29, 2020–january 12, 2021 (2021).

34. O'Toole, Á. *et al.* Tracking the international spread of sars-cov-2 lineages b. 1.1. 7 and b. 1.351/501y-v2. *Wellcome Open Res.* **6** (2021).

35. Bolze, A. *et al.* Rapid displacement of sars-cov-2 variant b. 1.1. 7 by b. 1.617. 2 and p. 1 in the united states. *medRxiv* (2021).

36. Mlcochova, P. *et al.* Sars-cov-2 b. 1.617. 2 delta variant emergence and vaccine breakthrough. *Researchsquare* (2021).

37. Network, P. New ay lineages and an update to ay.4-ay.12 (2021).

38. 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).

39. 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).

40. Lan, J. *et al.* Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature* **581**, 215–220 (2020).

41. 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).

42. 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).

43. 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).

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

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

46. 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).

47. Food, T. U. & (FDA), D. A. Fact sheet for health care providers emergency use authorization (eua) of bamlanivimab and etesevimab (2021).

48. Liu, H. *et al.* 501y. v2 and 501y. v3 variants of sars-cov-2 lose binding to bamlanivimab in vitro. In *Mabs*, vol. 13, 1919285 (Taylor & Francis, 2021).

49. Madhi, S. A. *et al.* Safety and efficacy of the chadox1 ncov-19 (azd1222) covid-19 vaccine against the b. 1.351 variant in south africa. *MedRxiv* (2021).

50. Food, T. U. & (FDA), D. A. Fact sheet for health care providers emergency use authorization (eua) of regen-covtm (casirivimab and imdevimab) (2021).

51. Pearson, C. *et al.* Estimates of severity and transmissibility of novel south africa sars-cov-2 variant 501y. v2. 2021.

52. Brief report: New variant strain of sars-cov-2 identified in travelers from brazil (2021).

53. Faria, N. Genomic characterisation of an emergent sars-cov-2 lineage in manaus: preliminary findings (2021).

54. Freitas, A. R. R. *et al.* The emergence of novel sars-cov-2 variant p. 1 in amazonas (brazil) was temporally associated with a change in the age and sex profile of covid-19 mortality: A population based ecological study. *The Lancet Reg. Heal.* 100021 (2021).

55. de Oliveira, M. H. S., Lippi, G. & Henry, B. M. Sudden rise in covid-19 case fatality among young and middle-aged adults in the south of brazil after identification of the novel b. 1.1. 28.1 (p. 1) sars-cov-2 strain: analysis of data from the state of parana. *medRxiv* (2021).

56. Wang, P. *et al.* Increased resistance of sars-cov-2 variant p. 1 to antibody neutralization. *Cell host & microbe* **29**, 747–751 (2021).

57. Felipe Naveca, e. a. Genomic characterisation of an emergent sars-cov-2 lineage in manaus: preliminary findings. *SARS-CoV-2 coronavirus* (2021).

58. Sabino, E. C. *et al.* Resurgence of covid-19 in manaus, brazil, despite high seroprevalence. *The Lancet* **397**, 452–455 (2021).

59. Prete, C. A. *et al.* Reinfection by the sars-cov-2 p. 1 variant in blood donors in manaus, brazil. *MedRxiv* (2021).

60. Wang, Z. *et al.* mrna vaccine-elicited antibodies to sars-cov-2 and circulating variants. *Nature* **592**, 616–622 (2021).

61. 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).

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

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

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

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

66. 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).

67. 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).

68. 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).

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

70. Cherian, S. *et al.* Sars-cov-2 spike mutations, l452r, t478k, e484q and p681r, in the second wave of covid-19 in maharashtra, india. *Microorganisms* **9**, 1542 (2021).

71. Allen, H., MSc, A. V., MFPH, J. F. & MSc, K. A. Increased household transmission of covid-19 cases associated with sars-cov-2 variant of concern b. 1.617. 2: a national casecontrol study. *Prepr. https://khub.net/documents/135939561/405676950/Increased+ Household+ Transm. of+ COVID-19+ Cases+-+ nationai+ case+ study. pdf/7f7764fb-ecb0-da31-77b3-bla8ef7be9aa (accessed 11 June 2021)* (2021).

72. Bolze, A. *et al.* Sars-cov-2 variant delta rapidly displaced variant alpha in the united states and led to higher viral loads. *medRxiv* 2021–06 (2021).

73. Luo, C. H. *et al.* Infection with the sars-cov-2 delta variant is associated with higher infectious virus loads compared to the alpha variant in both unvaccinated and vaccinated individuals. *medRxiv* (2021).

74. Twohig, K. A. *et al.* Hospital admission and emergency care attendance risk for sars-cov-2 delta (b. 1.617. 2) compared with alpha (b. 1.1. 7) variants of concern: a cohort study. *The Lancet Infect. Dis.* (2021).

75. Sheikh, A., McMenamin, J., Taylor, B. & Robertson, C. Sars-cov-2 delta voc in scotland: demographics, risk of hospital admission, and vaccine effectiveness. *The Lancet* (2021).

76. (PHE), P. H. E. Sars-cov-2 variants of concern and variants under investigation in england (2021).

77. of Health Metrics, I. & (IHME), E. Covid-19 results briefing (2021).

78. Hoffmann, M. *et al.* Sars-cov-2 variant b. 1.617 is resistant to bamlanivimab and evades antibodies induced by infection and vaccination. *Cell Reports* **36**, 109415 (2021).

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

80. Lopez Bernal, J. *et al.* Effectiveness of covid-19 vaccines against the b. 1.617. 2 (delta) variant. *New Engl. J. Medicine* **385**, 585–594 (2021).

81. Shastri, J. *et al.* Severe sars-cov-2 breakthrough reinfection with delta variant after recovery from breakthrough infection by alpha variant in a fully vaccinated health worker. *Front. Medicine* 1379 (2021).

82. Lopez Bernal, J. *et al.* Effectiveness of covid-19 vaccines against the b.1.617.2 (delta) variant. *New Engl. J. Medicine* **385**, 585–594, DOI: 10.1056/NEJMoa2108891 (2021). https://doi.org/10.1056/NEJMoa2108891.

83. Savulescu, J. Good reasons to vaccinate: mandatory or payment for risk? *J. Med. Ethics* **47**, 78–85 (2021).

84. Nanduri, S. Effectiveness of pfizer-biontech and moderna vaccines in preventing sars-cov-2 infection among nursing home residents before and during widespread circulation of the sars-cov-2 b. 1.617. 2 (delta) variant—national healthcare safety network, march 1–august 1, 2021. *MMWR. Morb. Mortal. Wkly. Rep.* **70** (2021).

85. Callaway, E. & Ledford, H. How to redesign covid vaccines so they protect against variants (2021).

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

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

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

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

90. Voloch, C. M. *et al.* Genomic characterization of a novel sars-cov-2 lineage from rio de janeiro, brazil. *medRxiv* (2020).

91. Wang, H. *et al.* Mutation-specific sars-cov-2 pcr screen: Rapid and accurate detection of variants of concern and the identification of a newly emerging variant with spike l452r mutation. *J. Clin. Microbiol.* JCM–00926 (2021).

92. Tierling, S. *et al.* Rapid base-specific calling of sars-cov-2 variants of concern using combined rt-pcr melting curve screening and sirph technology. *medRxiv* (2021).

93. Boudet, A., Stephan, R., Bravo, S., Sasso, M. & Lavigne, J.-P. Limitation of screening of different variants of sars-cov-2 by rt-pcr. *Diagnostics* **11**, 1241 (2021).

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

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

96. 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.

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

98. Organization, W. H. *et al.* Protocol: real-time rt-pcr assays for the detection of sars-cov-2, institut pasteur, paris. *World Heal. Organ. Geneva. Available via https://www. who. int/docs/default-source/coronaviruse/real-time-rt-pcr-assays-for-the-detection-of-sars-cov-2-institut-pasteur-paris. pdf* (2020).

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

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

101. 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).

102. 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.

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

104. 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).

105. 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. EC and PT made the laboratory testing. CAP and ALR wrote the the article, all authors contributed to editing of the article.

Supplementary Material

Table 2. Number of sequences by lineage downloaded from the GISAID repository on August 11^th, 2021.

Lineage	Number of Sequences
AY.3	17504
B.1	82953
B.1.1	47373
B.1.1.7	1051740
B.1.1.214	18066
B.1.1.519	23376
B.1.2	101442
B.1.160	27353
B.1.177	74475
B.1.177.21	13048
B.1.221	13510
B.1.243	12841
B.1.258	14061
B.1.351	30650
B.1.427	18239
B.1.429	39345
B.1.526	49786
B.1.596	10939
B.1.617.2	366831
D.2	12772
P.1	59692
R.1	10394
Total	2096390