Design of specific primer sets for SARS-CoV-2 variants using evolutionary algorithms

Rincon, A.L.; Romero, C.A.P.; Maldonado, L.M.; Claassen, E.; Garssen, J.; Kraneveld, A.D.; Tonda, A.

published in
GECCO '21
2021

DOI (link to publisher)
10.1145/3449639.3459359

document version
Publisher's PDF, also known as Version of record

Design of Specific Primer Sets for SARS-CoV-2 Variants Using Evolutionary Algorithms

Alejandro Lopez Rincon
a.lopezrincon@uu.nl
Division of Pharmacology,
University of Utrecht
Utrecht, The Netherlands

Carmina A. Perez Romero
Perezr.carmina@gmail.com
Universidad Central de Queretaro
Santiago de Queretaro, Queretaro
México

Lucero Mendoza Maldonado
mendoza.lucero91@gmail.com
Hospital Civil de Guadalajara
"Dr. Juan I. Menchaca"
Guadalajara, Jalisco, México

Eric Claassen
prof.eric.claassen@gmail.com
Athena Institute, Vrije Universiteit
Amsterdam, The Netherlands

Johan Garssen
johan.garssen@danone.com
University of Utrecht
& Danone Nutricia research
Utrecht, The Netherlands

Aletta D. Kraneveld
A.D.Kraneveld@uu.nl
Division of Pharmacology,
University of Utrecht
Utrecht, The Netherlands

Alberto Tonda
alberto.tonda@inrae.fr
UMR 518 MIA-Paris, INRAE,
Université Paris-Saclay
Paris, France

ABSTRACT

Primer sets are short DNA sequences of 18-22 base pairs, that can be used to verify the presence of a virus, and designed to attach to a specific part of a viral DNA. Designing a primer set requires choosing a region of DNA, avoiding the possibility of hybridization to a similar sequence, as well as considering its GC content and T_m (melting temperature). Coronaviruses, such as SARS-CoV-2, have a considerably large genome (around 30 thousand nucleotides) when compared to other viruses. With the rapid rise and spread of SARS-CoV-2 variants, it has become a priority to breach our lack of specific primers available for diagnosis of this new variants. Here, we propose an evolutionary-based approach to primer design, able to rapidly deliver a high-quality primer set for a target sequence of the virus variant. Starting from viral sequences collected from open repositories, the proposed approach is proven able to uncover a specific primer set for the B.1.1.7 SARS-CoV-2 variant. Only recently identified, B.1.1.7 is already considered potentially dangerous, as it presents a considerably higher transmissibility when compared to other variants.

CCS CONCEPTS

• **Applied computing** → **Genomics**; **Bioinformatics**.

GECCO '21, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8350-9/21/07…$15.00
https://doi.org/10.1145/3449639.3459359

KEYWORDS

primers, genetic algorithm, SARS-CoV-2, B.1.1.7

ACM Reference Format:
Alejandro Lopez Rincon, Carmina A. Perez Romero, Lucero Mendoza Maldonado, Eric Claassen, Johan Garssen, Aletta D. Kraneveld, and Alberto Tonda. 2021. Design of Specific Primer Sets for SARS-CoV-2 Variants Using Evolutionary Algorithms. In *2021 Genetic and Evolutionary Computation Conference (GECCO '21), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3449639.3459359

1 INTRODUCTION

As the pandemic of SARS-CoV-2 continues to affects the globe, researchers and public health teams around the world monitor the virus for acquired mutations that may lead to higher risks of developing COVID-19 or vaccine resistance. Prompt and widespread diffusion of information related to viral threats plays a critical role in research and mitigation of outbreaks [35]. This is especially true when viruses mutate rapidly under clinical, therapeutic or vaccine pressure.

Although SARS-CoV-2 acquires 1-2 mutations every month [10, 26], a new variant has been recently reported in the UK as a Variant of Concern (VOC) 202012/01 [7, 14], from the B.1.1.7 lineage [4, 7], Nextstrain clade 20B [18], or clade GR from GISAID (Global Initiative on Sharing All Influenza Data) [31]. The B.1.1.7 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; deletion △69-70, deletion △144, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H [4, 14]. 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 [4, 32] and to be more infectious in mice [17].

The mutation P681H is situated in key residues of the S protein needed for viral transmission and viral entry in lung cells [4, 19].

The rate of transmission of the B.1.1.7 SARS-CoV-2 variant has been estimated to be 56%-70% higher, and its reproductive number (Rt) seems to be up to 0.4 higher [7, 8]. Since it first identification in southern England, the presence of the B.1.1.7 variant has been rapidly increasing in the UK, displacing other SARS-CoV-2 variants [4]. Recently, Public Health England reported that infection with the B.1.1.7 variant can lead to an increase in risk of death of 1.65 when compared to infection with other non-VOC viruses [20]. Furthermore, 11 samples of the B.1.1.7 variant have been found to carry the E484K mutation in the S protein, which has been shown to moderately enhance binding affinity of the ACE2 receptor [32] and to significantly reduce convalescent serum neutralization [15, 16], raising concerns about the current monoclonal/polyclonal antibodies used for treating the COVID-19 disease, as well as the efficacy of new mRNA vaccines in combating the spread of this new SARS-CoV-2 variants [27, 38]. B.1.1.7 and other N501Y-carrying SARS-CoV-2 variants have also been identified in different parts of Europe, South Africa, Brazil, Japan, Australia, USA, and Egypt [3, 11, 18, 31].

Several molecular 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 (qRT-PCR) [2, 28], using *primers*. Primers are short RNA sequences, used to detect the presence of a specific virus in a host organism. Good qRT-PCR primer candidates need to possess several qualities, including having a PCR product size or amplicon between 100 - 200 bps, with a melting temperature (T_m) close to 60°C, a presence of C and G bases representing 40-60% of the sequence, and of a typical length of 18-22 nucleotides.

Public Health England was able to identify the surge of the B.1.1.7 SARS-CoV-2 variant only through the increase in S-gene target failure (negative results) from the otherwise positive target genes (N, ORF1ab) in their three target gene assay [7]. In other words, the primers designed for SARS-CoV-2 are not suited to identify its B.1.1.7 variant, due to fundamental differences in the RNA structure of the two virus strains. In order to detect and contain the potentially dangerous B.1.1.7, new primers need to be developed.

Primer design is traditionally a long process, where domain experts create a canonical sequence for the target virus, analyze the virus' structure, identify promising areas that might be unique to the strain, and create candidates that are then tested first in-silico, and then in the lab. The emerging adoption of machine learning techniques in the health sector, together with the free access to an unprecedented amount of data, has recently led to the first appearance of semi- and fully automated solutions for primer design, based on deep learning techniques [24, 25]. While effective, however, these approaches present several limitations, such as the inability of handling constraints on the candidate primers, and the training time required.

In this paper, we propose a novel, evolutionary-based approach to primer generation. An evolutionary algorithm (EA) [9] is applied to the task of finding meaningful sub-sequences of RNA that could represent good primer candidates. The fitness function is able to include constraints ranging from achieving a melting temperature in a given range, to presenting a sufficient quantity of specific bases. Tested on the B.1.1.7 SARS-CoV-2 variant, the approach proves to be much faster than previous machine learning techniques, delivering results of the same quality.

The rest of the paper is organized as follows: Section 2 provides minimal background information on primers and their design. Section 3 outlines the proposed approach to primer design, based on an evolutionary algorithm. Section 4 summarizes the experimental evaluation, while Section 5 discusses the results and Section 6 concludes the paper.

2 BACKGROUND

This section briefly introduces the notion of primers, and provides an overview of the techniques used to design them.

2.1 Primers

A primer or oligonucleotide is a short nucleic acid sequence that helps to start the DNA synthesis in a Polymerase Chain Reaction (PCR) assay. The presence of primers is necessary because the DNA polymerase can only attach new nucleotides to an existing strand of nucleotides to make copies of a DNA fragment[1]. Primers typically have a length of 18-22 nucleotides and are essential components of the procedure, as they determine the specificity of PCR [36]. In the case of SARS-CoV-2, the viral RNA will be first transformed into single-strand complementary DNA (cDNA), and then the section of interest we will amplified. The forward and reverse primer attach in different directions of the targeted section in the cDNA sequence. If the cDNA sequence does not contain the forward and reverse primer sequences, then they will not attach and consequently won't be amplified, see Fig. 1 for a summary of the procedure.

Primer design is a crucial step of the process, because of its relationship with the success and quality of PCR analyses. Different factors can affect primers' efficiency, including; dimer formation (high self-complementary formation), stem loop interference, GC content (presence of C and G bases in the DNA sequence), high 3' stability (presence of C/G pairs at the 3' end) and extreme melting temperatures. While different computer programs have been proposed to facilitate the design process, it is still necessary to develop new algorithms to select the best PCR primers and avoid errors when using them in the laboratory [23, 29].

2.2 Approaches to primer generation

The traditional approach to primer generation is to first use a canonical sample of the target variant, a summary of all samples available, and then analyze the differences with respect to the original virus. The differences, mainly mutations, are going to be possible locations of interest, and candidate primers can be designed by hand around the mutation. In the case of SARS-CoV-2, all the mutations are referenced to the canonical sequence *NC_*045512.2 [39]. In addition, several parameters need to be verified such as T_m close to 60°, %GC content close to 50%, self complementarity, and high 3' stability. Then, the designed primer has to be checked as a unique sub-sequence, not appearing more than once inside the same sample. This is traditionally carried out with the BLAST software, from the National Center for Biotechnology Information (NCBI) [6]. Nevertheless, if a sequence is not available in BLAST it cannot be

[1] https://www.nature.com/scitable/definition/primer-305/

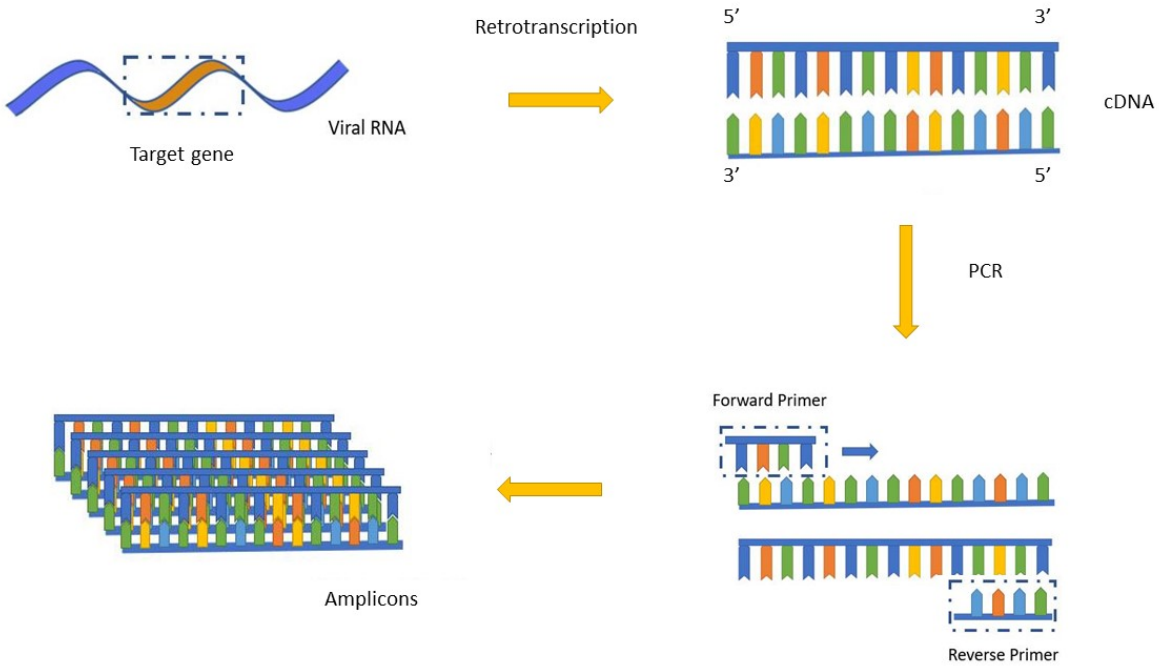

Figure 1: PCR amplification of a section of interest in viral RNA. A primer uniquely identifies a section.

identified using this method, as is the case for many of the new samples recently sequenced from the SARS-CoV-2 virus and its variants.

More recently, deep learning (DL) techniques for primer design have been applied to SARS-CoV-2 [25]. The DL approach is based on a Convolutional Neural Network (CNN) that is trained to separate the target virus' samples from the rest. Then, the filters of the CNN are analyzed to extract sequences of 21 base pairs, that are later tested for the suitability as primer candidates using the software *Primer3Plus* [34]. The same technique has been also applied to finding primers for the B.1.1.7 (British), the B.1.351 (South African), and the P.1 (Brazilian) SARS-CoV-2 variants. Around 16 hours are needed to find a suitable primer for each variant. Furthermore, most of the sequences generated by the DL approach are discarded due to undesirable properties (low presence of GC bases, high/low melting temperature, etc.) [24].

3 PROPOSED APPROACH

In this work, we present a novel evolutionary-based approach to primer generation. As each primer candidate can be seen as a sub-sequence of a SARS-CoV-2 RNA sequence, our aim is to efficiently explore the search space of all possible sub-sequences among all available samples of a target virus strain. The ideal primer candidate would be present in all samples of the target strain, and not present in any other virus, be it a variant of the target, or another virus entirely. Furthermore, the candidate should possess a high number of C and G bases in its genome, a number of N values (indicating sequencing errors) as small as possible, and a temperature of melting T_m as close as possible to 60°.

3.1 Individual representation

An individual in our approach is a primer candidate, thus a sub-sequence of a given virus sample. Assuming we have N samples of the target virus, and each sample is a sequence of L bps, an individual will be described by two integers, sample number k in the training dataset, and position number p inside the sample. The candidate primer will then be represented by the 21 bps in positions $\{p, p+1, ..., p+20\}$ of sample s_k, see Figure 2. While primers can be of any length between 18 and 22 bps, we selected 21 to be able to compare our final results with the primers produced by DL techniques [24, 25], that are of length 21.

3.2 Population initialization

The population is initialized with random individuals. More specifically, the i-th individual will be characterized by two integers, k_i drawn with uniform probability in $(1, N)$ and p_i drawn with uniform probability from $(1, L-21)$.

3.3 Fitness Function

The fitness function is a weighted sum, attempting to take into account all the criteria for a good primer candidate. The first term of the sum, F, is evaluating the presence of the sequence selected as candidate primer I inside training samples labeled with the target

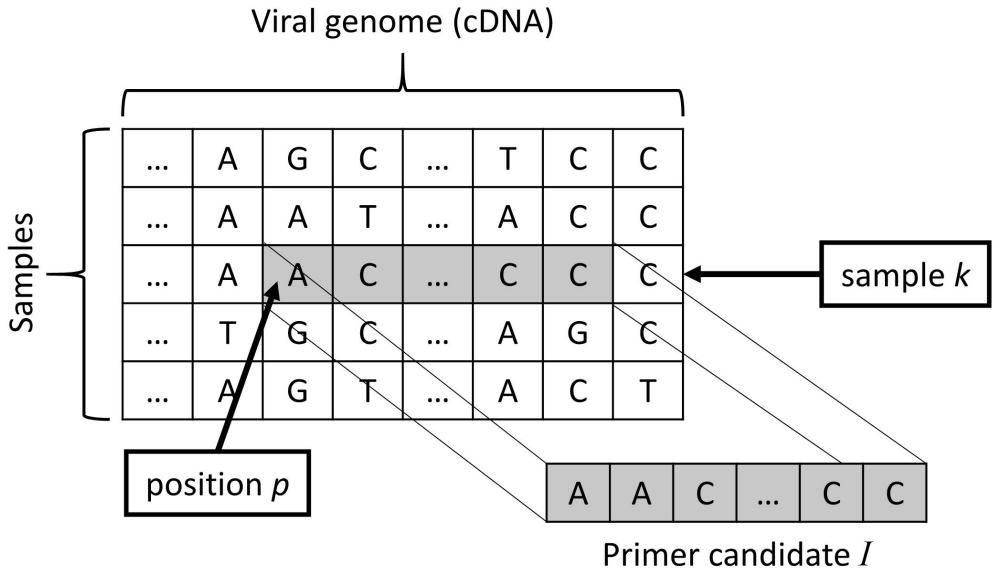

Figure 2: Subsequence I, representing a primer candidate, is uniquely identified by two integer values: index of the sample k in the training set, and position inside its genome p.

virus strain, and its absence from samples with a different label. In formal terms:

$$\mathcal{P}(I) = \sum_{i=0}^{T} P(I, s_i) \tag{1}$$

where T is the number of samples in the training set, s_i is the i-th sample in the training set, and function P is defined as:

$$P(I, s_i) = \begin{cases} 0, & \text{if } I \text{ is found inside } s_i \text{ and } L(s_i) == L(s_k) \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

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 candidate primer I is found inside a sample that does not belong to the target class, or is not found in a sample that belongs to the target class, the solution is penalized.

The second term of the weighted sum takes into account the GC content of the candidate primer, or in other words, the presence of bases G and C:

$$\mathcal{C}(I) = 0.5 - \sum_{i=0}^{21} \frac{C(I(i))}{21} \quad \text{where } C(b) = \begin{cases} 1, & \text{if base } b \text{ is C or G} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

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

$$\mathcal{N}(I) = \sum_{i=0}^{21} N(I(i)) \quad \text{where} \quad N(b) = \begin{cases} 1, & \text{if base } b \text{ is N} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

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 [34] provides an equation to compute T_m for a sequence I:

$$T_m(I) = 81.5 + 16.6 * log_{10}([Na+]) + 41 * C(I) - 600/l(I) \tag{5}$$

where $C(I)$ is the content of C and G bases in sequence I, as described in Equation 3, $[Na+]$ 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 [34], while $l(I) = 21$ by design. The term taking into account T_m will then be:

$$\mathcal{T}(I) = |60 - T_m(I)| \tag{6}$$

To summarize, the final weighted sum will be:

$$\mathcal{F}(I) = w_p \cdot \mathcal{P}(I) + w_c \cdot \mathcal{C}(I) + w_n \cdot \mathcal{N}(I) + w_t \cdot \mathcal{T}(I) \tag{7}$$

with w_p, w_c, w_n, w_t representing the weights associated to each term. Function $\mathcal{F}$ is to be minimized.

3.4 Selection

Individual selection for reproduction is performed through a tournament selection of size τ, where τ individuals are randomly drawn

from the population with uniform probability, their fitness is compared, and the individual with the best fitness is ultimately chosen.

3.5 Evolutionary operators

Mutations operators can act on the two integer values k, representing the index of a sample in the training set, and p, representing a position inside sample k's genome. Mutations acting on p draw from a multinomial distribution where the probabilities are calculated from a normal distribution, following the idea that small displacements of the beginning of a primer left or right in the genome might provide small changes in the fitness function. Mutations targeting k, on the other hand, draw uniformly from all available sample indexes: most genomes of a virus, resulting from a sequencing process, will be almost aligned with each other, with small differences resulting from additions or deletions in the cDNA; for this reason, we believe that the principle of locality will be preserved when creating a primer candidate from position p of two different samples k and k'.

The proposed crossover is a one-point crossover

3.6 Replacement

Individual replacement is carried out through a (μ, λ) scheme, where the entire population is replaced at each generation.

4 EXPERIMENTAL EVALUATION

The proposed approach is validated on real-world data, using samples collected from the GISAID repository. 10,712 SARS-CoV-2 sequences are downloaded on December 23rd, 2020. After removing repeated sequences, we obtain 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 variants are assigned class label 1, while all the remaining samples are labeled as class 0. Then, we divided the data into 8,030 samples for training and 893 for testing.

The algorithm described in Section 3 has been implemented in C#, and is available, along with all the data, in a GitHub open repository[2]. After a few preliminary trials to tune the EA's parameters, population size is set to $\mu = 100$, offspring size to $\lambda = 100$, and a termination condition based on a maximum number of generations, 20, probability of mutation to $p_m = 0.15$ and probability of crossover to $p_c = 0.85$, tournament selection size $\tau = 2$, and a (μ, λ) scheme, where the whole population is replaced at every generation. For the fitness function, the values for the weights are set as $w_p = 1.0, w_c = 100.0, w_n = 1000.0, w_t = 1.0$, to provide the same numerical importance to different parts of the evaluation, with the exception of $\mathcal{N}(I)$, as primers containing N symbols are unlikely to be acceptable. In principle, the fitness function could also be structured as a lexicographic comparison, first evaluating the presence of Ns in the sequence, and then proceeding with the rest of the evaluation. Using a combination of lexicographic acceptability and multi-objective evaluations could also be beneficial. These possibilities will be analyzed in future works.

In order to assess the stability of the results, the proposed approach is run 20 times, with each single run lasting around 62 minutes with 5 threads on a 64-bit Windows 10 laptop with Intel Xeon E-2186M. Thread parallelization is exploited to evaluate several individuals at the same time. The best individual of each run is then stored, to perform a later comparison with primers designed through other methods, discussed in Section 5.

The 20 runs of the proposed approach return different candidate primers, as shown in Table 1. Each of the resulting solutions is then simulated in the canonical version of B.1.1.7 in the GISAID sample *EPI_ISL_*601443 using Primer3Plus [34], to compute the reverse primer and perform a final assessment of its suitability as a primer. This last process is more computationally expensive than a simple evaluation of a fitness function, and it can last up to 20 minutes (Primer3Plus) depending of the software used to crosscheck.

From the 20 best individuals obtained, only 6 solutions can be used as forward primers. Further analysis on the sequences shows that the selected primers are in the region of 2 non-synonymous mutations (S982A, A570D) and 2 synonymous mutations (C15279T, C16176T), with the best candidate being the one based on mutation A570D (Table 2).

We then further validate our results on 487 samples of other coronaviruses, obtained from the National Genomics Data Center (NGDC) [5]. All the candidate primers only target B.1.1.7 SARS-CoV-2, with no appearance in any other coronavirus sample. Further validation on 20,571 samples belonging to other non-corona viruses from the National Center for Biotechnology Information (NCBI) [30], shows no appearance of the sequence in any other virus sample, providing further support for specificity of the obtained best candidate primers.

5 DISCUSSION

While the candidate primers identified by the proposed approach can be considered of high quality, it is worth it to compare them to other primer sets designed by conventional bioinformatics tools. A typical primer set would be represented by a sequence around a specific mutation, believed to be exclusive to the target virus variant.

In order to perform this comparison, we generate 21-bps sequences around mutations N501Y, A570D, D614G, P681H, T716I, S982A, D1118H, with 10 bps before and after the position of the mutation, using the canonical sequence of the B.1.1.7 variant: for example, mutation N501Y (A23063T) will correspond to sequence **CCAACCCACT T ATGGTGTTGG**. We then tested the frequency of appearance of these human-designed primers against the best primers found by the EA during the experimental evaluation. The results are reported in Table 3.

The 6 primer sets generated for the B.1.1.7 variant show almost the same frequency of appearance and similar specificity. Nevertheless, the sequence using mutation A570D is slightly better in frequency of appearance in the training set. The generated forward primers for the B.1.1.7 variant appear in 2,010 of 2,104 sequences, with an average frequency of 95.54%. A further analysis shows that only 2,014 of the sequences labeled as B.1.1.7 present 5 or more of the 7 studied mutations, which could point to an error in the annotation of the variant inside the GISAID dataset, or several extra mutations in the generated 21-bps sequences. If we consider as proper B.1.1.7 variants only sequences that show 5 or more of

[2]The repository address will be disclosed after review, to prevent identification of the authors, as per double-blind peer-review rules.

Table 1: Candidate primers found during 20 runs of the proposed approach, and Primer3Plus output simulated on the canonical reference sequence *EPI_ISL_*601443.

Sequence	Result
GCACGTCTTGACAAAGTTGAG	GAGGTGCTGACTGAGGGAAG
TGGCAGAGACATTGATGACAC	Left primer is unacceptable: High end self complementarity
CAGAGACATTGATGACACTAC	Left primer is unacceptable: Tm too low
GGCAGAGACATTGATGACACT	Left primer is unacceptable: High end self complementarity
CCTCAAGGTATTGGGAACCTG	CATCACAACCTGGAGCATTG
TTGGCAGAGACATTGATGACA	AGCAACAGGGACTTCTGTGC
ACCTCAAGGTATTGGGAACCT	CATCACAACCTGGAGCATTG
CACACAACACATTTGTGTCTG	Left primer is unacceptable: Tm too low/High end self complementarity
TTCAGTGCATCGATATCGGTA	Left primer is unacceptable: High end self complementarity
CTCAGACTAATTCTCATCGGC	Left primer is unacceptable: Tm too low/High 3' stability
TCAGACTAATTCTCATCGGCG	Left primer is unacceptable: High 3' stability
TCAGACTAATTCTCATCGGCG	Left primer is unacceptable: High 3' stability
GGCAGAGACATTGATGACACT	Left primer is unacceptable: High end self complementarity
CAGACTAATTCTCATCGGCGG	Left primer is unacceptable: High 3' stability
GTGATGTAGAAAACCCTCATC	Left primer is unacceptable: Tm too low
GTGATGTAGAAAACCCTCATC	Left primer is unacceptable: Tm too low
CTCAGACTAATTCTCATCGGC	Left primer is unacceptable: Tm too low/High 3' stability
CTCAGACTAATTCTCATCGGC	Left primer is unacceptable: Tm too low/High 3' stability
CTCATCTTATGGGTTGGGATT	GCCACACATGACCATTTCAC
CCTTGCACGTCTTGACAAAGT	GAGGTGCTGACTGAGGGAAG

Table 2: Frequency of appearance in the training and test set, for each of the six candidate primers validated by Primer3Plus.

Candidate Primer	Frequency in test set	Frequency in training set	Mutation
GCACGTCTTGACAAAGTTGAG	0.9922	0.9893	T24506G (S982A)
CCTCAAGGTATTGGGAACCTG	0.9922	0.9889	C16176T
TTGGCAGAGACATTGATGACA	0.9922	0.9895	C23271A (A570D)
ACCTCAAGGTATTGGGAACCT	0.9922	0.9888	C16176T
CTCATCTTATGGGTTGGGATT	0.9910	0.9893	C15279T
CCTTGCACGTCTTGACAAAGT	0.9922	0.9890	T24506G (S982A)

Table 3: Frequency of appearance of the most significant mutations and the forward primer in the 8,293 sequences from the GISAID dataset.

Sequence	B.1.1.7 # samples (%)	Other Variants # samples (%)
N501Y	1,985 (94.34%)	14 (0.02%)
A570D	2,013 (95.67%)	1 (< 0.01%)
D614G	2,096 (99.62%)	5,384 (78.96%)
P681H	2,014 (95.72%)	1 (< 0.01%)
T716I	2,005 (95.29%)	1 (< 0.01%)
S982A	2,008 (95.43%)	0 (0%)
D1118H	2,011 (95.57%)	0 (0%)
GCACGTCTTGACAAAGTTGAG	2,011 (95.57%)	0 (0%)
CCTCAAGGTATTGGGAACCTG	2,008 (95.43%)	0 (0%)
TTGGCAGAGACATTGATGACA	2,014 (95.72%)	1 (< 0.01%)
ACCTCAAGGTATTGGGAACCT	2,007 (95.38%)	0 (0%)
CTCATCTTATGGGTTGGGATT	2,012 (95.62%)	1 (<0.01%)
CCTTGCACGTCTTGACAAAGT	2,009 (95.48%)	0 (0%)
Total Samples	2104	6137

the mutations, the average of the generated primers correctly identifies 2,095 out of 2,104 samples, for a final 99.57% frequency of appearance.

While the primers generated by our approach have a performance similar to primers obtained by conventional PCR design, it is extremely important to remark that this primers can only exist after the canonical sequence of a virus variant is defined. The canonical sequence represents the reference genome of a virus, and a considerable amount of effort from experts is necessary to reach a consensus on this sequence. Our approach, like other ML techniques for primer design [24, 25] has the advantage of automatically finding suitable primers in a matter of hours; but differently from similar ML approaches, it is much faster (about an order of magnitude faster on comparable hardware, from the experiments reported in the previously cited publications) and can handle a wider variety of constraints describing the desired features of a candidate primer.

As of February 5th, other SARS-CoV-2 variants of concern have been identified and are on the rise across the globe, such as the one originating from Brazil (P.1) [12, 13] and the one generated in South Africa (B.1.351) [1, 18, 33]. These new SARS-CoV-2 variants also carry the N501Y and D614G mutations, similarly to the B.1.1.7 variant. Thus, it is important to verify that the primers generated in this work are able to differentiate between the variants.

From the GISAID repository, we downloaded all available 326 sequences of the B.1.351 variant on January 7, 2021 and we downloaded all 28 non-repeated sequences of the P.1 variant on January 19, 2021. Next, we verified the frequency of appearance of the primers in samples from other variants. From Table 4, it is evident that while the primers found with our approach are exclusive to the B.1.1.7 variant, other sequences built around mutations (such as the one built around N501Y) are often also found in different variants. thus negatively impacting their specificity.

Finally, from a comparison between the frequency of appearance of reverse primers in UK samples (Table 5) and the frequency of appearance in other variants (Table 4), we are able to conclude that the best candidates for a primer set are sequences **GCA CGT CTT GAC AAA GTT GAG** as forward primer, based on mutation S982A, and **GAG GTG CTG ACT GAG GGA AG** as reverse primer. While the results are encouraging, it is important to remember that an *in-silico* validation is not enough to provide a final answer. The next step in the process will be to test the primers in the lab.

6 CONCLUSIONS

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 [2]. The arising prevalence of new SARS-CoV-2 variants such as B.1.1.7 has become of great concern, as most RT-PCR test to date will not be able to distinguish these new variants because they where not designed for such a purpose. Therefore, public health officials most rely on their current testing systems and their sequencing results to draw conclusions on the prevalence of new variants in their territories. An example of such cases 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 [7]. 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 on B.1.1.7 [7]. This △69-70 deletion, which affects its N-terminal domain, has been recurrently occurring in different SARS-CoV-2 variants around the world [14, 18, 31] and has been associated with other spike protein receptor binding domain changes [4]. Due to the likeliness of mutation in the S-gene, assays relying solely on its detection are not recommended, and a multiplex approach is required [2, 28, 37]. This is consistent with other existing designs, like CoV2R-3 in the S-gene [22], 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. [21]. Given the concern for the increase in prevalence of the new SARS-CoV-2 B.1.1.7 variant and its possible clinical implication in the ongoing pandemic, diagnosing and monitoring the prevalence of such variant in the general population will be of critical importance to help fight its spread and develop new policies.

In this work, we propose an evolutionary approach to the design of primer sets. The approach is tested on the identification of the B.1.1.7 SARS-CoV-2 variant. The primer set developed using our approach proves to perform on par with primers found through other methods, and can be obtained in a reduced amount of time. We believe that the primer set can be used 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 variant. In this way, health authorities can better evaluate the medical outcome of this patients, and adapt or inform new policies that can help curve the rise of infections by variants of interest. Although the primer sets delivered by our automated methodology will still require laboratory testing to be validated, our approach 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.

REFERENCES

[1] 2020. SARS-CoV-2 Variants. https://www.who.int/csr/don/31-december-2020-sars-cov2-variants/en

[2] Adeel Afzal. 2020. Molecular diagnostic technologies for COVID-19: Limitations and challenges. *Journal of advanced research* (2020).

[3] Erik Alm, Eeva K Broberg, Thomas Connor, Emma B Hodcroft, Andrey B Komissarov, Sebastian Maurer-Stroh, Angeliki Melidou, Richard A Neher, Áine O'Toole, Dmitriy Pereyaslov, et al. 2020. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Eurosurveillance* 25, 32 (2020), 2001410.

[4] Arambaut, Garmstrong, and Isabel. 2020. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563

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

[6] Grzegorz M Boratyn, Christiam Camacho, Peter S Cooper, George Coulouris, Amelia Fong, Ning Ma, Thomas L Madden, Wayne T Matten, Scott D McGinnis, Yuri Merezhuk, et al. 2013. BLAST: a more efficient report with usability improvements. *Nucleic acids research* 41, W1 (2013), W29–W33.

[7] Meera Chand, Susan Hopkins, and Gavin Dabrera. 2020. Investigation of novel SARS-COV-2 variant: Variant of Concern 202012/01.

Table 4: Frequency of appearance of the characteristic mutations for the UK (B.1.1.7), South African (B.1.351), and Brazilian (P.1) variants in B.1.1.7 sequences (2,104), B.1.351 sequences (337), P.1 sequences (28) and sequences of other variants (6,808).

Sequence	Other	B.1.1.7	B.1351	P.1
T1001I	0.01%	95.53%	0.00%	0.00%
A1708D	0.00%	91.83%	0.00%	0.00%
I2230T	0.01%	94.39%	0.00%	0.00%
N501Y	0.04%	94.34%	99.70%	82.14%
A570D	0.01%	95.67%	0.00%	0.00%
P681H	0.01%	95.72%	0.30%	0.00%
T716I	0.01%	95.29%	0.00%	0.00%
S982A	0.00%	95.44%	0.00%	0.00%
D1118H	0.00%	95.58%	0.00%	0.00%
Q27stop	0.04%	90.64%	0.30%	0.00%
R52I	0.00%	90.64%	0.00%	0.00%
Y73C	0.00%	90.78%	0.00%	0.00%
S235F	0.03%	95.58%	0.00%	0.00%
GCACGTCTTGACAAAGTTGAG	0.00%	95.58%	0.00%	0.00%
CCTCAAGGTATTGGGAACCTG	0.00%	95.44%	0.00%	0.00%
TTGGCAGAGACATTGATGACA	0.01%	95.72%	0.00%	0.00%
ACCTCAAGGTATTGGGAACCT	0.00%	95.39%	0.00%	0.00%
CTCATCTTATGGGTTGGGATT	0.03%	95.63%	0.00%	0.00%
CCTTGCACGTCTTGACAAAGT	0.00%	95.48%	0.00%	0.00%

Table 5: Frequency of appearance of the forward and reverse primers found by the EA algorithm.

Forward Primer	Frequency	Reverse Primer	Frequency	Average
GCACGTCTTGACAAAGTTGAG	95.58%	GAGGTGCTGACTGAGGGAAG	99.71%	97.65%
CCTCAAGGTATTGGGAACCTG	95.44%	CATCACAACCTGGAGCATTG	98.62%	97.03%
TTGGCAGAGACATTGATGACA	95.72%	AGCAACAGGGACTTCTGTGC	99.62%	97.67%
ACCTCAAGGTATTGGGAACCT	95.39%	CATCACAACCTGGAGCATTG	98.62%	97.01%
CTCATCTTATGGGTTGGGATT	95.63%	GCCACACATGACCATTTCAC	99.81%	97.72%
CCTTGCACGTCTTGACAAAGT	95.48%	GAGGTGCTGACTGAGGGAAG	99.71%	97.60%

[8] Nicholas Davies and et al. 2020. Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. https://cmmid.github.io/topics/covid19/uk-novel-variant.html. *Available in Github* (2020). Online; accessed 26 December 2020.

[9] Kenneth De Jong. 2016. *Evolutionary Computation: A Unified Approach.* Bradford Books, Cambridge, Massachusetts.

[10] Sebastian Duchene, Leo Featherstone, Melina Haritopoulou-Sinanidou, Andrew Rambaut, Philippe Lemey, and Guy Baele. 2020. Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus evolution* 6, 2 (2020), veaa061.

[11] H Elghazaly, S El-Nakeep, A Moustafa, MH El-Sayed, H Hafez, M Matboli, M Saad, AM Elsamie, FSE Ebeid, S Makkeyah, et al. 2020. 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).

[12] Nuno Faria. 2021. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586?fbclid=IwAR3Xe13JtwUoWrHAl82ekepQAvvvKK1kGYixmjbkq6_YEkk1fzsdDYXF6hM

[13] Centre for Disease and Control Prevention. 2021. Emergence of SARS-CoV-2 B.1.1.7 Lineage — United States, December 29, 2020–January 12, 2021. https://www.cdc.gov/mmwr/volumes/70/wr/mm7003e2.htm?s_cid=mm7003e2_e

[14] European Centre for Disease Prevention and Control. 2020. Rapid increase of a SARS-CoV-2 variant with multiple spike protein mutations observed in the United Kingdom. https://www.ecdc.europa.eu/sites/default/files/documents/SARS-CoV-2-variant-multiple-spike-protein-mutations-United-Kingdom.pdf

[15] Allison J Greaney, Andrea N Loes, Katharine HD Crawford, Tyler N Starr, Keara D Malone, Helen Y Chu, and Jesse D Bloom. 2020. Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. *bioRxiv* (2020), 2020–12.

[16] Allison J Greaney, Tyler N Starr, Pavlo Gilchuk, Seth J Zost, Elad Binshtein, Andrea N Loes, Sarah K Hilton, John Huddleston, Rachel Eguia, Katharine HD Crawford, et al. 2020. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell host & microbe* (2020).

[17] Hongjing Gu, Qi Chen, Guan Yang, Lei He, Hang Fan, Yong-Qiang Deng, Yanxiao Wang, Yue Teng, Zhongpeng Zhao, Yujun Cui, et al. 2020. Adaptation of SARS-CoV-2 in BALB/c mice for testing vaccine efficacy. *Science* 369, 6511 (2020).

[18] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 23 (2018), 4121–4123.

[19] Markus Hoffmann, Hannah Kleine-Weber, and Stefan Pöhlmann. 2020. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. *Molecular Cell* (2020).

[20] Peter Horby, Catherine Huntley, Nick Davies, John Edmunds, Neil Ferguson, Graham Medley, Andrew Hayward, Muge Cevik, and Calum Semple. 2021. NERVTAG note on B.1.1.7 severity.

[21] Michael Kidd, Alex Richter, Angus Best, Jeremy Mirza, Benita Percival, Megan Mayhew, Oliver Megram, Fiona Ashford, Thomas White, Emma Moles-Garcia, Liam Crawford, Andrew Bosworth, Tim Plant, and Alan McNally. 2020. S-variant SARS-CoV-2 is associated with significantly higher viral loads in samples tested by ThermoFisher TaqPath RT-QPCR. *medRxiv* (2020). https://doi.org/10.1101/2020.12.24.20248834

[22] Sinae Kim, Jong Ho Lee, Siyoung Lee, Saerok Shim, Tam T Nguyen, Jihyeong Hwang, Heijun Kim, Yeo-Ok Choi, Jaewoo Hong, Suyoung Bae, et al. 2020. The

Progression of SARS Coronavirus 2 (SARS-CoV2): Mutation in the Receptor Binding Domain of Spike Gene. *Immune Network* 20, 5 (2020).

[23] Kelvin Li and Anushka Brownley. 2010. Primer design for RT-PCR. In *RT-PCR Protocols*. Springer, 271–299.

[24] Alejandro Lopez-Rincon, Carmina A. Perez-Romero, Alberto Tonda, Lucero Mendoza-Maldonado, Eric Claassen, Johan Garssen, and Aletta D. Kraneveld. 2020. Design of Specific Primer Set for Detection of B.1.1.7 SARS-CoV-2 Variant using Deep Learning. (Dec. 2020). https://doi.org/10.1101/2020.12.29.424715

[25] Alejandro Lopez-Rincon, Alberto Tonda, Lucero Mendoza-Maldonado, Daphne G. J. C. Mulders, Richard Molenkamp, Carmina A. Perez-Romero, Eric Claassen, Johan Garssen, and Aletta D. Kraneveld. 2021. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. *Scientific Reports* 11, 1 (Jan. 2021). https://doi.org/10.1038/s41598-020-80363-5

[26] Roujian Lu, Xiang Zhao, Juan Li, Peihua Niu, Bo Yang, Honglong Wu, Wenling Wang, Hao Song, Baoying Huang, Na Zhu, et al. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* 395, 10224 (2020), 565–574.

[27] Gard Nelson, Oleksandr Buzko, Patricia Spilman, Kayvan Niazi, Shahrooz Rabizadeh, and Patrick Soon-Shiong. 2021. 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* (2021). https://doi.org/10.1101/2021.01.13.426558

[28] World Health Organization et al. 2020. Molecular assays to diagnose COVID-19: summary table of available protocols.

[29] Alicia Rodríguez, Mar Rodríguez, Juan J Córdoba, and María J Andrade. 2015. Design of primers and probes for quantitative real-time PCR methods. In *PCR Primer Design*. Springer, 31–56.

[30] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 29, 1 (2001), 308–311.

[31] Yuelong Shu and John McCauley. 2017. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* 22, 13 (2017).

[32] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al. 2020. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* 182, 5 (2020), 1295–1310.

[33] Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, Sureshnee Pillay, Emmanuel James San, Nokukhanya Msomi, et al. 2020. 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).

[34] Andreas Untergasser, Harm Nijveen, Xiangyu Rao, Ton Bisseling, René Geurts, and Jack AM Leunissen. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic acids research* 35, suppl_2 (2007), W71–W74.

[35] Mark B van der Waal, Carolina dos S Ribeiro, Moses Ma, George B Haringhuizen, Eric Claassen, and Linda HM van de Burgwal. 2020. Blockchain-facilitated sharing to advance outbreak R&D. *Science* 368, 6492 (2020), 719–721.

[36] Elizabeth van Pelt-Verkuil, Alex van Belkum, and John P. Hays. 2008. PCR Primers. In *Principles and Technical Aspects of PCR Amplification*. Springer Netherlands, 63–90. https://doi.org/10.1007/978-1-4020-6241-4_5

[37] Rui Wang, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. 2020. Mutations on COVID-19 diagnostic targets. *arXiv preprint arXiv:2005.02188* (2020).

[38] Zijun Wang, Fabian Schmidt, Yiska Weisblum, Frauke Muecksch, Christopher O Barnes, Shlomo Finkin, Dennis Schaefer-Babajew, Melissa Cipolla, Christian Gaebler, Jenna A Lieberman, Zhi Yang, Morgan E Abernathy, Kathryn E Huey-Tubman, Arlene Hurley, Martina Turroja, Kamille A West, Kristie Gordon, Katrina G Millard, Victor Ramos, Justin Da Silva, Jianliang Xu, Robert A Colbert, Roshni Patel, Juan P Dizon, Cecille Unson-O'Brien, Irina Shimeliovich, Anna Gazumyan, Marina Caskey, Pamela J Bjorkman, Rafael Casellas, Theodora Hatziioannou, Paul D Bieniasz, and Michel C Nussenzweig. 2021. mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. *bioRxiv* (2021). https://doi.org/10.1101/2021.01.15.426911

[39] Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, et al. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579, 7798 (2020), 265–269.