PCR primer diagnostic kits design for Covid-19 (Alpha- Delta- Omicron) in Africa

Abstract

The diversity of SARS-CoV-2 variants of concern ran unprecedently high day after day owing to (S) Spike protein gene rapid mutations which cause escape from SARs-CoV-2 neutralizing antibodies. Thereby, the infection rate of Omicron in Africa is higher than that of Delta and Alpha variants, respectively. Thus, numerous companies are continuously striving with almost diligence to set out RT-PCR specific primers for emerging variants during the past months, mainly applying diverse algorithms for MSA programs besides evolutionary algorithms to detect individual variants efficiently. The following study is targeting conserved regions of VOCs, where there is less evolutionary freedom in such regions as mutations would result in a partial or complete virus-replication-defect. Phylogenetic tree construction to infer links between reference and descendent sequences and nodes in the tree represent sequence haplotypes that match one or more sequences in the data. In the course of proposed in-silico PCR assays, we used the Muscle tool on selected Alpha, Delta, and Omicron sequences in Africa, implemented the matplotlib and phyllo packages in python script to extract conserved regions among them, and construct our phylogenetic tree, Moreover, producing conserved sequences for primer Blast on NCBI to predict specific primers for the selection of the best primer based on GC% content criteria, Self-complementarity within Ref seq RNA database. Eventually, primer validation using PCR primer stats and UCSC In-Silico PCR.

Introduction

Coronaviruses cause respiratory infections. They infect humans, mammals, and other species such as livestock and are not only a public health challenge but also an economic problem. The first member is the Middle East respiratory syndrome coronavirus (MERS-COV) first reported in Saudi Arabia in 2012 [(1)] the second one is (SARS) was first reported in 2003 in Asia. [(2)] And we are working on our research on the third novel of coronavirus (SARS-COV-2).symptoms of this disease are Fever, Cough and Congestion, Shortness of breath, body aches, and loss of taste or smell. It is a zoonotic disease, discovered in Wuhan China in December 2019 and was declared a pandemic by the World Health Organization (WHO) on March 11, 2020, spreading to more than 200 countries [(3)] there have been at least 11,122,258 reported Infections with 10,325,656 recoveries and 245,070 reported deaths caused by COVID-19 in Africa till now [(4)] we are specifically talking about the variants of Covid-19 variants which have become the most concern issue in the scientific community nowadays. The Alpha variant with lineage (B.1.1.7) of SARS-CoV-2 was the first identified major (VOC) variant of concern First document in the United Kingdom in November 2020[(5)] and the first confirmed case in Africa recorded in The Gambia on 14 January 2021 [(6)] this variant spreads approximately 50% better than the original one. Delta variant is known as B.1.617.2 First document in India in late 2020 The Delta variant is more transmissible around 40% to 60% more than the Alpha variant according to The studies. Its spread very fast because of its capability to attack the host's immune system compared to the original strain [(7)]. The Omicron variant (B.1.1.529) first documented in South Africa in November 2021 Omicron has several mutations that might show how quickly it spreads or how to damage it causes [(8)]. Omicron

Pango lineage B.1.1.529 and it has sub Pango lineages BA.1. 1, BA.1, BA.2, and BA.3. BA. 1 has the majority of Omicron sequences till now and it's the most well-traveled member of this Variant; it spread very fast. [5] .in our study we aim to design PCR primer diagnostic kits for exclusive parts in the three variants of Covid-19 (Alpha, Delta, Omicron) using a fully automated script. We used pango lineage (B.1.1.7) to present the alpha variant and for the delta, we used (B.1.617.2) and finally, for the Omicron variant, we used (BA. 1).

Methods

- **Data collection:**

First, we identified each Covid-19 Pango lineage searching for NCBI virus much easier which is shown in figure [2], and used the Pango lineages (BA.1, B.1.617.2, B.1.1.7) representing Omicron, Delta, and Alpha.

Table 1: SARS-CoV-2 Variants of Concern (VOCs) and Variants of Interest (VOIs) [4]

WHO label	Pango lineage	GISAID clade	Nextstrain clade	Earliest documented samples	Date of designation
Variants of Concern (VOCs)					
Alpha	B.1.1.7	GRY (formerly GR/501Y.V1)	20I/501Y.V1	United Kingdom, Sep-2020	18-Dec-2020
Beta	B.1.351	GH/501Y.V2	20H/501Y.V2	South Africa, May-2020	18-Dec-2020
Gamma	P.1	GR/501Y.V3	20J/501Y.V3	Brazil, Nov-2020	11-Jan-2021
Delta	B.1.617.2	G/452R.V3	21A/S:478K	India, Oct-2020	VOI: 4-Apr-2021 VOC: 11-May-2021

Figure 1. Shows the Pango linage for different variants.
Vineet Sharma, Himanshu Rai, Dev N. S. Gautam, Pradeep K. Prajapati, Rohit Sharma, Emerging evidence on Omicron (B.1.1.529) SARS-CoV-2 variant, Journal of Medical Virology, 10.1002/jmv.27626, (2022).

After carefully choosing closer dates of our sequences avoiding any bias in the study, due to the lack of suitable computational power, we only downloaded 200 whole-genome sequences ID for every variant from Africa and inserted them into our bash script to begin downloading the sequences with the path of the desired folder to save our results using the flag arguments. Also, we made a new file for the sequences changing the header to only accession ID making the header easier to observe than deleting the old File. The code is shown in figure [2]. In most of the following steps, we will be using our automated bash and python script to get our results.

```
efetch -db nucleotide -format fasta -id $ID >$file_path/sequence.fasta
cut -d ' ' -f1 $file_path/sequence.fasta > $file_path/Sequence.fasta
rm $file_path/sequence.fasta
```

Figure 2. Shows the bash script used to download the sequences and then removes the header keeping only accession ID in a new file followed by removing the old file.

- **Multiple sequence alignment:**

Multiple alignments were performed using the muscle command line after installing it on our machine. This was done on every variant's sequences to get three different alignments and three different tree files one for each variant. The results were shown in clustalW format.

```
muscle -in $file_path/Sequence.fasta -out $file_path/Alignment.fasta \
 -clw -tree1 $file_path/Tree.phy
```

Figure 3. Shows the bash script used to do the MSA and the Tree file.

- **Conserved region extraction:**

We started extracting our conserved region using the integrated python functionality in the bash script with a python script as an input which produced two files one has all the conserved regions, the other has the longest conserved region which we will use in our primer design. This was done automatically by the script using the NumPy and Biopython packages.

```
for SeqRecord in AlignIO.read(filename, 'clustal'):
    A.append(SeqRecord.seq)
profile = np.array(A)
difference = True
for x in range(len(A[0])):
    if '-' in profile[:, x]:
        difference = True
    if len(set(profile[:, x])) == 1:
        difference = False
        y.append(x)
    if len(set(profile[:, x])) != 1:
        difference = True
    if difference or x == (len(A[0]) - 1):
        if len(y) == 1:
            Conserved_region_txt_open.write('>Conserved_Nucleotide(index=%s) \n%s \n\n' % (y[0], A[0][y[0]]))
            y.clear()
        if len(y) == 0:
            continue
        if len(y) != 1:
            New_seq=textwrap.fill("".join(A[0][y[0]:(y[-1] + 1)]),70)
            Conserved_region_txt_open.write(
                '>Conserved_region(index=%s:%s)_length=%s \n%s \n\n' % (y[0], y[-1],
                                                                       (y[-1] - y[0] + 1), New_seq))
            y.clear()
```

Figure 4. Shows part of the python code used to get conserved which will also display the index range and length of each conserved region.

- **Drawing phylogenetic tree:**

By using “phylo” and “matplotlib” packages in python we generated the phylogenetic tree showing the relationship between all the sequences with branch length displayed and saved it.

```
tree_file = ("%s/Delta_tree.phy" % filepath)
tree = Phylo.read(tree_file, "newick")
fig_2 = plt.figure(figsize=(10, 20), dpi=100)
fig_2.suptitle('The Phylogenetic Tree',
               fontsize=20)
axes = fig_2.add_subplot(1, 1, 1)
Phylo.draw(tree, axes=axes,branch_labels=lambda c: c.branch_length,do_show=False)
fig_2.savefig("%s/The_Phylogenetic_Tree"% filepath,figsize=(10, 20))
```

Figure 5. Shows the python script used to save the phylogenetic tree picture on the specified folder at the start.

- **Functional products/interpretations:**

After obtaining the longest conserved region for each alignment data our script used this data to get all open reading frames and then we accessed the Pfam website to find if there were any corresponding functional products in the conserved region to know its properties. Also in our script, we added multiple parameters that can be modified using flag arguments making it more flexible and easier for the user.

```
getorf -sequence $file_path/Longest_conserved.fasta -outseq $file_path/ORFs_ou\
t.txt \-table $table -minsize $minsize -find 3
```

Figure 6. Shows the bash script used to get the ORFs from the longest conserved region also shows the flag parameters we can change.

- **Primer design:**

Using Primer-BLAST we imported every longest conserved region for every variant on it to get a suitable primer for each variant Taking into consideration that the temperature difference must not be more than three and the database was Refseq RNA.

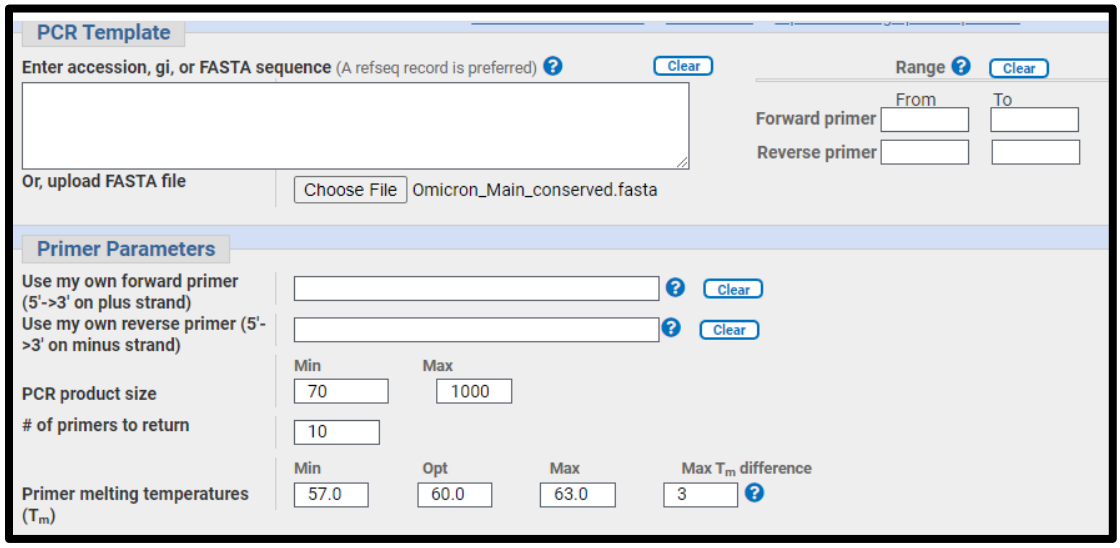

Figure 7. Shows in the usage of omicron main conserved region (Longest) in primer blast against the SAR-2(Taxid: 2697049) Refseq RNA database.

- **Primer Validation:**

After the selection of the suitable primer according to GC content, melting temperature, and self-complementarity we validate0d our results on UCSC In-Silico PCR and PCR primer stats.

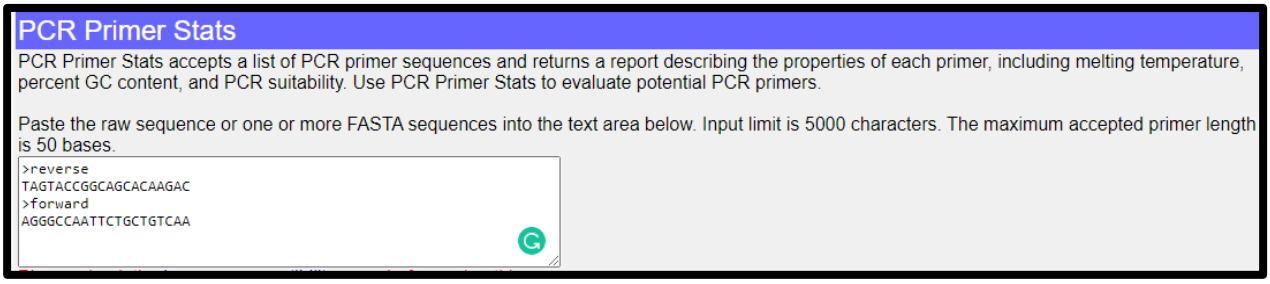

Figure 8. Shows the PCR primer stats in the validation of omicron prime.

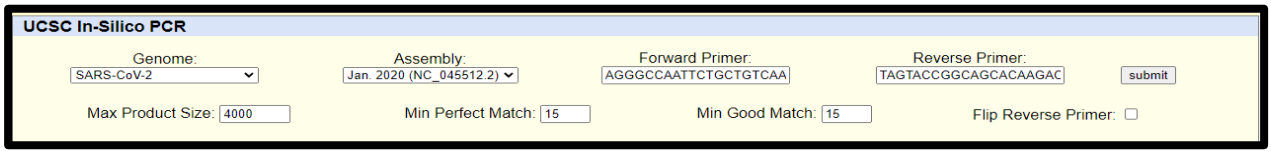

Figure 9. Shows the UCSC In-Silico PCR in the validation of omicron primer.

- **Universal Primer:**

We proceeded with building a universal primer that can detect any Alpha, Delta, and Omicron variant, but due to our low computational power we used 70 sequences from each variant into a single FASTA file, then we began doing the multiple sequence alignment followed by constructing the phylogenetic tree next obtaining the conserved regions and its ORFs. From that point, we moved to form our universal primer using Primer-BLAST with the same parameters as the rest of the primers and testing it on UCSC in-Silico PCR and PCR primer stats.

Results:

As we have so much data to be analyzed and primer design needs so much accuracy , there is no avoiding a bit of good algorithmic design to achieve automation which empowers us to get most accurate results in the lowest possible time The bulk of our work is dependent on good alignment program by which we can accomplish both fast and accurate results Depending on benchmark study we found that muscle program is the best to use [9], though We identified highly conserved residues, which were then extracted using a Python script. Primer-BLAST Primers for finding templates for major conserved regions using the RefSeq RNA database allowed Primer-BLAST to better identify the template and thus perform better primer specificity checking. As three templates are shorter than 50,000 bp there is no need to identify primer range by using the default parameters [10]. The primers are shown in figures (11, 12, 13)

Figure 10. Shows part of the Omicron alignment in clustalW format.

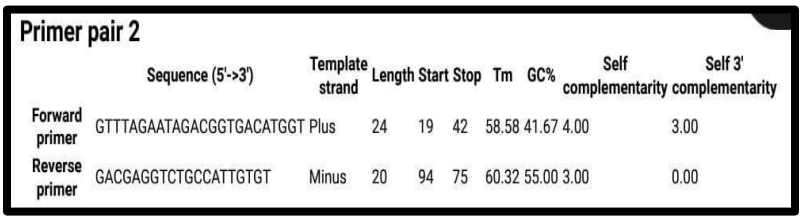

Figure 11. Shows primer of Omicron variant.

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AGGGCCAATTCTGCTGTCAA	Plus	20	540	559	59.89	50.00	4.00	1.00
Reverse primer	TAGTACCGGCAGCACAAGAC	Minus	20	621	602	59.75	55.00	4.00	1.00

Figure 11. Shows primer of Delta variant.

Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	ACACAATGGCAGACCTCGTC	Plus	20	180	199	60.32	55.00	3.00	3.00
Reverse primer	CAGCATTTCGCATGGCATCA	Minus	20	413	394	59.90	50.00	5.00	1.00

Figure 12. Shows primer of Alpha variant.

Then primers validation was done via primer stat shown in figure (13). All primers passed the PCR suitability test. The three pairs fulfilled the criteria for good primers Length of 18-24 bases, 40-60% G/C content, melting temperature (Tm) of 50-60°C, Tm difference within 5°C of each other, and no complementary region [10].

```
PCR suitability tests (Pass / Warning):
-----------------------------------
               Single base runs: Pass
         Dinucleotide base runs: Pass
                        Length: Pass
                    Percent GC: Pass
          Tm (Nearest neighbor): Warning: Tm is greater than 58;
                      GC clamp: Pass
                 Self-annealing: Pass
               Hairpin formation: Pass
-------------------------------------------------------------
```

Figure 13. Shows PCR primer stats result of forward primer strand of Omicron.

We constructed a Phylogenetic tree with Itol web tool, the program Convert the Newick data produced from the muscle alignment into a branching diagram ("tree") that shows the relationships between the sequences. Branch length represents the evolutionary time between two nodal lines.

Alpha: https://itol.embl.de/tree/1562054415647001164579919 3

Delta: https://itol.embl.de/tree/1562175015724701645209153

Omicron: https://itol.embl.de/tree/1562171381691997816457932 74

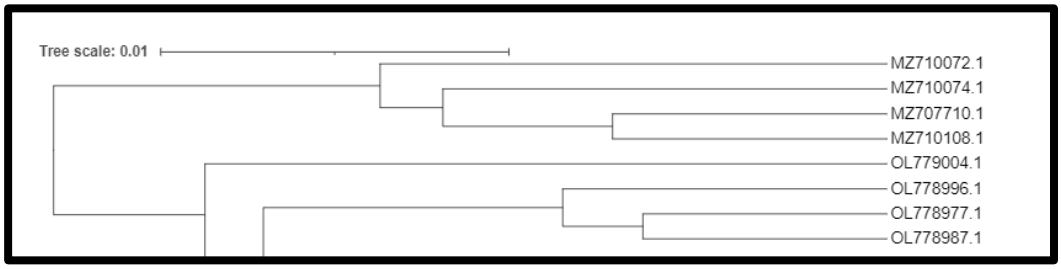

Figure 14. Shows part of the phylogenetic tree of the Delta variant.

It was found that the SARS-CoV-2 virus has high mutations rate in Spike protein Binds to host cell receptor, mediates viral and host membrane attachment and fusion.ORF2 (21,563–25,384) [11].The Pfam showed that the extracted ORF from the conserved region of Alpha and beta variant both had the protein Coronavirus RNA-dependent RNA polymerase, N-terminal which is responsible for the replication and transcription of the viral genome, and in omicron, the protein Coronavirus replicase NSP8 play a role in the stabilization of regions involved in RNA binding and are essential for a highly active polymerase complex.

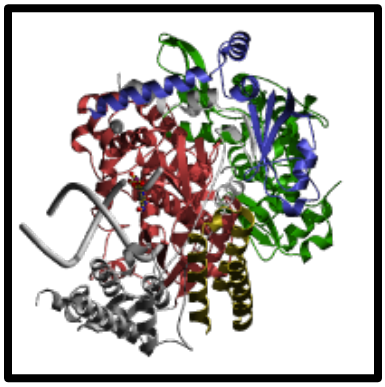

Figure 14. 3D structure of RNA polymerase, N-terminal

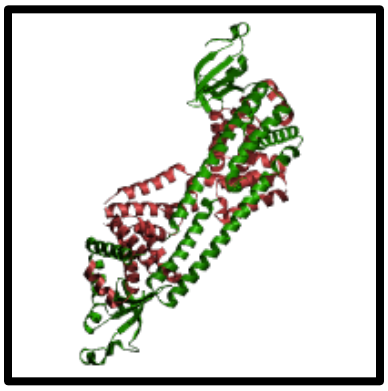

Figure 14. 3D structure of Coronavirus replicase NSP8

Universal Primer:

Using only 70 sequences from each variant we believe our produced primer can detect any Alpha, Beta, or Omicron strain. And upon validation, our primer passed all the PCR primer stats tests and UCSC in-Silico PCR.

Primer pair 9

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GCATGGATTGGCTTCGATGT	Plus	20	25	44	58.97	50.00	4.00	2.00
Reverse primer	AGGCGGTGGTTTAGCACT	Minus	18	198	181	58.84	55.56	4.00	2.00
Product length	174								

Figure 15. Shows the universal Primer in Primer-BLAST results.

Conclusion

It is extremely challenging PCR primers design for highly variable DNA sequences of SARS-CoV-2 variants as these are considered the most genetically variable viruses ever known nowadays. Deciphering region of consensus or interest is the gold key in primer specificity for each of Alpha, delta, and Omicron variants. The constraints of formulating sites to maximize population coverage, matching of melting temperatures in primer pair to drive to experimental conditions, and many other parameters. Whilst employing several developing algorithms to achieve optimal detection threshold, the extracted primers get homologous with SARs-CoV-2 genes E, N, Orf1a/1b, and *RdRp*. However, attempting to specify primer pairs for Alpha, Delta and Omicron haven't ever been experimented with previously in published research. Recently, allele-specific PCR and multiplex RT-PCR assays have extensively been used in many countries although false positive and negative results have been confirmed among variants as these three simultaneous variants failed to be tested by these kits. Ultimately in-vitro experimental trials with the primers gotten out of our study are pending adoption.

From the sequences selected from NCBI for each variant mentioned previously. Multiple sequence alignment algorithm proceeds in 3 stages. First draft progressive, improved progressive & refinement. In the first step, draft multiple alignments, emphasizing speed. The second stage uses the Kimura distance for the estimation of the binary tree for the alignment. Finally, refining the alignment made in the second step. In the first two stages, time complexity and space complexity. The last stage adds another term to time complexity. Simply speaking, aligned sequence pair with computed pairwise identity and conversion to additive distance estimate, applying Kimura correction for multiple substitutions at a single site. Distance matrices clustered by Unweighted Group Method with Arithmetic Mean, which needs distance matrix of the analyzed taxa calculated from multiple alignments. Or adopting a Neighbor-joining that will give a better estimate of the evolutionary tree.

Furthermore, exploiting NCBI primer blast results, that precisely depend on efficient manipulation of primers parameters to get acceptable primers ever.

Moreover, no other tools are more accurate and sophisticated in handing primers parameters as the one mentioned above. The impact of mutations on provided primers specificity in PCR is revealed by

the difference of cycle threshold (Ct) value with or without mismatch between primer and template is at minimum meaning that mismatch moves away from 3' end of the primers. When a mismatch occurred in the middle of the primer, ΔCt is within ±1, indicating PCR does not, and these measures are determined by every unique PCR kit according to manufacturing conditions.

In the SARS-CoV-2 genome, there are 2 ORFs, ORF1a & ORF1ab consisting of 23 of the genomic map which is translated into 2 polyproteins, PP1a (NSP1-NSP11) & PP1ab (NSP1-NSP16). Between 5' UTR and 3' UTRs, there exist several Non-Structural Proteins (NSPs) at the 5' end and a few structural proteins at the 3' end of the genome as envelope protein, spike glycoprotein, nucleocapsid, and membrane proteins. Relevant to conserved regions translation via Pfam into specific amino acid sequence findings, we turned out to find that Alpha and Delta have RNA dependent RNA polymerase (RdRp) in their conserved region, and Omicron has a few more gene targets in the conserved region including NSP8 & NSP9, which preliminarily function for primase activity and viral RNA replication respectively. As far, BA.2 subvariant, which lacks 69-70 deletion in S protein, dominates in recent weeks in South Africa, that may render PCR test false negative as depending on currently SGTF marker (3 targets (S), (NP) & (ORF1ab), absence of spike protein and detecting nucleocapsid and ORF1ab genes). Moreover, increased discrepancies in cycle threshold value (Ct) between different gene targets & failure to detect specific gene targets such as those containing sequences that coincide with documented mutations with false-negative results.

By multiple in-silico trials in primer-Blast for conventional and RT-PCR, obtaining primers and subsequent primers' validation, optimal primers for experimental laboratory are in hand to release into diagnostic kits.

References:

1. about MERS. (2022). Retrieved 9 February 2022, from https://www.cdc.gov/coronavirus/mers/about/index.html
2. about MERS. (2022). Retrieved 9 February 2022, from https://www.cdc.gov/coronavirus/mers/about/index.html
3. Coronavirus disease (COVID-19) – World Health Organization. (2022). Retrieved 15 February 2022, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019
4. Coronavirus (COVID-19). (2022). Retrieved 18 February 2022, from https://www.afro.who.int/health-topics/coronavirus-covid-19
5. Cov-Lineages. (2022). Retrieved 15 February 2022, from https://cov-lineages.org/global_report_B.1.1.7.html
6. (2022). Retrieved 18 February 2022, from https://www.reuters.com/business/healthcare-pharmaceuticals/gambia-records-first-two-cases-uk-covid-19-variant-2021-01-14/
7. Tracking SARS-CoV-2 variants. (2022). Retrieved 9 February 2022, from https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/
8.Shiehzadegan S, Alaghemand N, Fox M, Venketaraman V. Analysis of the Delta Variant

B.1.617.2 COVID-19. Clinics and Practice. 2021; 11(4):778-784. https://doi.org/10.3390/clinpract11040093
9. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113 (2004). https://doi.org/10.1186/1471-2105-5-113
10. Addgene: Protocol - How to Design Primers. (2022). Retrieved 25 February 2022, from https://www.addgene.org/protocols/primer-design/
11. Bai, C., Zhong, Q. & Gao, G.F. Overview of SARS-CoV-2 genome-encoded proteins. Sci. China Life Sci. 65, 280–294 (2022). https://doi.org/10.1007/s11427-021-1964-4