NILE UNIVERSITY

Covid 19 PCR diagnostic kits

A BMD301 Project Report

By

Dina Yahia Zahran 202001218

Nada Osama Fikry 202002277

Dina Mohammed Sharwida 202002647

Alaa Hussein Mohamed 202001728

Fatema Gamal Soliman 202002674

Submitted in partial fulfillment of the requirements.

for BMD301 Project

January 30, 2023

Abstract

In our experiment, we used three different waves to obtain three sequences of the h-Cov19 for each wave. Waves 1, 2, and 3 occurred between 15 March 2020 and 30 June 2020, 1st September 2020 and 30 April 2021, and 1st July 2021 and 26 September 2021, respectively. The nine sequences were combined into a single file, and multiple sequence alignment was performed on them. We then created our tree to determine which sequences of the three waves are related with each other. NCBI website was used to create our primers as the following step. We selected six consecutive blocks, each of which has full stars at the end and took the first line of each one to represent the conserved region. Then, the conserved region was added to the NCBI Primer-BLAST tool. We received ten primers for the individual region, each with a forward and reverse sequence. To validate our primers, we extracted the forward and reverse sequences for each primer, gave them names, and entered those sequences with their names into a sequence manipulation suite tool. All ten of our primers that were validated had certain warnings, but the best one had only one warning which was that the temperature was over 58 degrees. To find the best conserved region with its best primers, we repeated the conserved region step three times, but each time the temperature warning appeared in the results.

Table of Contents

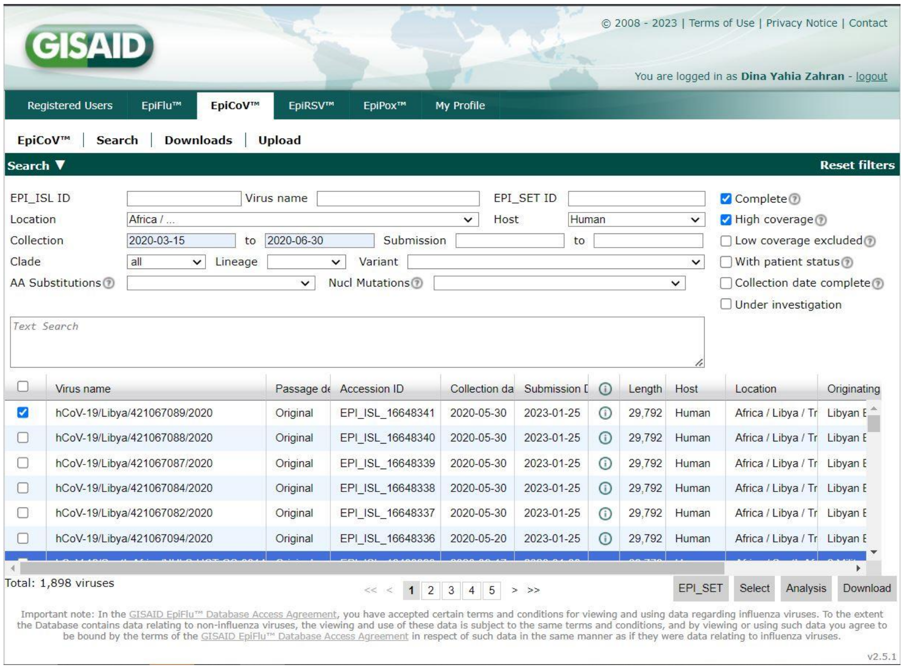
1. Introduction

The discipline of bioinformatics has been expanding globally recently. A branch of biology and computer science called bioinformatics deals with the collection, organization, analysis, and communication of biological data, most frequently DNA and amino acid sequences. Therefore, it is a branch of computational science that deals with the examination of biological molecule sequences [4]. Usually, it relates to proteins, DNA, RNA, or genes. In order to aid with biological concerns, the newly growing field of bioinformatics integrates biology, information science, and mathematics. The majority of bioinformatics researchers use Linux. Because it is unrestricted in terms of production, free, quick, secure, and does not slow down with time. Additionally, it includes tens of thousands of free products, does not require additional antivirus protection beyond routine updates, and has a vibrant community that is eager to assist you with any issues you may have. Because of this, Linux is required for genomics and a variety of other practical tasks. Numerous Linux distributions exist; they are referred to as "distros" since Ubuntu Linux is the most widely used. Linux also features a shell or terminal [4]. A shell is a software that takes user instructions, sends them to the OS for processing, and displays the results. Its key component is the Linux shell. Although some of its distributions include a GUI (graphical user interface), Linux primarily has a CLI (command line interface). Your computer's text user interface is called the Linux command line. A computer software designed to understand instructions is sometimes referred to by the names shell, terminal, console, command prompts, and several other names. Command lines are simple to use and may complete tasks in a few lines. They carry out actions in a substantial amount of code lines, unlike any other programming language. It offers the option of autocomplete. You may explore your computer's files and directories using the command line [4]. Nearly all of the operations that can be done with a GUI can be done using the command line. However, many activities may be completed more quickly and more easily with automation and remote work.

The project's goal is to predict and compare covid 19 sequences across three waves in order to determine which wave has an effect on people and which has a negative impact on DNA sequences. To observe high coverage and complete sequence in genome, we compare sequences using multiple sequence alignment. In addition, to build multiple sequence alignments, we should examine several conserved regions to determine which sequences are related to each other in order to predict which waves have similar properties and therefore have the same effects on genes.

2. Methodology

First, we downloaded three sequences for each one of the three waves of hCov-19 from GISAID [2] based on the date of each one of them, where the first wave ranged from 15 March 2020 till 30 June 2020 [6]. While the second wave ranged from 1st September 2020 till 30 April 2021 [5]. Finally, the third wave ranged between 1st July 2021 till 26 September 2021 [1]. We specified the location of the virus to be Africa and we chose the host to be human. Also, we filtered the sequences by choosing only the ones that are complete and with high coverage so that there won't be gaps in the sequences chosen and to have them matched with each other.



EPI_ISL ID		Virus name		EPI_SET ID	
Location	Africa / ...			Host	Human
Collection	2020-03-15	to	2020-06-30	Submission	to
Clade	all	Lineage		Variant	
AA Substitutions		Nucl Mutations			

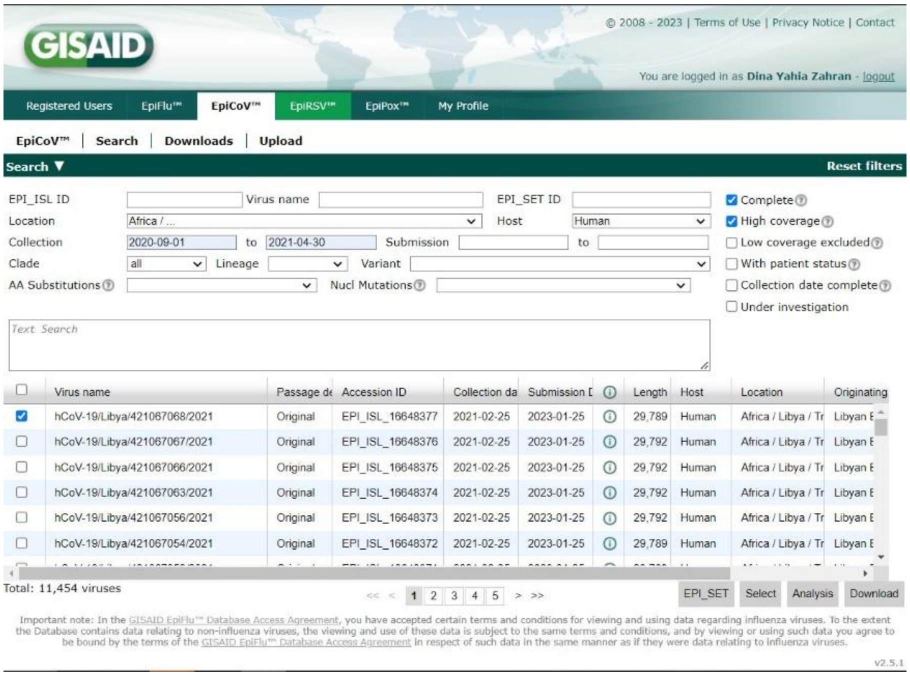
☑ Complete
☑ High coverage
☐ Low coverage excluded
☐ With patient status
☐ Collection date complete
☐ Under investigation

☐	Virus name	Passage de	Accession ID	Collection da	Submission D	ⓘ	Length	Host	Location	Originating
☑	hCoV-19/Libya/421067089/2020	Original	EPI_ISL_16648341	2020-05-30	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067088/2020	Original	EPI_ISL_16648340	2020-05-30	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067087/2020	Original	EPI_ISL_16648339	2020-05-30	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067084/2020	Original	EPI_ISL_16648338	2020-05-30	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067082/2020	Original	EPI_ISL_16648337	2020-05-30	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067094/2020	Original	EPI_ISL_16648336	2020-05-20	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E

Total: 1,898 viruses

Important note: In the GISAID EpiFlu™ Database Access Agreement, you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the GISAID EpiFlu™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses.

Figure 1: Wave 1



EPI_ISL ID		Virus name		EPI_SET ID	
Location	Africa / ...			Host	Human
Collection	2020-09-01	to	2021-04-30	Submission	to
Clade	all	Lineage		Variant	
AA Substitutions		Nucl Mutations			

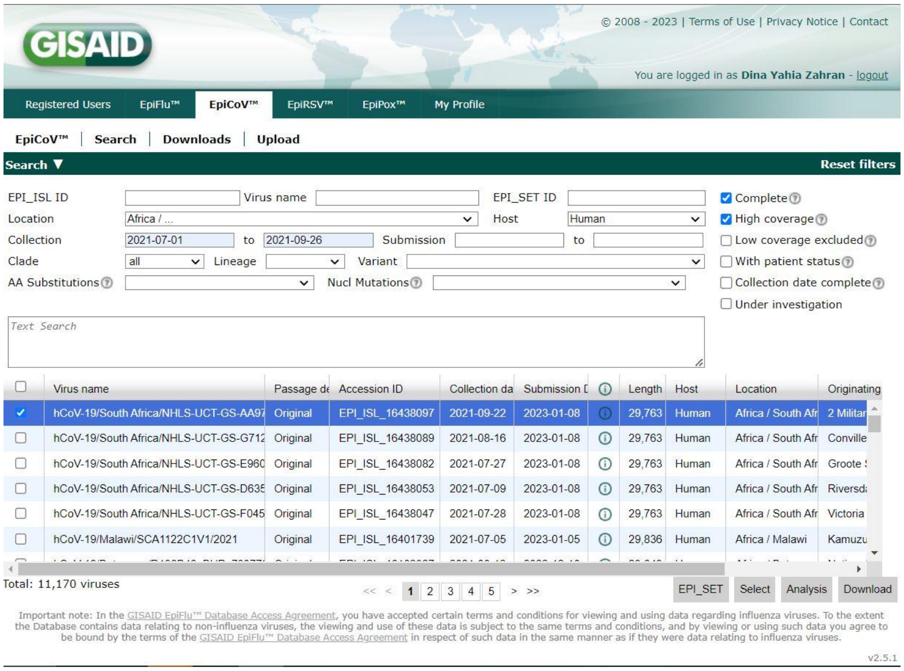
☑ Complete
☑ High coverage
☐ Low coverage excluded
☐ With patient status
☐ Collection date complete
☐ Under investigation

☐	Virus name	Passage de	Accession ID	Collection da	Submission D	ⓘ	Length	Host	Location	Originating
☑	hCoV-19/Libya/421067068/2021	Original	EPI_ISL_16648377	2021-02-25	2023-01-25	ⓘ	29,789	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067067/2021	Original	EPI_ISL_16648376	2021-02-25	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067066/2021	Original	EPI_ISL_16648375	2021-02-25	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067063/2021	Original	EPI_ISL_16648374	2021-02-25	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067056/2021	Original	EPI_ISL_16648373	2021-02-25	2023-01-25	ⓘ	29,792	Human	Africa / Libya / Tr	Libyan E
☐	hCoV-19/Libya/421067054/2021	Original	EPI_ISL_16648372	2021-02-25	2023-01-25	ⓘ	29,789	Human	Africa / Libya / Tr	Libyan E

Total: 11,454 viruses

Important note: In the GISAID EpiFlu™ Database Access Agreement, you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the GISAID EpiFlu™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses.

Figure 2: Wave 2


GISAID
© 2008 - 2023 | Terms of Use | Privacy Notice | Contact
You are logged in as Dina Yahia Zahran - logout
Registered Users
EpiFlu™
EpiCoV™
EpiRSV™
EpiPox™
My Profile
EpiCoV™ | Search | Downloads | Upload
Search ▼
Reset filters
EPI_SET
Select
Analysis
Download
<< < 1 2 3 4 5 > >>
v2.5.1

EPI_ISL ID: Virus name: EPI_SET ID:
Location: Africa / ... Host: Human
Collection: 2021-07-01 to 2021-09-26 Submission: to
Clade: all Lineage: Variant:
AA Substitutions: Nucl Mutations:

☑ Complete
☑ High coverage
□ Low coverage excluded
□ With patient status
□ Collection date complete
□ Under investigation

Text Search

□	Virus name	Passage de	Accession ID	Collection da	Submission D	ⓘ	Length	Host	Location	Originating
☑	hCoV-19/South Africa/NHLS-UCT-GS-AA97	Original	EPI_ISL_16438097	2021-09-22	2023-01-08	ⓘ	29,763	Human	Africa / South Afr	2 Militar
□	hCoV-19/South Africa/NHLS-UCT-GS-G712	Original	EPI_ISL_16438089	2021-08-16	2023-01-08	ⓘ	29,763	Human	Africa / South Afr	Conville
□	hCoV-19/South Africa/NHLS-UCT-GS-E960	Original	EPI_ISL_16438082	2021-07-27	2023-01-08	ⓘ	29,763	Human	Africa / South Afr	Groote S
□	hCoV-19/South Africa/NHLS-UCT-GS-D635	Original	EPI_ISL_16438053	2021-07-09	2023-01-08	ⓘ	29,763	Human	Africa / South Afr	Riversda
□	hCoV-19/South Africa/NHLS-UCT-GS-F045	Original	EPI_ISL_16438047	2021-07-28	2023-01-08	ⓘ	29,763	Human	Africa / South Afr	Victoria
□	hCoV-19/Malawi/SCA1122C1V1/2021	Original	EPI_ISL_16401739	2021-07-05	2023-01-05	ⓘ	29,836	Human	Africa / Malawi	Kamuzu

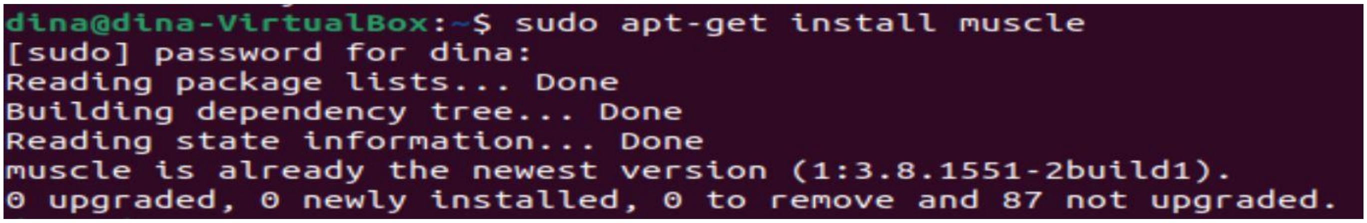
Total: 11,170 viruses

Important note: In the GISAID EpiFlu™ Database Access Agreement, you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the GISAID EpiFlu™ Database Access Agreement in respect of such data in the same manner as if they were data relating to influenza viruses.

Figure 3: Wave 3

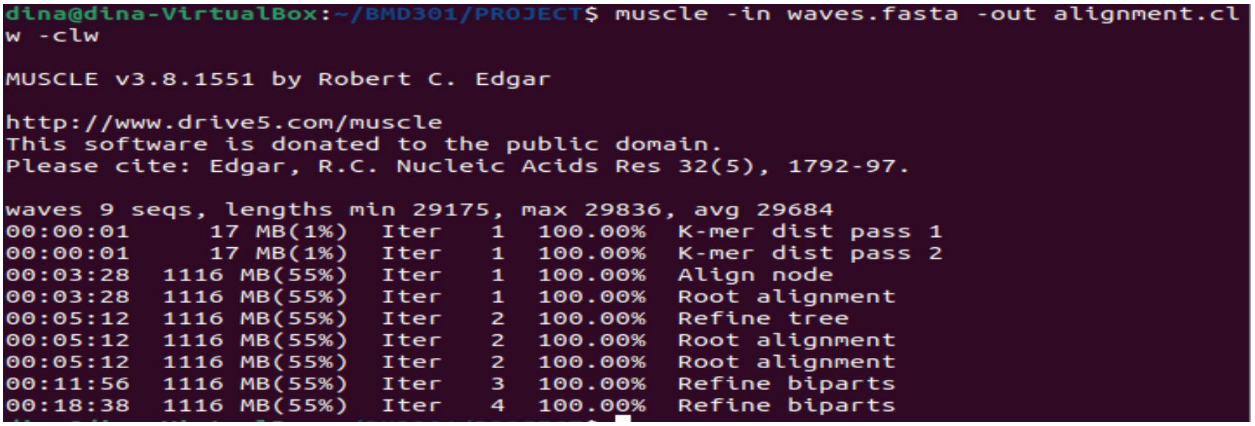
2.1. Multiple Sequence Alignment

Then we installed muscle tool to perform multiple sequence alignment between nine sequences of the three waves.



```
dina@dina-VirtualBox:~$ sudo apt-get install muscle
[sudo] password for dina:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
muscle is already the newest version (1:3.8.1551-2build1).
0 upgraded, 0 newly installed, 0 to remove and 87 not upgraded.
```

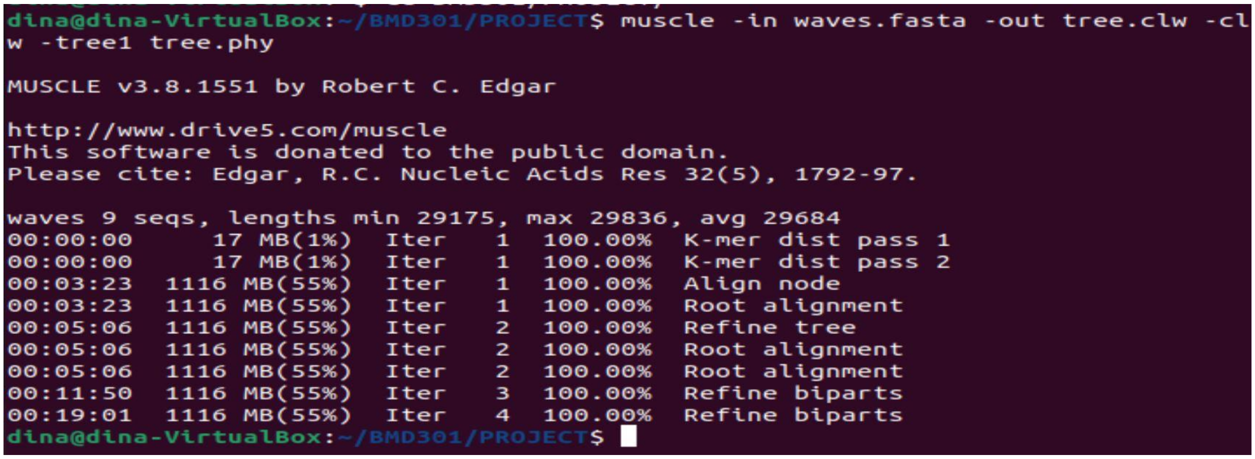
Next, we converted the Fasta file including the aligned nine sequences to Clusterlw file to get sequences in blocks and extract the conserved region from them in order to design the primers suitable for this region.



```
dina@dina-VirtualBox:~/BMD301/PROJECT$ muscle -in waves.fasta -out alignment.cl
w -clw

MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

waves 9 seqs, lengths min 29175, max 29836, avg 29684
00:00:01      17 MB(1%)   Iter   1  100.00%  K-mer dist pass 1
00:00:01      17 MB(1%)   Iter   1  100.00%  K-mer dist pass 2
00:03:28   1116 MB(55%)   Iter   1  100.00%  Align node
00:03:28   1116 MB(55%)   Iter   1  100.00%  Root alignment
00:05:12   1116 MB(55%)   Iter   2  100.00%  Refine tree
00:05:12   1116 MB(55%)   Iter   2  100.00%  Root alignment
00:05:12   1116 MB(55%)   Iter   2  100.00%  Root alignment
00:11:56   1116 MB(55%)   Iter   3  100.00%  Refine biparts
00:18:38   1116 MB(55%)   Iter   4  100.00%  Refine biparts
```

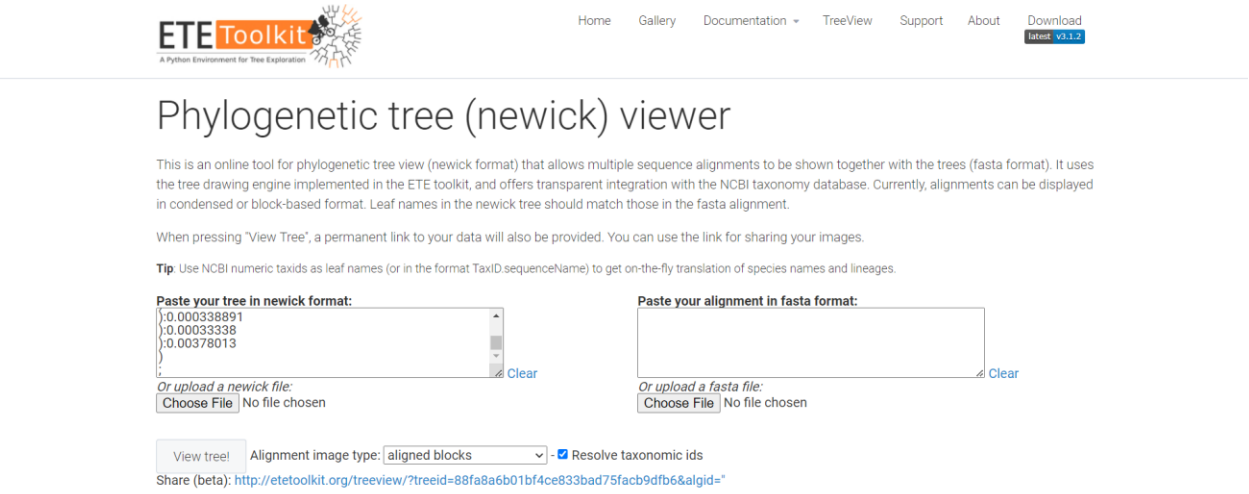
2.2. Phylogenetic Tree

We also created the phylogenetic tree from the Clusterlw file to view matched and aligned sequences.



```
dina@dina-VirtualBox:~/BMD301/PROJECT$ muscle -in waves.fasta -out tree.clw -cl
w -tree1 tree.phy

MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

waves 9 seqs, lengths min 29175, max 29836, avg 29684
00:00:00      17 MB(1%)   Iter   1  100.00%  K-mer dist pass 1
00:00:00      17 MB(1%)   Iter   1  100.00%  K-mer dist pass 2
00:03:23   1116 MB(55%)   Iter   1  100.00%  Align node
00:03:23   1116 MB(55%)   Iter   1  100.00%  Root alignment
00:05:06   1116 MB(55%)   Iter   2  100.00%  Refine tree
00:05:06   1116 MB(55%)   Iter   2  100.00%  Root alignment
00:05:06   1116 MB(55%)   Iter   2  100.00%  Root alignment
00:11:50   1116 MB(55%)   Iter   3  100.00%  Refine biparts
00:19:01   1116 MB(55%)   Iter   4  100.00%  Refine biparts
dina@dina-VirtualBox:~/BMD301/PROJECT$ 
```

Phylogenetic tree (newick) viewer is a tool that was used in the project for visualizing the phylogenetic tree as a dendrogram to make the interpretation process of the tree easier.



Phylogenetic tree (newick) viewer

This is an online tool for phylogenetic tree view (newick format) that allows multiple sequence alignments to be shown together with the trees (fasta format). It uses the tree drawing engine implemented in the ETE toolkit, and offers transparent integration with the NCBI taxonomy database. Currently, alignments can be displayed in condensed or block-based format. Leaf names in the newick tree should match those in the fasta alignment.

When pressing "View Tree", a permanent link to your data will also be provided. You can use the link for sharing your images.

Tip: Use NCBI numeric taxids as leaf names (or in the format TaxID.sequenceName) to get on-the-fly translation of species names and lineages.

Paste your tree in newick format:

```
):0.000338891
):0.00033338
):0.00378013
)
;
```

Or upload a newick file: No file chosen

Paste your alignment in fasta format:

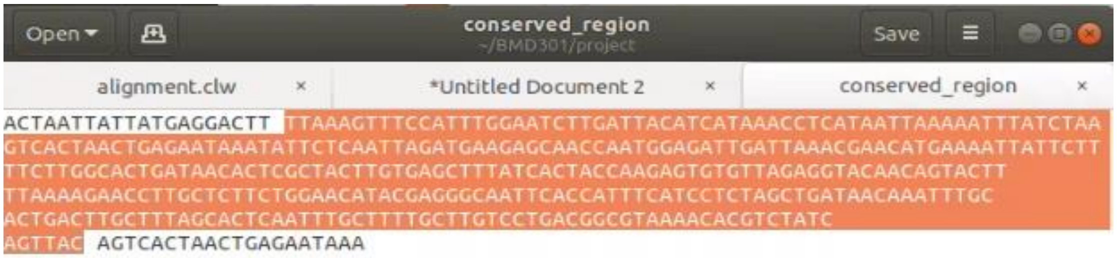
Or upload a fasta file: No file chosen

Alignment image type: aligned blocks - ☑ Resolve taxonomic ids

Share (beta): http://etetoolkit.org/treeview/?treeid=88fa8a6b01bf4ce833bad75facb9dfb6&algid="

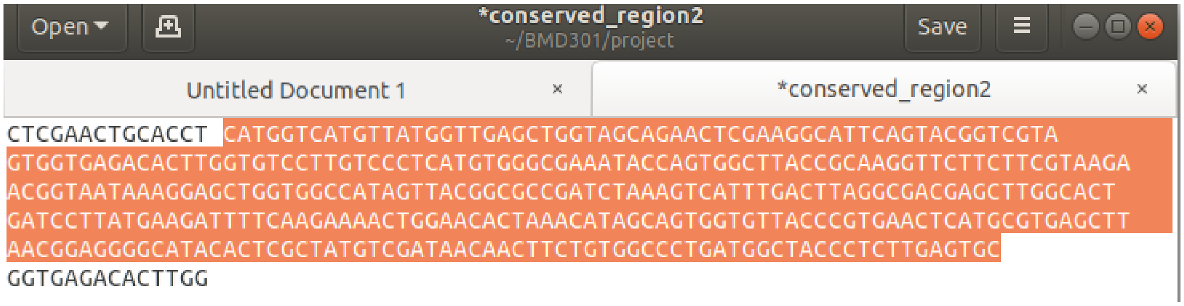
2.3. Conserved Regions

We specified three different conserved regions from different consecutive blocks in Clusterlw file to see which one will have best primers. For the first conserved region, we chose six lines consecutively from the file to represent the conserved region of the multiple sequences and put it in a separated text file and selected from a random original sequence 20 bases before the first line chosen and 20 bases after that represent forward and reverse bases.



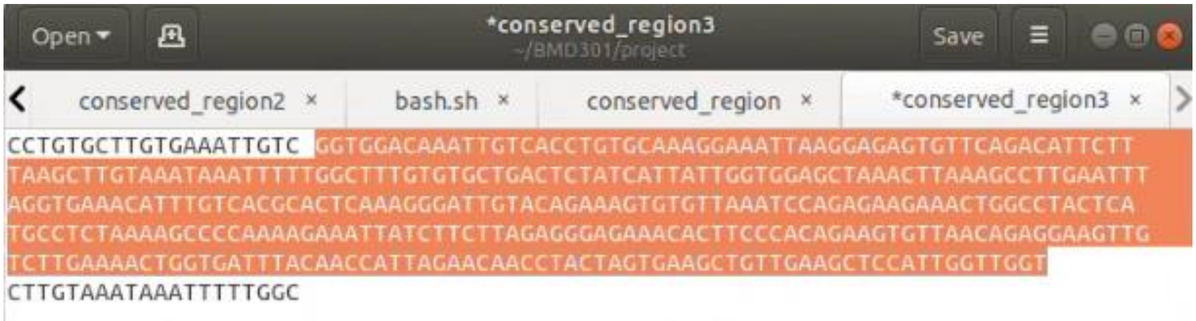
```
ACTAATTATTATGAGGACTT TTAAAGTTTCCATTTGGAATCTTGATTACATCATAAACCTCATAATTAAAAATTTATCTAA
GTCACTAACTGAGAATAAATATTCTCAATTAGATGAAGAGCAACCAATGGAGATTGATTAAACGAACATGAAAATTATTCTT
TTCTTGGCACTGATAACACTCGCTACTTGTGAGCTTTATCACTACCAAGAGTGTGTTAGAGGTACAACAGTACTT
TTAAAAGAACCTTGCTCTTCTGGAACATACGAGGGCAATTCACCATTTCATCCTCTAGCTGATAACAAATTTGC
ACTGACTTGCTTTAGCACTCAATTTGCTTTTGCTTGTCCTGACGGCGTAAAACACGTCTATC
AGTTAC AGTCACTAACTGAGAATAAA
```

We performed the same steps for specifying the other two conserved regions and designing their own primers so that at the end, we can compare which resulted primers are the best.



```
CTCGAACTGCACCT CATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCAGTACGGTCGTA
GTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGA
ACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACT
GATCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTT
AACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGGCCCTGATGGCTACCCTCTTGAGTGC
GGTGAGACACTTGG
```

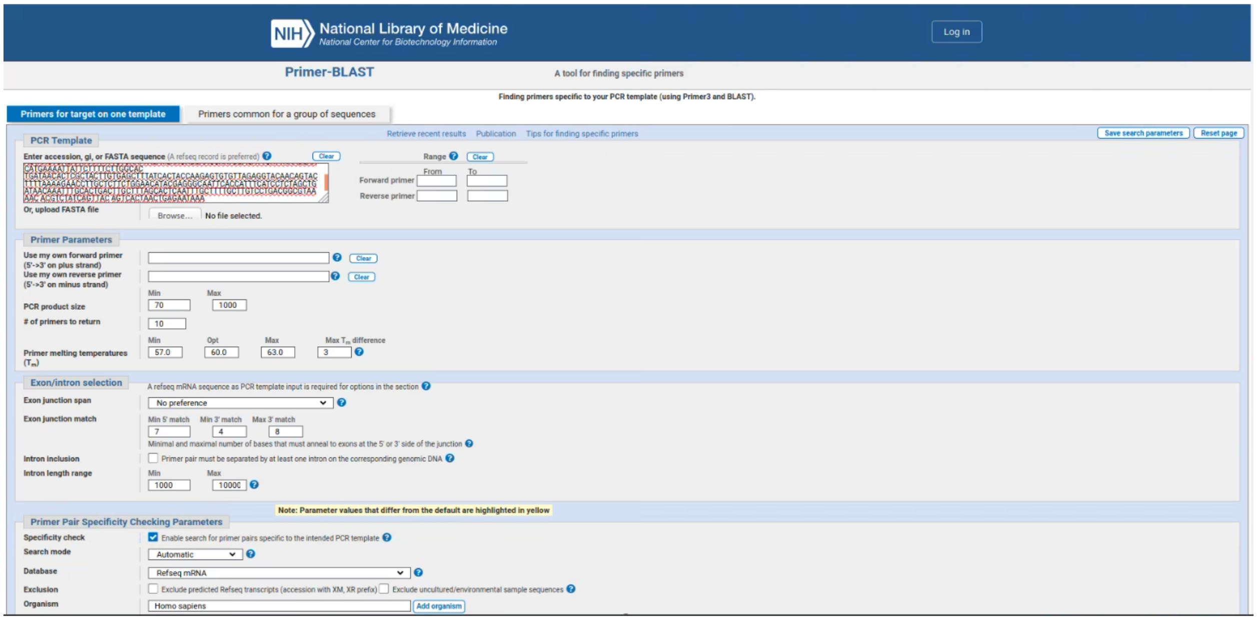
Figure 5: Conserved Region 2



```
CCTGTGCTTGTGAAATTGTC GGTGGACAAATTGTCACCTGTGCAAAGGAAATTAAGGAGAGTGTTCAGACATTCTT
TAAGCTTGTAAATAAATTTTTGGCTTTGTGTGCTGACTCTATCATTATTGGTGGAGCTAAACTTAAAGCCTTGAATTT
AGGTGAAACATTTGTCACGCACTCAAAGGGATTGTACAGAAAGTGTGTTAAATCCAGAGAAGAAACTGGCCTACTCA
TGCCTCTAAAAGCCCCAAAAGAAATTATCTTCTTAGAGGGAGAAACACTTCCCACAGAAGTGTTAACAGAGGAAGTTG
TCTTGAAAACTGGTGATTTACAACCATTAGAACAACCTACTAGTGAAGCTGTTGAAGCTCCATTGGTTGGT
CTTGTAAATAAATTTTTGGC
```

Figure 4: Conserved Region 3

After specifying the first conserved region, we put it on NCBI Primer-BLAST [7] to design 10 primers for the specified conserved region.


NIH
National Library of Medicine
National Center for Biotechnology Information
Log in
Primer-BLAST
A tool for finding specific primers
Primers for target on one template
Primers common for a group of sequences
Retrieve recent results
Publication
Tips for finding specific primers
Save search parameters
Reset page
Clear
Browse...
Add organism

Finding primers specific to your PCR template (using Primer3 and BLAST).

PCR Template

Enter accession, gi, or FASTA sequence (A refseq record is preferred)

```
CATGAAAATTATTCTTTTCTTGGCAC
TGATAACACTCGCTACTTGTGAGCTTTATCACTACCAAGAGTGTGTTAGAGGTACAACAGTAC
TTTTAAAAGAACCTTGCTCTTCTGGAACATACGAGGGCAATTCACCATTTCATCCTCTAGCTG
ATAACAAATTTGCACTGACTTGCTTTAGCACTCAATTTGCTTTTGCTTGTCCTGACGGCGTAA
AAC ACGTCTATCAGTTAC AGTCACTAACTGAGAATAAA
```

Or, upload FASTA file: No file selected.

Range — From / To
Forward primer
Reverse primer

Primer Parameters

Use my own forward primer (5'->3' on plus strand)
Use my own reverse primer (5'->3' on minus strand)

	Min	Max		
PCR product size	70	1000		
# of primers to return	10			
	Min	Opt	Max	Max T_m difference
Primer melting temperatures (T_m)	57.0	60.0	63.0	3

Exon/intron selection

A refseq mRNA sequence as PCR template input is required for options in the section

Exon junction span	No preference		
	Min 5' match	Min 3' match	Max 3' match
Exon junction match	7	4	8

Minimal and maximal number of bases that must anneal to exons at the 5' or 3' side of the junction

Intron inclusion: □ Primer pair must be separated by at least one intron on the corresponding genomic DNA

	Min	Max
Intron length range	1000	10000

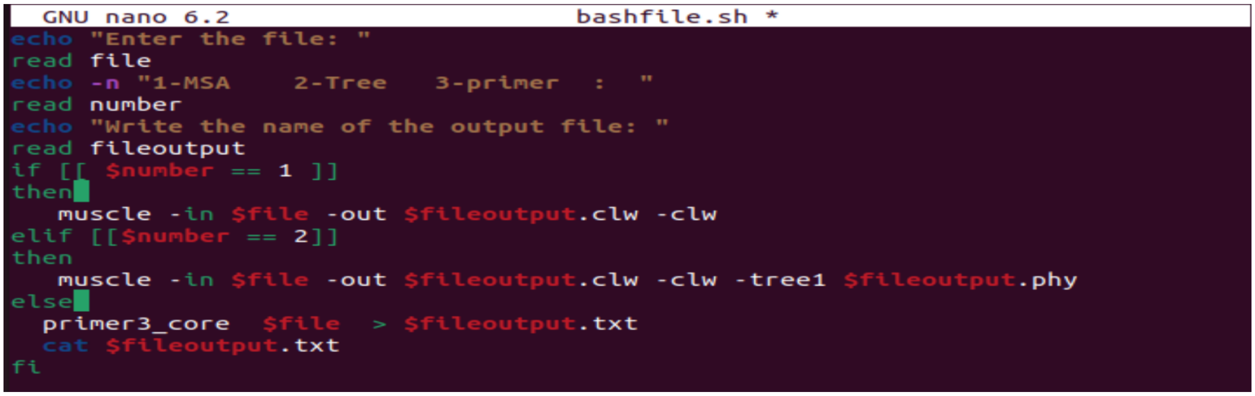
Note: Parameter values that differ from the default are highlighted in yellow

Primer Pair Specificity Checking Parameters

Specificity check: ☑ Enable search for primer pairs specific to the intended PCR template
Search mode: Automatic
Database: Refseq mRNA
Exclusion: □ Exclude predicted Refseq transcripts (accession with XM, XR prefix) □ Exclude uncultured/environmental sample sequences
Organism: Homo sapiens

2.4. Automated Bash Script

At the end, we automated the multiple sequence alignment, phylogenetic tree and primer design parts using bash script, where the user is asked to enter the specific file, put a name for the output file and choose the process he/she wants to perform on the sequence, in order to make the process more generic and not specific for a certain file. This was done using “nano bashscript.sh” and writing the commands for performing such processes inside it.

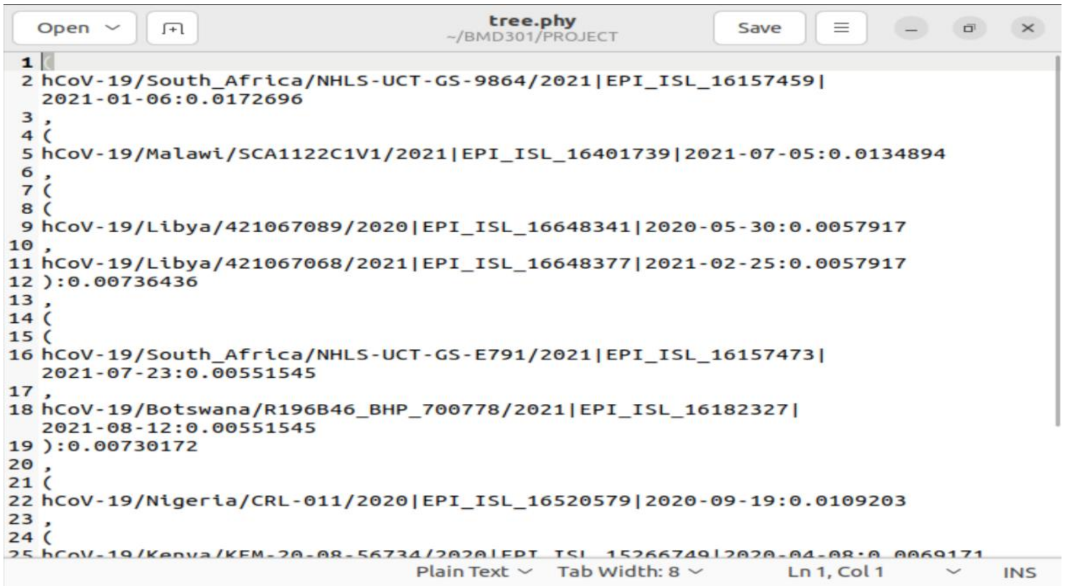


```
  GNU nano 6.2                           bashfile.sh *
echo "Enter the file: "
read file
echo -n "1-MSA    2-Tree   3-primer  :  "
read number
echo "Write the name of the output file: "
read fileoutput
if [[ $number == 1 ]]
then
   muscle -in $file -out $fileoutput.clw -clw
elif [[$number == 2]]
then
   muscle -in $file -out $fileoutput.clw -clw -tree1 $fileoutput.phy
else
  primer3_core  $file  > $fileoutput.txt
  cat $fileoutput.txt
fi
```

3. Results

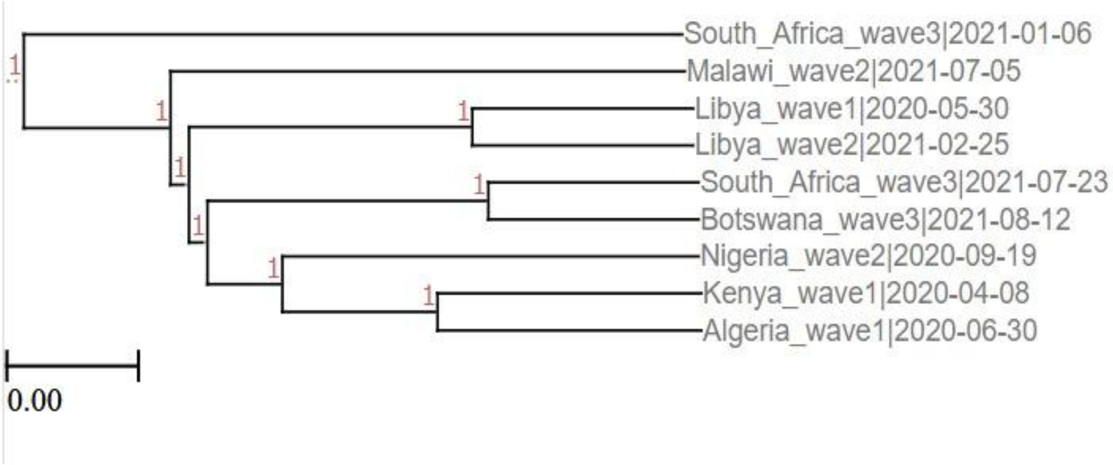
3.1. Phylogenetic Tree

The following figure represents the phylogenetic tree produced from the nine sequences of the three different waves. It was found that there are two different sequences from two different waves are similar to each other which are the two sequences from Libya in wave one and two, where both of them are complete and of high coverage. The rest of the similar sequences are in the same waves but might be from different countries.



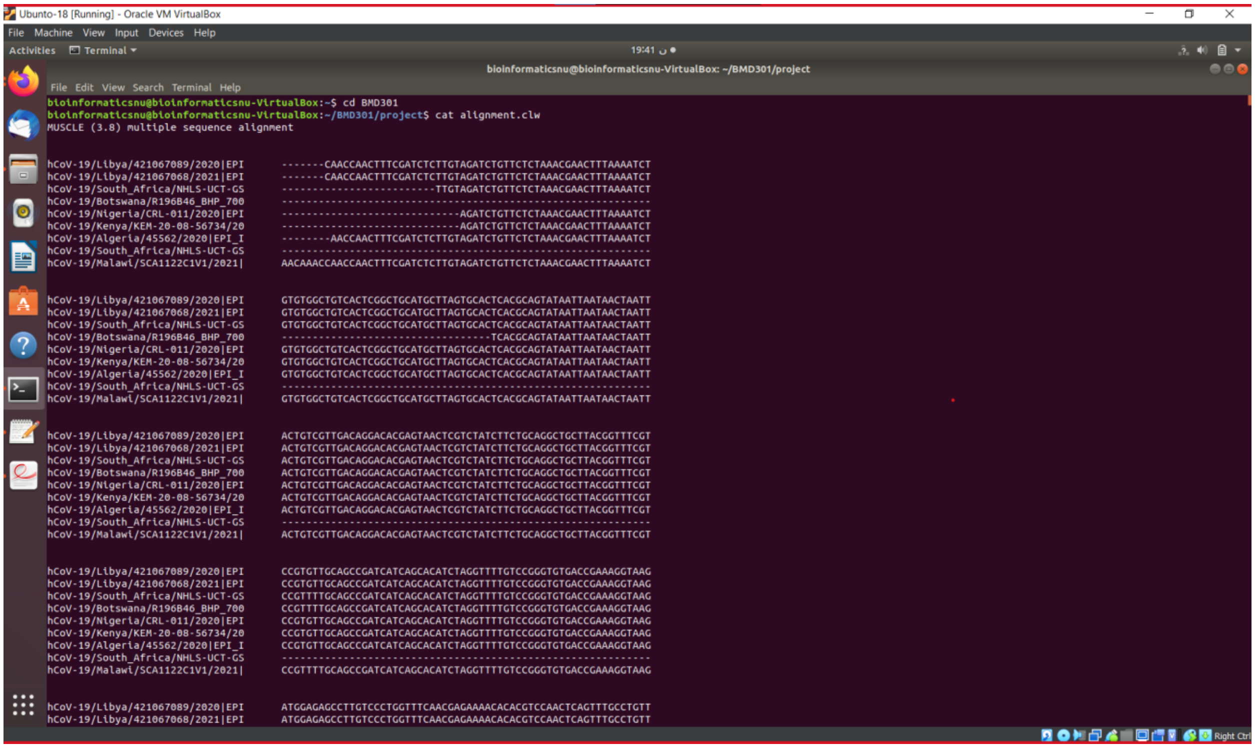
```
(
hCoV-19/South_Africa/NHLS-UCT-GS-9864/2021|EPI_ISL_16157459|2021-01-06:0.0172696
,
(
hCoV-19/Malawi/SCA1122C1V1/2021|EPI_ISL_16401739|2021-07-05:0.0134894
,
(
(
hCoV-19/Libya/421067089/2020|EPI_ISL_16648341|2020-05-30:0.0057917
,
hCoV-19/Libya/421067068/2021|EPI_ISL_16648377|2021-02-25:0.0057917
):0.00736436
,
(
(
hCoV-19/South_Africa/NHLS-UCT-GS-E791/2021|EPI_ISL_16157473|2021-07-23:0.00551545
,
hCoV-19/Botswana/R196B46_BHP_700778/2021|EPI_ISL_16182327|2021-08-12:0.00551545
):0.00730172
,
(
hCoV-19/Nigeria/CRL-011/2020|EPI_ISL_16520579|2020-09-19:0.0109203
,
(
hCoV-19/Kenya/KEM-20-08-56734/2020|EPI_ISL_15266749|2020-04-08:0.0069171
```

The output of the Phylogenetic tree (newick) viewer tool [8] is represented in the following figure.



3.2. Multiple Sequence Alignment

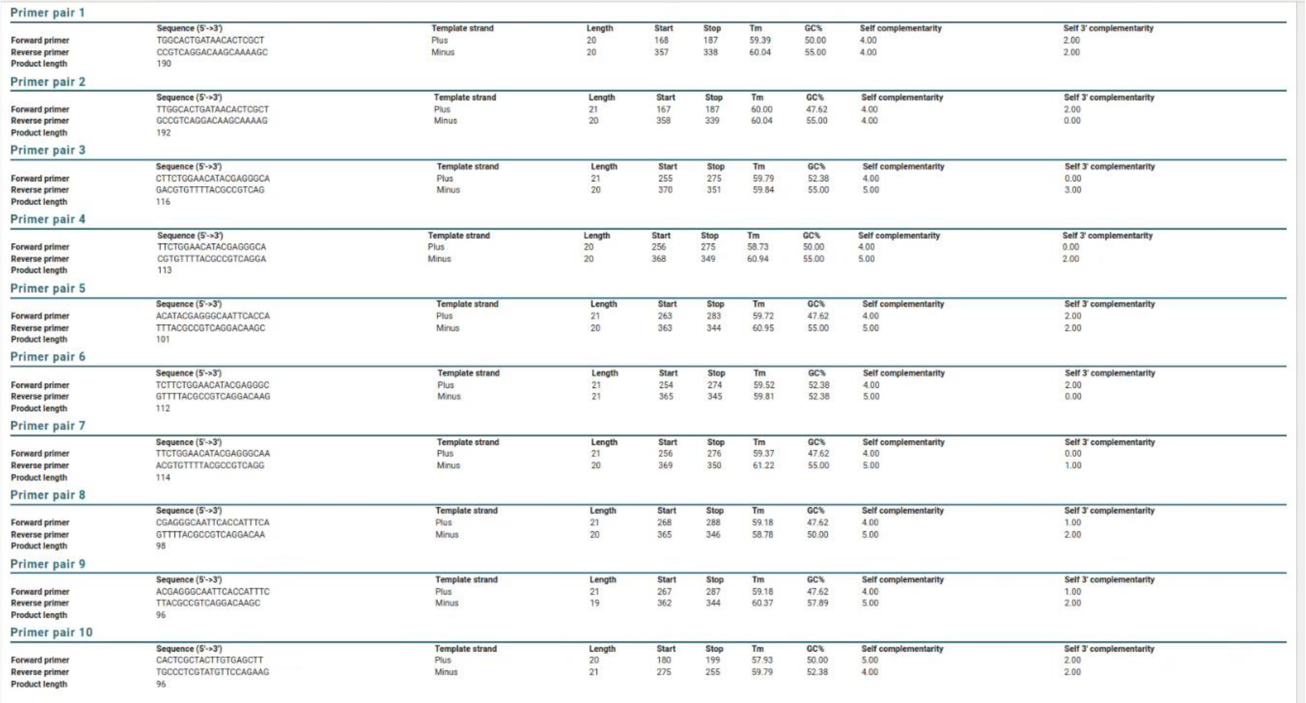
We use command ‘cat’ to visualize all aligned sequences in file alignment.clw to extract the conserved regions for designing the primers.


Ubunto-18 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
19:41
bioinformaticsnu@bioinformaticsnu-VirtualBox: ~/BMD301/project
File Edit View Search Terminal Help
Right Ctrl

```
bioinformaticsnu@bioinformaticsnu-VirtualBox:~$ cd BMD301
bioinformaticsnu@bioinformaticsnu-VirtualBox:~/BMD301/project$ cat alignment.clw
MUSCLE (3.8) multiple sequence alignment


hCoV-19/Libya/421067089/2020|EPI      -------CAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCT
hCoV-19/Libya/421067068/2021|EPI      -------CAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCT
hCoV-19/South_Africa/NHLS-UCT-GS      -------------------------TTGTAGATCTGTTCTCTAAACGAACTTTAAAATCT
hCoV-19/Botswana/R196B46_BHP_700      ------------------------------------------------------------
hCoV-19/Nigeria/CRL-011/2020|EPI      -----------------------------AGATCTGTTCTCTAAACGAACTTTAAAATCT
hCoV-19/Kenya/KEM-20-08-56734/20      -----------------------------AGATCTGTTCTCTAAACGAACTTTAAAATCT
hCoV-19/Algeria/45562/2020|EPI_I      --------AACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCT
hCoV-19/South_Africa/NHLS-UCT-GS      ------------------------------------------------------------
hCoV-19/Malawi/SCA1122C1V1/2021|      AACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCT


hCoV-19/Libya/421067089/2020|EPI      GTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATT
hCoV-19/Libya/421067068/2021|EPI      GTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATT
hCoV-19/South_Africa/NHLS-UCT-GS      GTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATT
hCoV-19/Botswana/R196B46_BHP_700      ----------------------------------TCACGCAGTATAATTAATAACTAATT
hCoV-19/Nigeria/CRL-011/2020|EPI      GTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATT
hCoV-19/Kenya/KEM-20-08-56734/20      GTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATT
hCoV-19/Algeria/45562/2020|EPI_I      GTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATT
hCoV-19/South_Africa/NHLS-UCT-GS      ------------------------------------------------------------
hCoV-19/Malawi/SCA1122C1V1/2021|      GTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAACTAATT


hCoV-19/Libya/421067089/2020|EPI      ACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGT
hCoV-19/Libya/421067068/2021|EPI      ACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGT
hCoV-19/South_Africa/NHLS-UCT-GS      ACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGT
hCoV-19/Botswana/R196B46_BHP_700      ACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGT
hCoV-19/Nigeria/CRL-011/2020|EPI      ACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGT
hCoV-19/Kenya/KEM-20-08-56734/20      ACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGT
hCoV-19/Algeria/45562/2020|EPI_I      ACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGT
hCoV-19/South_Africa/NHLS-UCT-GS      ------------------------------------------------------------
hCoV-19/Malawi/SCA1122C1V1/2021|      ACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGT


hCoV-19/Libya/421067089/2020|EPI      CCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAG
hCoV-19/Libya/421067068/2021|EPI      CCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAG
hCoV-19/South_Africa/NHLS-UCT-GS      CCGTTTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAG
hCoV-19/Botswana/R196B46_BHP_700      CCGTTTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAG
hCoV-19/Nigeria/CRL-011/2020|EPI      CCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAG
hCoV-19/Kenya/KEM-20-08-56734/20      CCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAG
hCoV-19/Algeria/45562/2020|EPI_I      CCGTGTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAG
hCoV-19/South_Africa/NHLS-UCT-GS      ------------------------------------------------------------
hCoV-19/Malawi/SCA1122C1V1/2021|      CCGTTTTGCAGCCGATCATCAGCACATCTAGGTTTTGTCCGGGTGTGACCGAAAGGTAAG


hCoV-19/Libya/421067089/2020|EPI      ATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTT
hCoV-19/Libya/421067068/2021|EPI      ATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTT
```

3.3. Conserved Regions

According to the three conserved regions selected from the aligned sequence file, we designed 10 primers for each region.



Primer pair 1

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TGGCACTGATAACACTCGCT	Plus	20	168	187	59.39	50.00	4.00	2.00
Reverse primer	CCGTCAGGACAAGCAAAAGC	Minus	20	357	338	60.04	55.00	4.00	2.00
Product length	190								

Primer pair 2

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TTGGCACTGATAACACTCGCT	Plus	21	167	187	60.00	47.62	4.00	2.00
Reverse primer	GCCGTCAGGACAAGCAAAAG	Minus	20	358	339	60.04	55.00	4.00	0.00
Product length	192								

Primer pair 3

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CTTCTGGAACATACGAGGGCA	Plus	21	255	275	59.79	52.38	4.00	0.00
Reverse primer	GACGTGTTTTACGCCGTCAG	Minus	20	370	351	59.84	55.00	5.00	3.00
Product length	116								

Primer pair 4

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TTCTGGAACATACGAGGGCA	Plus	20	256	275	58.73	50.00	4.00	0.00
Reverse primer	CGTGTTTTACGCCGTCAGGA	Minus	20	368	349	60.94	55.00	5.00	2.00
Product length	113								

Primer pair 5

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	ACATACGAGGGCAATTCACCA	Plus	21	263	283	59.72	47.62	4.00	2.00
Reverse primer	TTTACGCCGTCAGGACAAGC	Minus	20	363	344	60.95	55.00	5.00	2.00
Product length	101								

Primer pair 6

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TCTTCTGGAACATACGAGGGC	Plus	21	254	274	59.52	52.38	4.00	2.00
Reverse primer	GTTTTACGCCGTCAGGACAAG	Minus	21	365	345	59.81	52.38	5.00	0.00
Product length	112								

Primer pair 7

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	TTCTGGAACATACGAGGGCAA	Plus	21	256	276	59.37	47.62	4.00	0.00
Reverse primer	ACGTGTTTTACGCCGTCAGG	Minus	20	369	350	61.22	55.00	5.00	1.00
Product length	114								

Primer pair 8

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CGAGGGCAATTCACCATTTCA	Plus	21	268	288	59.18	47.62	4.00	1.00
Reverse primer	GTTTTACGCCGTCAGGACAA	Minus	20	365	346	58.78	50.00	5.00	2.00
Product length	98								

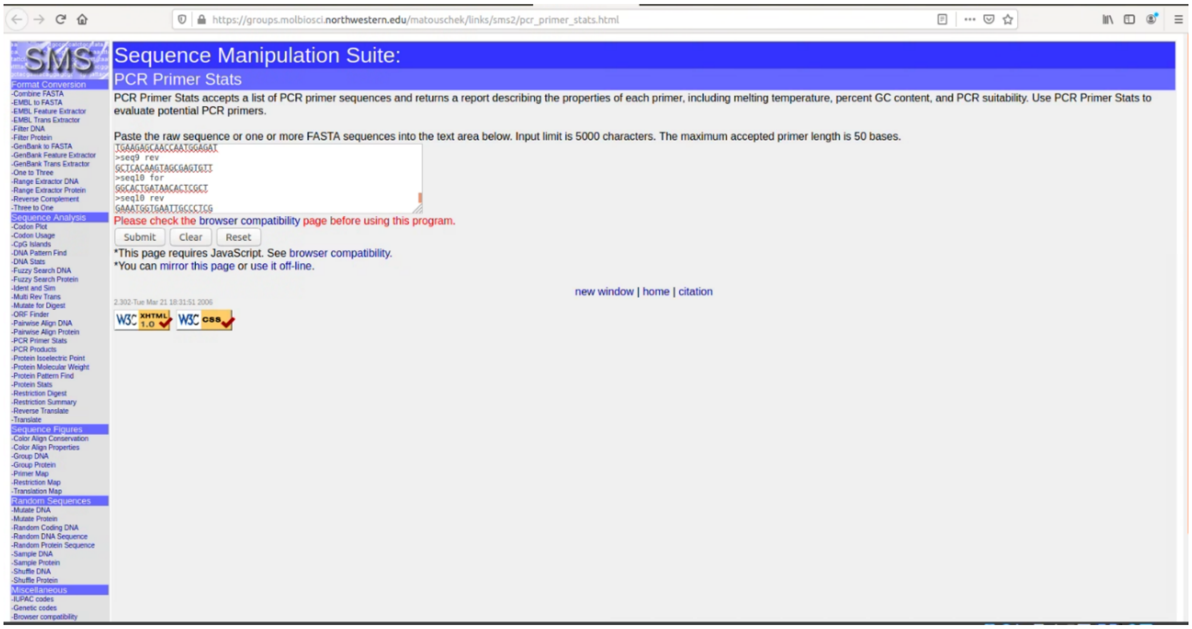
Primer pair 9

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	ACGAGGGCAATTCACCATTTC	Plus	21	267	287	59.18	47.62	4.00	1.00
Reverse primer	TTACGCCGTCAGGACAAGC	Minus	19	362	344	60.37	57.89	5.00	2.00
Product length	96								

Primer pair 10

	Sequence (5'->3')	Template strand	Length	Start	Stop	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	CACTCGCTACTTGTGAGCTT	Plus	20	180	199	57.93	50.00	5.00	2.00
Reverse primer	TGCCCTCGTATGTTCCAGAAG	Minus	21	275	255	59.79	52.38	4.00	2.00
Product length	96								

We then used Sequence Manipulation Suite tool (PSR Primer State) to validate the 10 primers of each conserved region and compare between them.



Sequence Manipulation Suite:

PCR Primer Stats

PCR Primer Stats accepts a list of PCR primer sequences and returns a report describing the properties of each primer, including melting temperature, percent GC content, and PCR suitability. Use PCR Primer Stats to evaluate potential PCR primers.

Paste the raw sequence or one or more FASTA sequences into the text area below. Input limit is 5000 characters. The maximum accepted primer length is 50 bases.

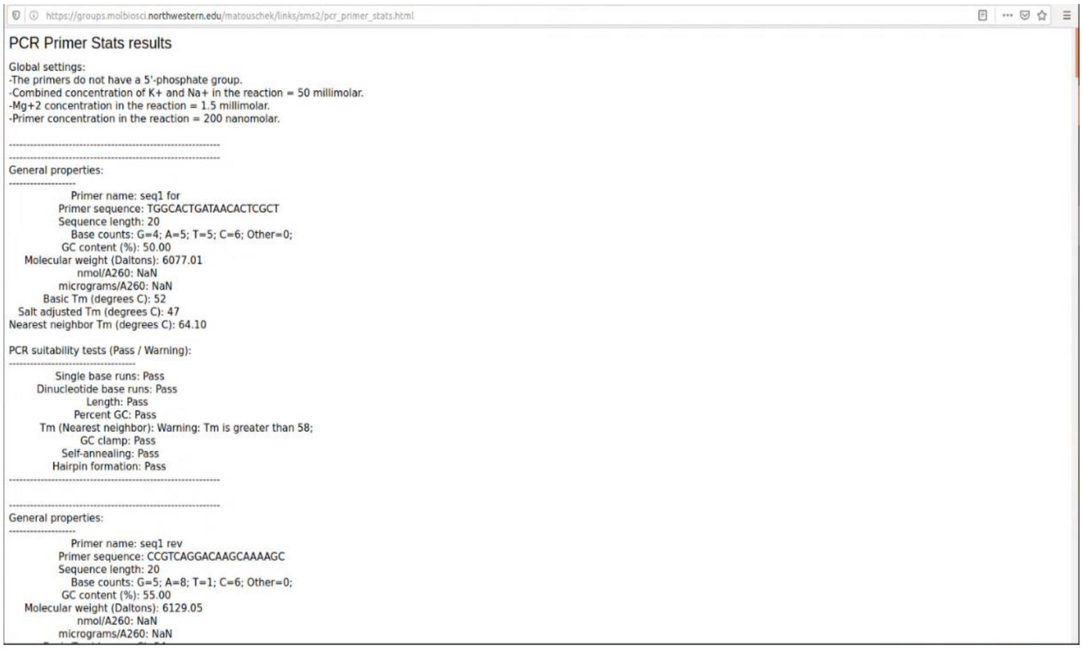
```
TGAAGAGCAACCAATGGAGAT
>seq9 rev
GCTCACAAGTAGCGAGTGTT
>seq10 for
GGCACTGATAACACTCGCT
>seq10 rev
GAAATGGTGAATTGCCCTCG
```

Please check the browser compatibility page before using this program.

*This page requires JavaScript. See browser compatibility.
*You can mirror this page or use it off-line.

2.302-Tue Mar 21 18:31:51 2006

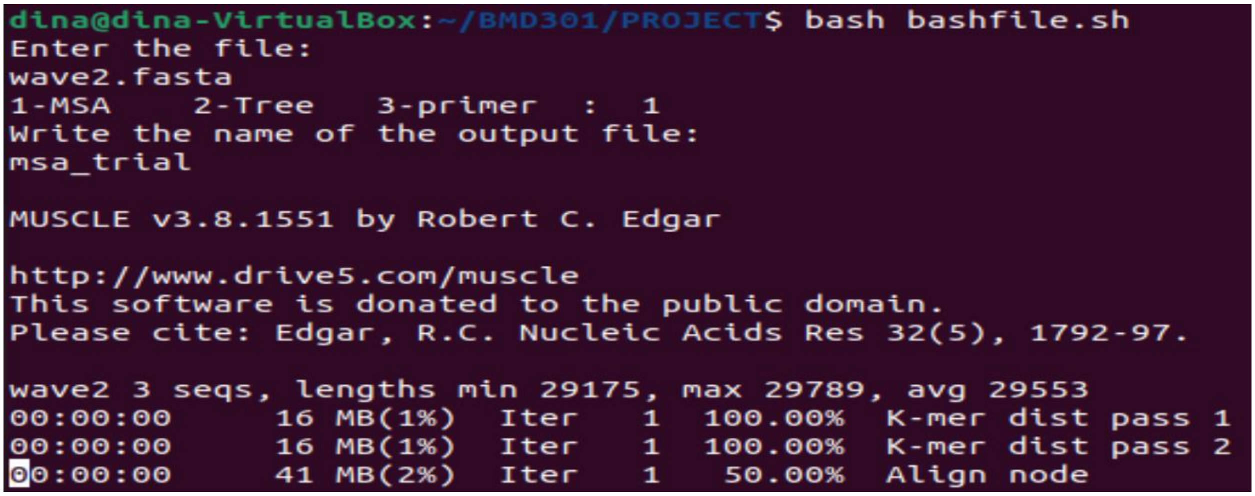
Although we chose different conserved regions throughout the multiple sequence alignment of the nine sequences, the primers designed for each one of them had the same results in all the criteria they are evaluated upon, where the best primer in each had only warning in the temperature part saying that "Tm is greater than 58", while the rest of the criteria such as GC clamp, self-annealing, and hairpin formation were good without any warnings.


https://groups.molbiosci.northwestern.edu/matouschek/links/sms2/pcr_primer_stats.html

```
PCR Primer Stats results

Global settings:
-The primers do not have a 5'-phosphate group.
-Combined concentration of K+ and Na+ in the reaction = 50 millimolar.
-Mg+2 concentration in the reaction = 1.5 millimolar.
-Primer concentration in the reaction = 200 nanomolar.

----------------------------------------------------------
----------------------------------------------------------
General properties:
------------------
              Primer name: seq1 for
           Primer sequence: TGGCACTGATAACACTCGCT
           Sequence length: 20
              Base counts: G=4; A=5; T=5; C=6; Other=0;
           GC content (%): 50.00
   Molecular weight (Daltons): 6077.01
               nmol/A260: NaN
           micrograms/A260: NaN
       Basic Tm (degrees C): 52
  Salt adjusted Tm (degrees C): 47
Nearest neighbor Tm (degrees C): 64.10

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
----------------------------------------------------------

----------------------------------------------------------
General properties:
------------------
              Primer name: seq1 rev
           Primer sequence: CCGTCAGGACAAGCAAAAGC
           Sequence length: 20
              Base counts: G=5; A=8; T=1; C=6; Other=0;
           GC content (%): 55.00
   Molecular weight (Daltons): 6129.05
               nmol/A260: NaN
           micrograms/A260: NaN
```

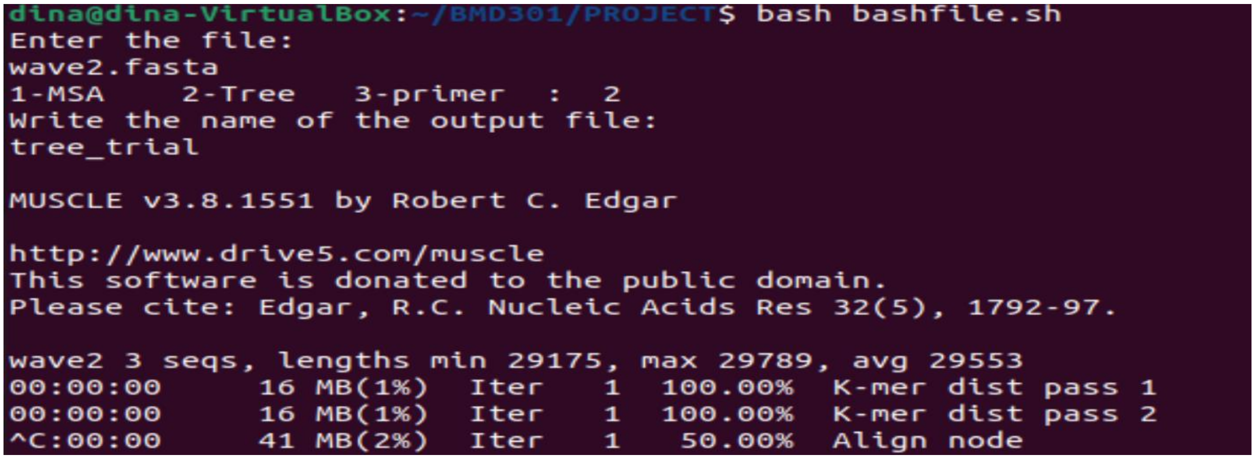
3.4. Automated Bash Script

In the bash script part, if the user chose the option of performing multiple sequence alignment on the file he/she desires, the output will be as the following resulting in a “file.afa”.



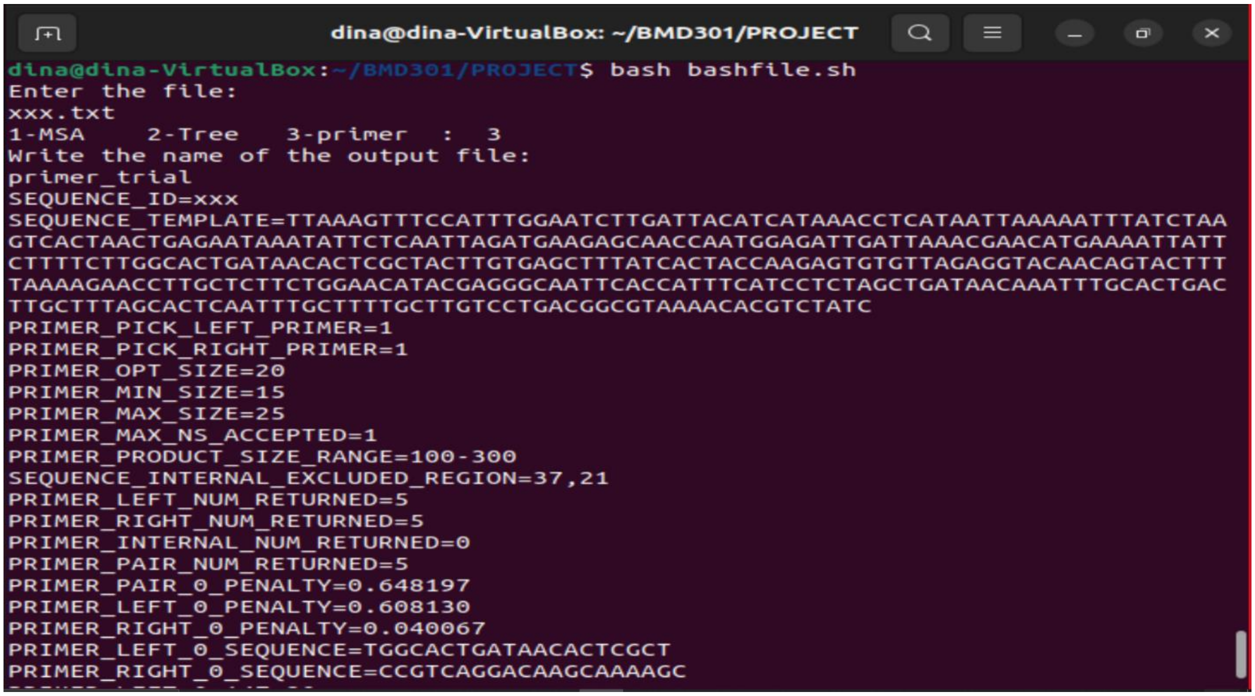
```
dina@dina-VirtualBox:~/BMD301/PROJECT$ bash bashfile.sh
Enter the file:
wave2.fasta
1-MSA    2-Tree   3-primer  :  1
Write the name of the output file:
msa_trial

MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

wave2 3 seqs, lengths min 29175, max 29789, avg 29553
00:00:00      16 MB(1%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00      16 MB(1%)  Iter   1  100.00%  K-mer dist pass 2
00:00:00      41 MB(2%)  Iter   1   50.00%  Align node
```

On the other hand, if option two was chosen, which is making a phylogenetic tree for the specified sequence, the output will be the following figure resulting in a “file.phy”.



```
dina@dina-VirtualBox:~/BMD301/PROJECT$ bash bashfile.sh
Enter the file:
wave2.fasta
1-MSA    2-Tree   3-primer  :  2
Write the name of the output file:
tree_trial

MUSCLE v3.8.1551 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

wave2 3 seqs, lengths min 29175, max 29789, avg 29553
00:00:00      16 MB(1%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00      16 MB(1%)  Iter   1  100.00%  K-mer dist pass 2
^C:00:00      41 MB(2%)  Iter   1   50.00%  Align node
```

Finally, if option three, which is designing the primer for the conserved region, is chosen then the output will be a number of primers designed for the region including the forward and reverse strands of each one of them as well as other parameters for each primer.



```
dina@dina-VirtualBox:~/BMD301/PROJECT$ bash bashfile.sh
Enter the file:
xxx.txt
1-MSA    2-Tree   3-primer  :  3
Write the name of the output file:
primer_trial
SEQUENCE_ID=xxx
SEQUENCE_TEMPLATE=TTAAAGTTTCCATTTGGAATCTTGATTACATCATAAACCTCATAATTAAAAATTTATCTAA
GTCACTAACTGAGAATAAATATTCTCAATTAGATGAAGAGCAACCAATGGAGATTGATTAAACGAACATGAAAATTATT
CTTTTCTTGGCACTGATAACACTCGCTACTTGTGAGCTTTATCACTACCAAGAGTGTGTTAGAGGTACAACAGTACTTT
TAAAAGAACCTTGCTCTTCTGGAACATACGAGGGCAATTCACCATTTCATCCTCTAGCTGATAACAAATTTGCACTGAC
TTGCTTTAGCACTCAATTTGCTTTTGCTTGTCCTGACGGCGTAAAACACGTCTATC
PRIMER_PICK_LEFT_PRIMER=1
PRIMER_PICK_RIGHT_PRIMER=1
PRIMER_OPT_SIZE=20
PRIMER_MIN_SIZE=15
PRIMER_MAX_SIZE=25
PRIMER_MAX_NS_ACCEPTED=1
PRIMER_PRODUCT_SIZE_RANGE=100-300
SEQUENCE_INTERNAL_EXCLUDED_REGION=37,21
PRIMER_LEFT_NUM_RETURNED=5
PRIMER_RIGHT_NUM_RETURNED=5
PRIMER_INTERNAL_NUM_RETURNED=0
PRIMER_PAIR_NUM_RETURNED=5
PRIMER_PAIR_0_PENALTY=0.648197
PRIMER_LEFT_0_PENALTY=0.608130
PRIMER_RIGHT_0_PENALTY=0.040067
PRIMER_LEFT_0_SEQUENCE=TGGCACTGATAACACTCGCT
PRIMER_RIGHT_0_SEQUENCE=CCGTCAGGACAAGCAAAAGC
```

4. Conclusion

After we made alignment, we concluded that there are multiple sequences similar to each other in the same wave but in different countries. In addition, when we specified the conserved regions from the multiple sequence alignment, it was also observed that the conserved region is considered as similar sequences in same wave but in different countries. Depending on that, when we made multiple primers for different conserved regions, this gave the same result, so we conducted that the alignment was done correctly.

5. References

[1] "Africa Faces Steepest COVID-19 Surge Yet." WHO | Regional Office for Africa, www.afro.who.int/news/africa-faces-steepest-covid-19-surge-yet. Accessed 30 Jan. 2023.

[2] Epicov.org, 2023, www.epicov.org/epi3/frontend#4c9d24.

[3] https://www.bioinformatics.org/sms2/

[4] National Human Genome Research Institute. "Bioinformatics." Genome.gov, 6 Sept. 2022, www.genome.gov/genetics-glossary/Bioinformatics.

[5] "New COVID-19 Variants Fuelling Africa's Second Wave." WHO | Regional Office for Africa, www.afro.who.int/news/new-covid-19-variants-fuelling-africas-second-wave.

[6] "PCR Primer Stats." Groups.molbiosci.northwestern.edu, groups.molbiosci.northwestern.edu/matouschek/links/sms2/pcr_primer_stats.html. Accessed 30 Jan. 2023.

[7] "Primer Designing Tool." Nih.gov, 2019, www.ncbi.nlm.nih.gov/tools/primer-blast/

[8] "Tree Viewer - Online Visualization of Phylogenetic Trees (Newick) and Alignments." Etetoolkit.org, etetoolkit.org/treeview/.

6. Team Members Contributions

- **Fatema Gamal Soliman:** Downloaded the sequences from different waves and performed multiple sequence alignment.
- **Dina Mohammed Sharwida:** Made the phylogenetic tree for the nine sequences downloaded and analyzed its results.
- **Dina Yahia Zahran:** Extracted conserved regions from multiple sequence alignment file and designed primers for such regions.
- **Alaa Hussein Mohamed:** Performed Primer validation among different conserved regions and compared between their results.
- **Nada Osama Fikry:** Automated bash script for multiple sequence alignment, phylogenetic tree and primer designing.