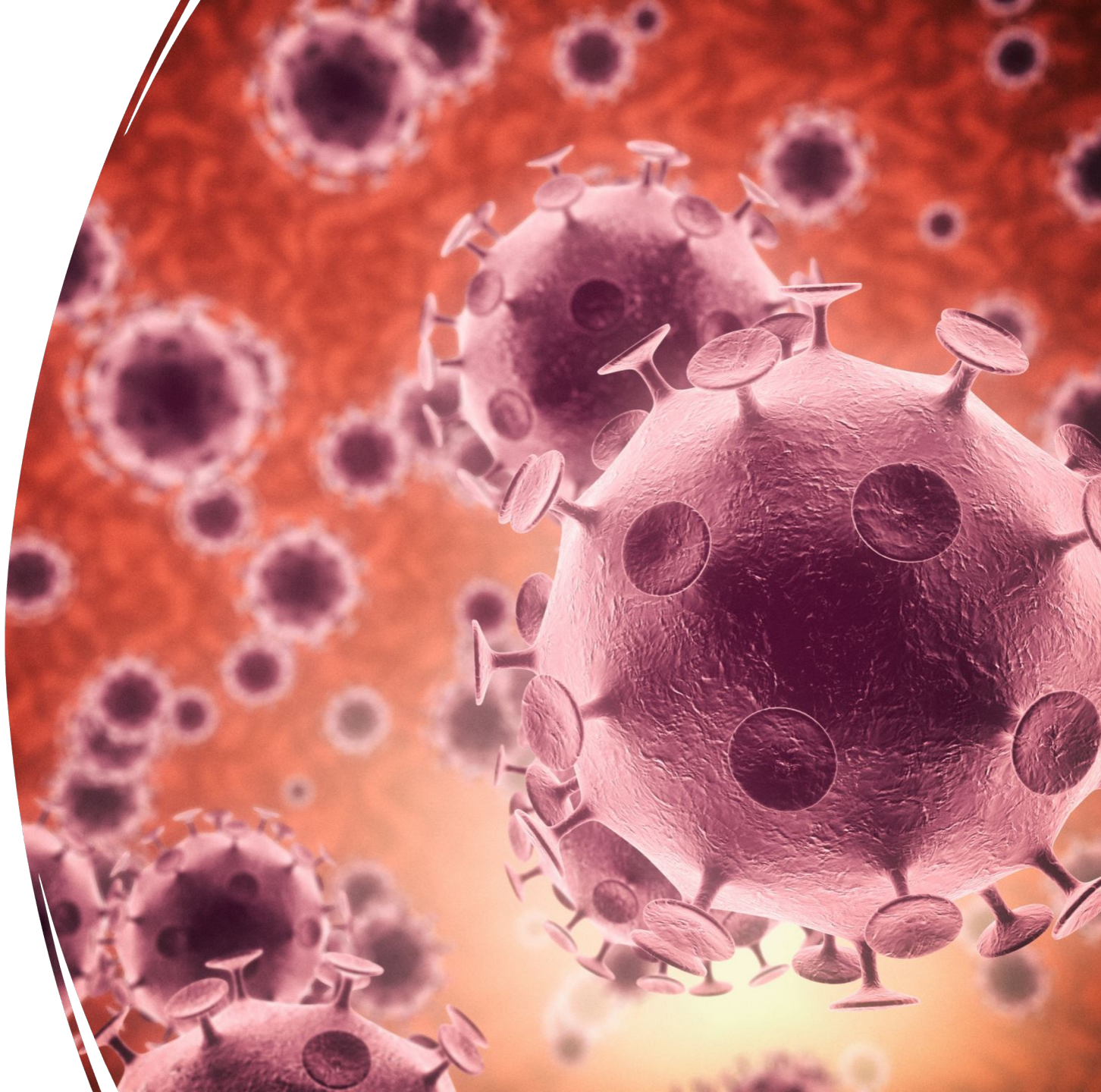
PCR primer diagnostic kits design for Covid-19 (Alpha- Delta- Omicron) in Africa

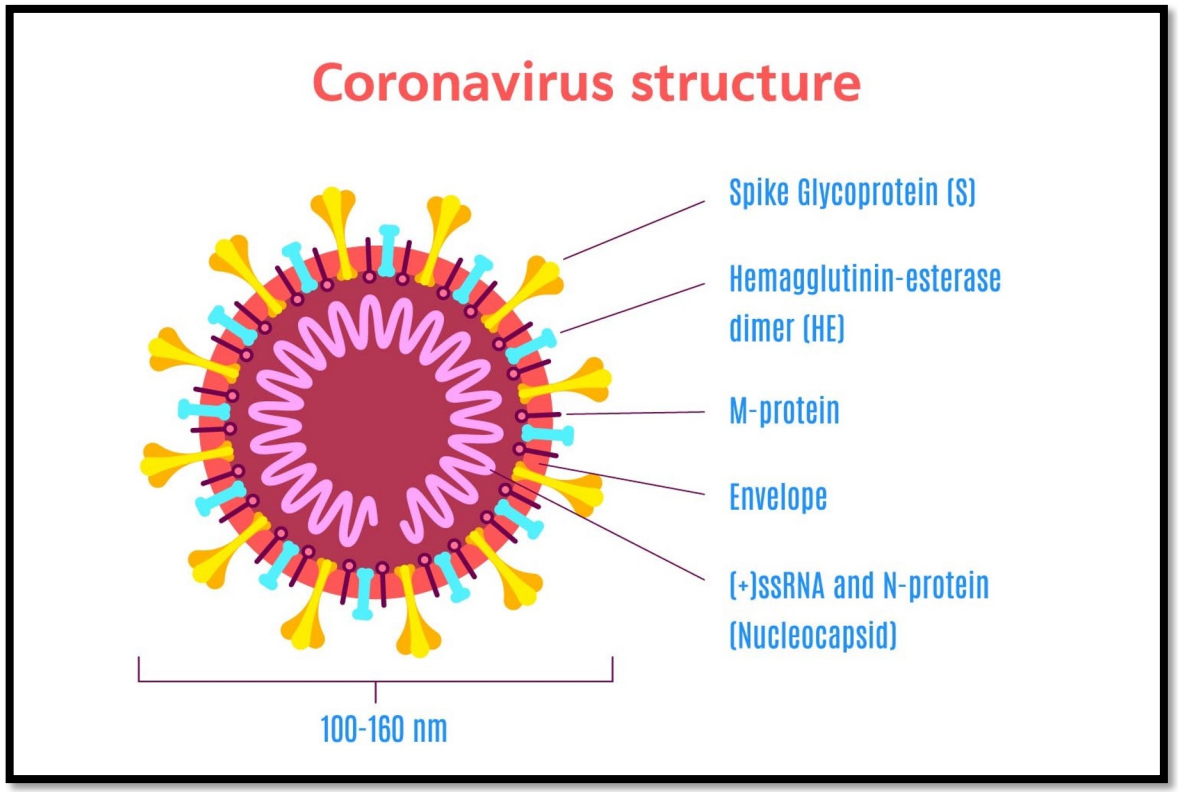
By:

- **Mohamed Elmanzalawi**
- **Ahmed Elghamry**
- **Nour Bahaa**
- **Ahmed Hussein**
- **Yasmine Hosny**
- **Kareem Muhammad**



Introduction

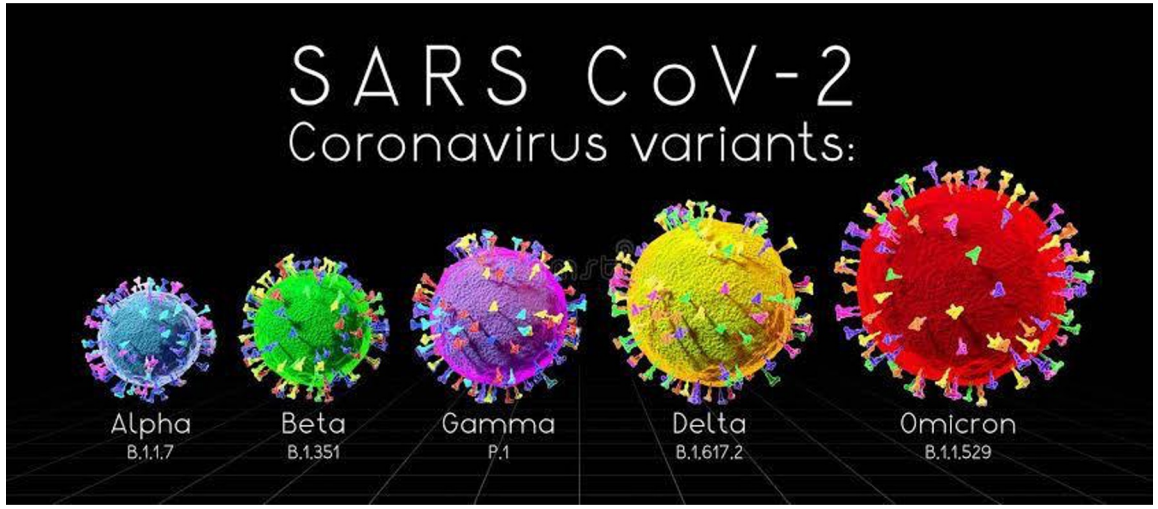
- **Corona Virus Nature.**
  - Coronaviruses are a large family of viruses that cause respiratory infections.
  - **Genetic Material**: there are a highly diverse family of enveloped positive-sense single-stranded RNA viruses.
  - **Host range**: they infect humans, other mammals and other species.
  - **Diversity**: MERS-COV, SARS, SARS-COV-2.
  - **Symptoms**: notably shortness of breath or difficulty breathing, fever or chills and fatigue.



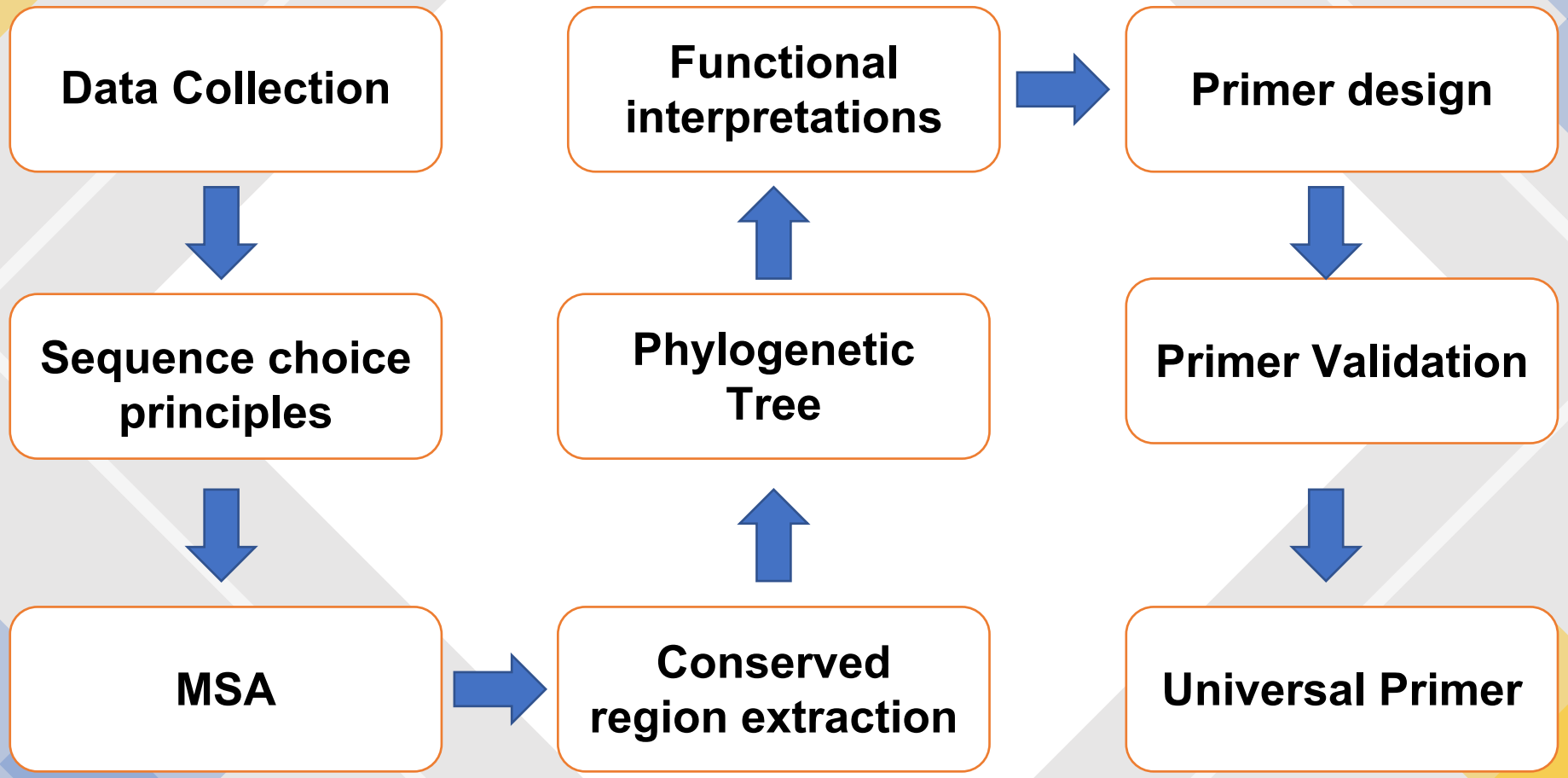
Introduction

- **COVID-19:**
  - 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.
  - Variants:
    - **Alpha variant** with lineage (B.1.1.7) of SARS-CoV-2: the first identified.
    - **Delta variant** with lineage (B.1.617.2): more transmissible around than the Alpha variant.
    - **Omicron variant** with lineage (B.1.1.529): it includes **BA.1**, BA.1.1, BA.2 and BA.3 lineages.

- **Aim**: designing PCR primer kits for the exclusive parts in each Covid-19 variant.



Workflow


Data Collection
Sequence choice principles
MSA
Conserved region extraction
Phylogenetic Tree
Functional interpretations
Primer design
Primer Validation
Universal Primer

Data Collection

Data Collection

- First, we identified each Covid-19 Pango lineage and used the Pango lineages (BA.1, B.1.617.2, B.1.1.7) representing Omicron, Delta, and Alpha respectively.

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

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).

Data Collection

- 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.
- We 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.
- We then 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.
- 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
```

MSA

Multiple sequence alignment

- Multiple alignments were performed using “muscle” command-line after installing it on our machine.
- This was done on every variant’s sequences to get 3 different alignments and 3 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
```

```
OM141321.1      ACAACTTAGCTCCAAATTTGGTGCAATTTCAAGTGTTTTAAATGATATCTTTTCACGTCT
OM141323.1      ACAACTTAGCTCCAAATTTGGTGCAATTTCAAGTGTTTTAAATGATATCTTTTCACGTCT
OM141317.1      ACAACTTAGCTCCAAATTTGGTGCAATTTCAAGTGTTTTAAATGATATCTTTTCACGTCT
OM141461.1      ACAACTTAGCTCCAAATTTGGTGCAATTTCAAGTGTTTTAAATGATATCTTTTCACGTCT
                ********** ********************************** **********
```

Conserved region extraction

Conserved region extraction

- We started extracting our conserved region using our integrated python script in the bash script.
- It produces 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()
```

Conserved region extraction

- Conserved regions were identified, and the longest stretch were used for primer design (as it was more likely that it had a functional role in virus replication.)

```
>Conserved_region(index=13285:13299) length=15
GGTTTTACACTTAAA

>Conserved_region(index=13562:14055) length=494
TGACAATTTAATTGATTCTTACTTTGTAGTTAAGAGACACACTTTCTCTAACTACCAACATGAAGAAACAATTTATAATTTACTTAAGGATTGTCCAGCTGTTGCTAAACATGACTTCTTTAAGTTTAGAATAGACGGT
GACATGGTACCACATATATCACGTCAACGTCTTACTAAATACACAATGGCAGACCTCGTCTATGCTTTAAGGCATTTTGATGAAGGTAATTGTGACACATTAAAAGAAATACTTGTCACATACAATTGTTGTGATGATG
ATTATTTCAATAAAAAGGACTGGTATGATTTTGTAGAAAACCCAGATATATTACGCGTATACGCCAACTTAGGTGAACGTGTACGCCAAGCTTTGTTAAAAACAGTACAATTCTGTGATGCCATGCGAAATGCTGGTAT
TGTTGGTGTACTGACATTAGATAATCAAGATCTCAATGGTAACTGGTATGATTTCGGTGATTTCATACAAACCACGC

>Conserved_region(index=14057:14113) length=57
AGGTAGTGGAGTTCCTGTTGTAGATTCTTATTATTCATTGTTAATGCCTATATTAAC

>Conserved_region(index=14115:14207) length=93
TTGACCAGGGCTTTAACTGCAGAGTCACATGTTGACACTGACTTAACAAAGCCTTACATTAAGTGGGATTTGTTAAAATATGACTTCACGGAA
```

Phylogenetic Tree

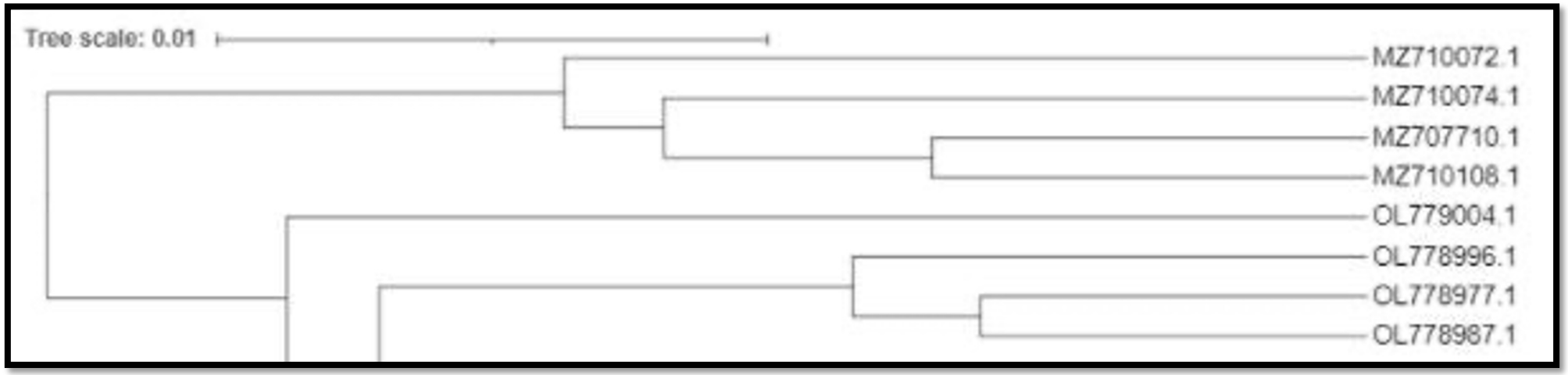
Phylogenetic Tree

- By using “phylo” and “matplotlib” packages in python we got 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))
```

Phylogenetic Tree

- Omicron variant samples phylogenetic tree



Functional products interpretations

Functional products/interpretations

- After obtaining the longest conserved region for each alignment data our script used this data to get all open reading frames.
- 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 argument.

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

Functional products/interpretations

Open Reading Frame Viewer

Help

Sequence

ORFs found: 3 Genetic code: 1 Start codon: 'ATG' only

1 Find:

(U) ORFfinder_2.23.20592835

ORF1

1: 1..494 (494 nt)

Six-frame translation...

ORF1 (116 aa) Display ORF as... Mark

Mark subset... Marked: 0 Download marked s

```
>lcl|ORF1
MVPHISRQRLTKYTMADLVYALRHFDEGNCDTLKEILVTYNCCDDDYFNK
KDWYDFVENPDILRVYANLGERVRQALLKTVQFCDAMRNAGIVGVLTLDN
QDLNGNWYDFGDFIQTT
```

Label	Strand	Frame	Start	Stop
ORF1	+	2	143	
ORF2	-	1	434	
ORF3	-	3	402	

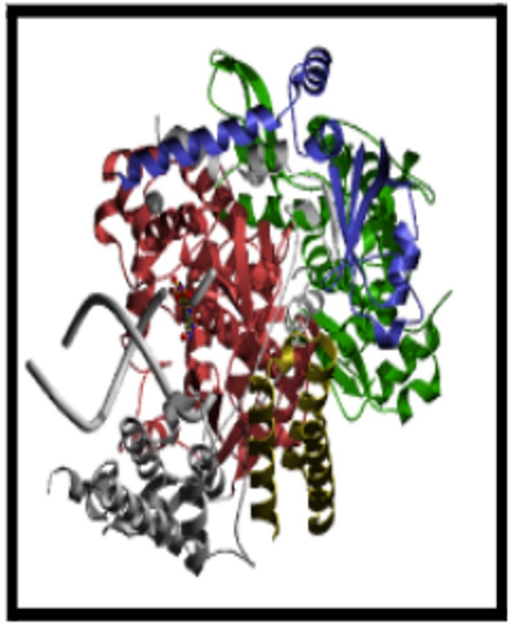


Figure 14. 3D structure of RNA polymerase, N-terminal

Alpha / Delta

Figure 14. 3D structure of Coronavirus replicase NSP8

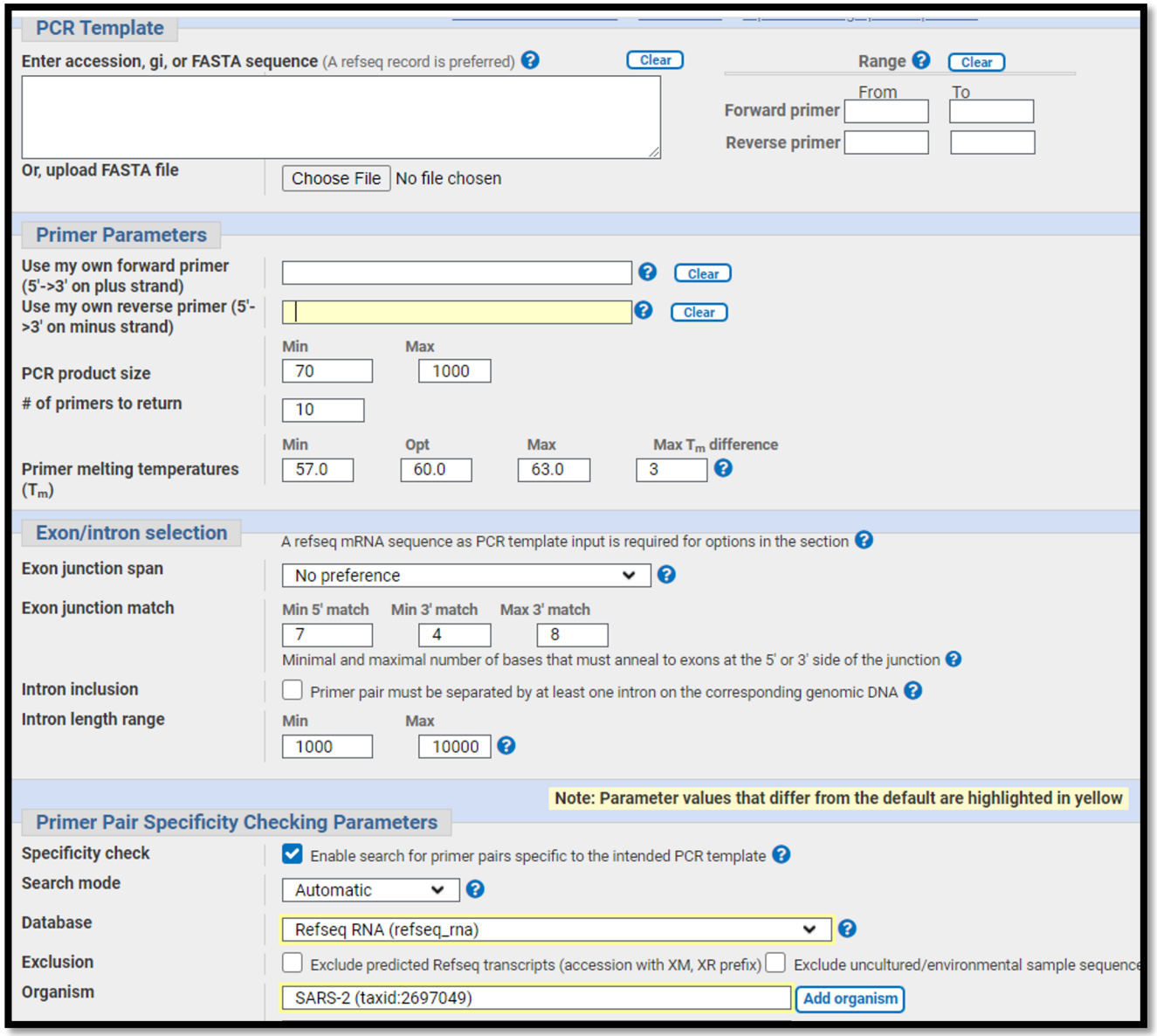
Omicron

Primer design

Primer design

- Using Primer-BLAST we imported every longest conserved region for every variant on it to get a suitable primer for each variant.

- Considerations:
  - Temperature difference must not be more than 3.
  - RefSeq database was chosen.



Alpha 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

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

Delta Variant

Primer pair 2

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	GTTTAGAATAGACGGTGACATGGT	Plus	24	19	42	58.58	41.67	4.00	3.00
Reverse primer	GACGAGGTCTGCCATTGTGT	Minus	20	94	75	60.32	55.00	3.00	0.00

Primer Validation

Primer Validation

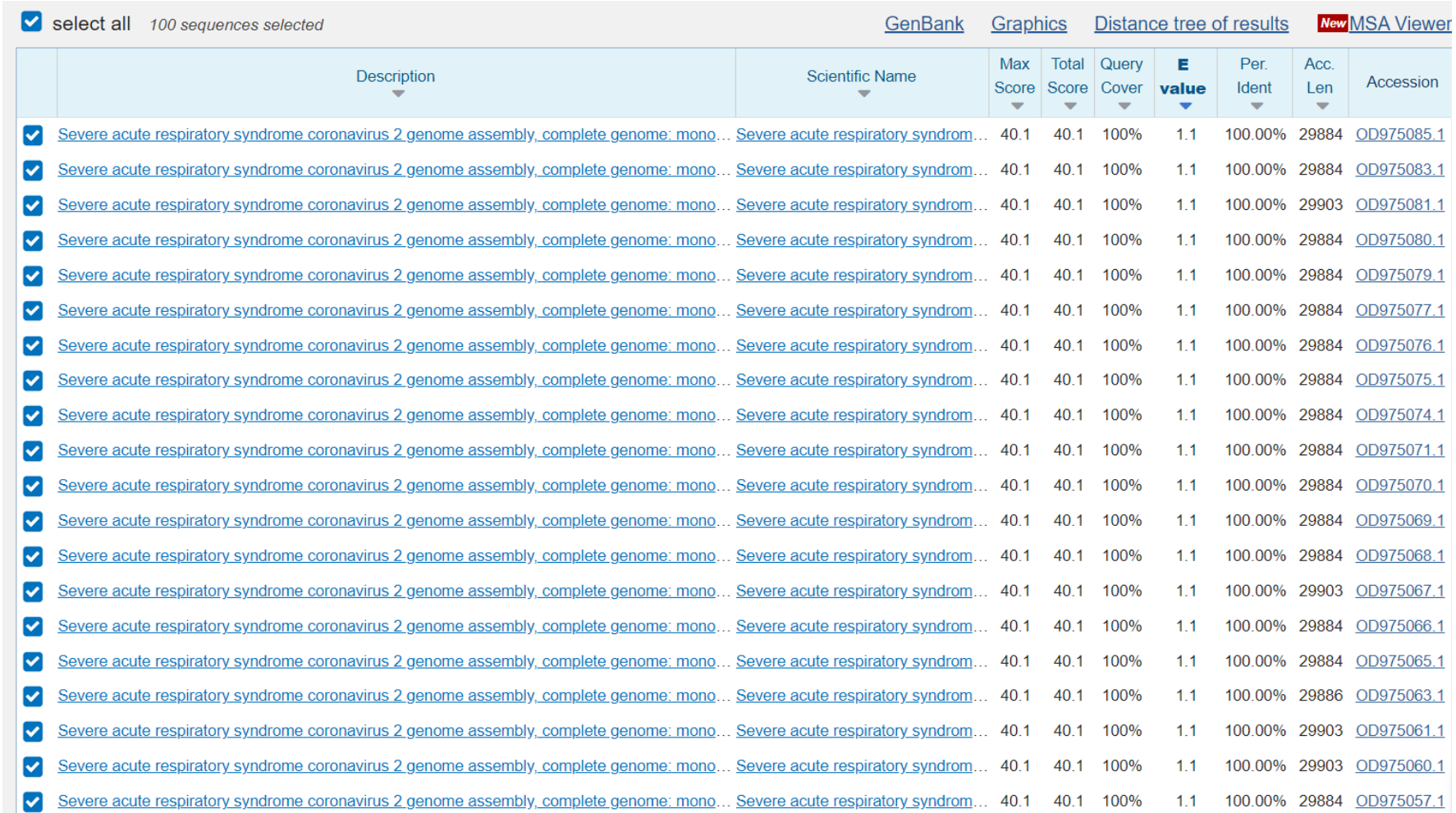
- After the selection of the suitable primer according to GC content, temperature, and self-complementary we validated our results on UCSC In-Silico PCR and PCR primer stats.

```
------------------------------------------------------------
General properties:
-------------------
                    Primer name: forward
                Primer sequence: AGGGCCAATTCTGCTGTCAA
                Sequence length: 20
                    Base counts: G=5; A=5; T=5; C=5; Other=0;
                 GC content (%): 50.00
     Molecular weight (Daltons): 6117.04
                      nmol/A260: NaN
                micrograms/A260: NaN
           Basic Tm (degrees C): 52
   Salt adjusted Tm (degrees C): 47
Nearest neighbor Tm (degrees C): 64.69

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
------------------------------------------------------------
```

Primer Validation

- Validation using blastn search



select all 100 sequences selected GenBank Graphics Distance tree of results New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975085.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975083.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29903	OD975081.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975080.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975079.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975077.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975076.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975075.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975074.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975071.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975070.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975069.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975068.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29903	OD975067.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975066.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975065.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29886	OD975063.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29903	OD975061.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29903	OD975060.1
Severe acute respiratory syndrome coronavirus 2 genome assembly, complete genome: mono…	Severe acute respiratory syndrom…	40.1	40.1	100%	1.1	100.00%	29884	OD975057.1

Universal Primer

Universal Primer

- We proceeded with building a universal primer that can detect any Alpha, Delta, and Omicron variant.
- Due to our low computation power, we used 70 sequences from each variant into a single FASTA file.
- Then we did the MSA following by constructing the phylogenetic tree next obtaining the conserved regions and its ORFs, all this was done by our automated bash/python script.
- 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.

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								

Conclusion

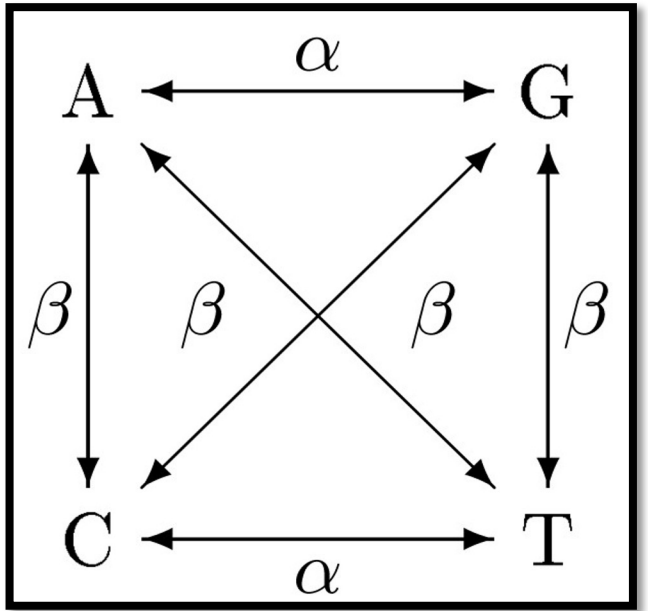
Conclusion

- By deciphering region of consensus for each of Alpha, delta and Omicron variants, and by taking into considerations the constraints of formulating sites to maximize population coverage, matching of melting temperatures in primer pair to drive to experimental conditions and much other parameters.

- Attempting to specify primer pair for Alpha, Delta and Omicron haven’t ever been experimented previously in published researches.

Conclusion

- By deciphering region of consensus for each of Alpha, delta and Omicron variants, and by taking into From the sequences selected from NCBI for each variant mentioned previously. multiple sequence alignment muscle algorithm proceeds in stages. First draft progressive, improved progressive & refinement. In the first step, draft multiple alignment, emphasizing speed.
- The second stage, Kimura distance used to reestimate the binary tree for draft alignment, in turn producing a more accurate multiple alignment.
- Finally, refining the alignment made in second step. In first two stages, time complexity and space complexity. the last stage adds another term to time complexity.



Conclusion

- 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.

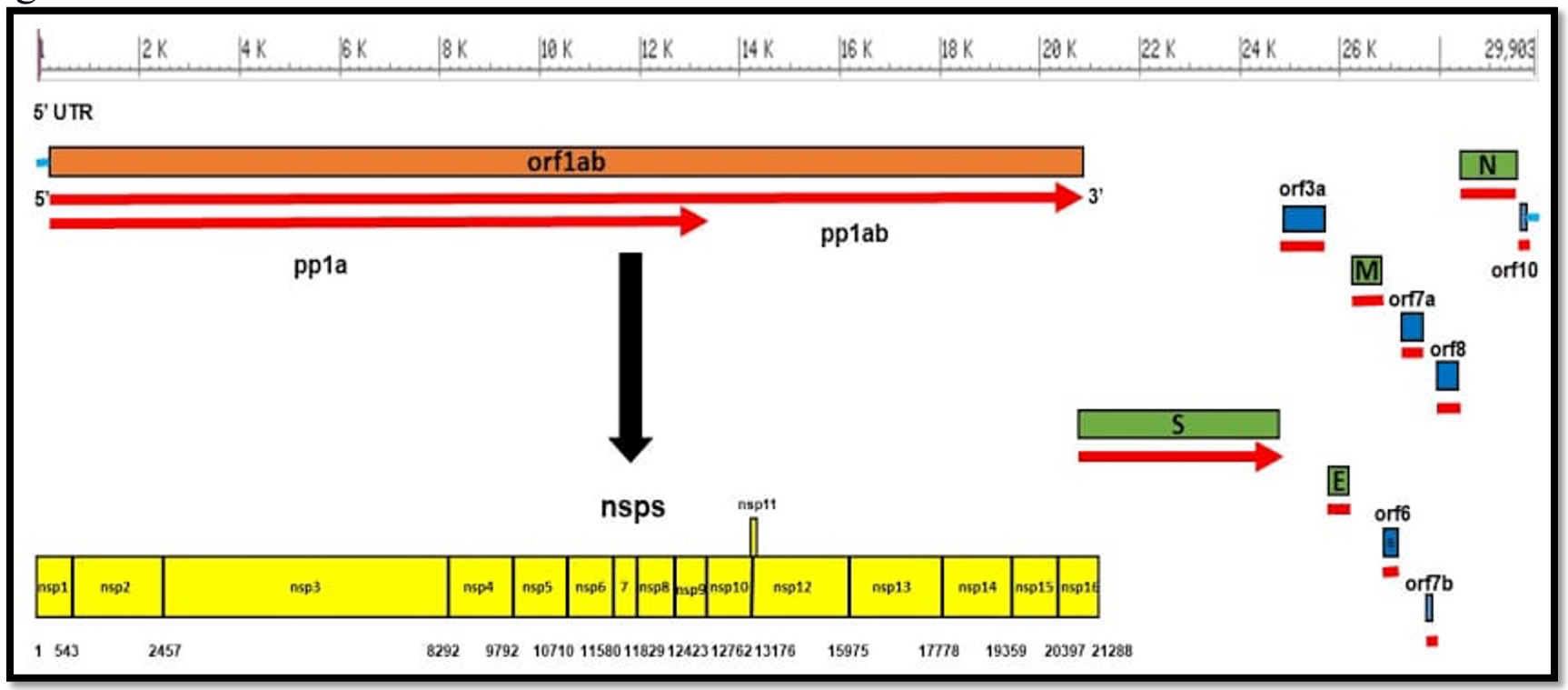
- Aligned sequence pair with computed pairwise identity and conversion to additive distance estimate, applying

- 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.

Conclusion

- 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. into diagnostic kits.



Future suggestions

Future Suggestions

- Increasing the number of sequences covering more countries for our primer design.
- Employing more powerful and robust computational power.
- Testing and characterization of our primers kit in vitro.

Script Demo

Thank you