# A6_Tianyi_Zuo

Tianyi Zuo

2022/3/1

## GitHub username: Lydia12138

## Date: 2022-01-26
## Repo:https://github.com/Lydia12138/A6_Tianyi_Zuo (https://github.com/Lydia12138/A6_Tianyi_Zuo)

# load the packages

```
library(BiocManager)
library(genbankr)
library(Biostrings)
library(annotate) # pairwise alignments
library(ape) # Multiple Alignments
library(muscle) # Align the sequences
library(rentrez)
library(ggtree)# phylogenetic tree
library(dplyr)
library(ggplot2)
library(reshape2)
```

# Input the Sequence

```
# load the sequence : >human isolate, unknown sequence
UKSeq <-"ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCA
GTAACCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACTGCGTCTTGGTTCAC
CGCTCTCACTCAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGAC
CAAATTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCT
ACTACCTAGGAACTGGGCCAGAAGCTGGACTTCCCTATGGTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTT
GAATACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACTTCCTCAAGGAACAACATTG
CCAAAAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCCTCATCACGTAGTCGCAACAGTTCAAGAA
ATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCT
TGACAGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCT
GAGGCTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAAC
AAACCCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCC
CAGCGCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGACCTACACAGGTGCCATC
AAATTGGATGACAAAGATCCAAATTTCAAAGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAA
CAGAGCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCT
TCTTCCTGCTGCAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAA"
```

# Pairwise Alignments: BLAST

```
UKGbkBLAST<-blastSequences(paste(UKSeq),as = 'data.frame',
                           hitListSize = 20, timeout = 600)
```

```
## estimated response time 6 seconds
```

```
## elapsed time 6 seconds
```

```
## elapsed time 17 seconds
```

# Mutiple Alignment

```
# Make a simple vector of accession numbers from the BLAST results above and make the
m into a simple data frame object with two columns
UKHitsDF<-data.frame(ID=UKGbkBLAST$Hit_accession,Seq=UKGbkBLAST$Hsp_hseq,
                     stringsAsFactors = FALSE)
head(UKHitsDF)
```

```
##         ID
## 1 OM797449
## 2 OM793753
## 3 OM779898
## 4 OM766143
## 5 OM766139
## 6 OM766138
##
Seq
## 1 ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAA
CCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACTGCGTCTTGGTTCACCGCT
CTCACTCAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAA
TTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTA
CCTAGGAACTGGGCCAGAAGCTGGACTTCCCTATGGTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAAT
ACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACTTCCTCAAGGAACAACATTGCCAA
AAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCCTCATCACGTAGTCGCAACAGTTCAAGAAATTC
AACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGAC
AGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCTGAGG
CTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAAAC
CCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGC
GCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGACCTACACAGGTGCCATCAAAT
TGGATGACAAAGATCCAAATTTCAAAGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGA
GCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTT
CCTGCTGCAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAA
## 2 ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAA
CCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACTGCGTCTTGGTTCACCGCT
CTCACTCAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAA
TTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTA
CCTAGGAACTGGGCCAGAAGCTGGACTTCCCTATGGTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAAT
ACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACTTCCTCAAGGAACAACATTGCCAA
AAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCCTCATCACGTAGTCGCAACAGTTCAAGAAATTC
AACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGAC
AGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCTGAGG
CTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAAAC
CCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGC
GCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGACCTACACAGGTGCCATCAAAT
TGGATGACAAAGATCCAAATTTCAAAGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGA
GCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTT
CCTGCTGCAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAA
## 3 ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAA
CCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACTGCGTCTTGGTTCACCGCT
CTCACTCAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAA
TTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTA
CCTAGGAACTGGGCCAGAAGCTGGACTTCCCTATGGTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAAT
ACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACTTCCTCAAGGAACAACATTGCCAA
AAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCCTCATCACGTAGTCGCAACAGTTCAAGAAATTC
AACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGAC
AGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCTGAGG
CTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAAAC
CCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGC
GCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGACCTACACAGGTGCCATCAAAT
TGGATGACAAAGATCCAAATTTCAAAGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGA
GCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTT
CCTGCTGCAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAA
## 4 ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAA
```

```
CCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACTGCGTCTTGGTTCACCGCT
CTCACTCAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAA
TTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTA
CCTAGGAACTGGGCCAGAAGCTGGACTTCCCTATGGTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAAT
ACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACTTCCTCAAGGAACAACATTGCCAA
AAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCCTCATCACGTAGTCGCAACAGTTCAAGAAATTC
AACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGAC
AGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCTGAGG
CTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAAAC
CCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGC
GCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGACCTACACAGGTGCCATCAAAT
TGGATGACAAAGATCCAAATTTCAAAGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGA
GCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTT
CCTGCTGCAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAA
## 5 ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAA
CCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACTGCGTCTTGGTTCACCGCT
CTCACTCAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAA
TTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTA
CCTAGGAACTGGGCCAGAAGCTGGACTTCCCTATGGTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAAT
ACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACTTCCTCAAGGAACAACATTGCCAA
AAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCCTCATCACGTAGTCGCAACAGTTCAAGAAATTC
AACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGAC
AGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCTGAGG
CTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAAAC
CCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGC
GCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGACCTACACAGGTGCCATCAAAT
TGGATGACAAAGATCCAAATTTCAAAGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGA
GCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTT
CCTGCTGCAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAA
## 6 ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAA
CCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACTGCGTCTTGGTTCACCGCT
CTCACTCAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAA
TTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTA
CCTAGGAACTGGGCCAGAAGCTGGACTTCCCTATGGTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAAT
ACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACTTCCTCAAGGAACAACATTGCCAA
AAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCCTCATCACGTAGTCGCAACAGTTCAAGAAATTC
AACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGAC
AGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCTGAGG
CTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAAAC
CCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGC
GCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGACCTACACAGGTGCCATCAAAT
TGGATGACAAAGATCCAAATTTCAAAGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGA
GCCTAAAAAGGACAAAAAGAAGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTT
CCTGCTGCAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAA
```

ID of each sequences is their BLAST hit accession. We can fins all the sequences that the BLAST finds are virtually the same, that means that there are multiple accessions in NCBI for the same DNA sequence

```
## check the length of each sequence
UKGbkBLAST$Hit_len
```

```
##  [1] "29777" "29829" "29792" "29815" "29799" "29817" "29882" "29785" "29791"
## [10] "29801" "29797" "29815" "29821" "29903" "29903" "29782" "29782" "29782"
## [19] "29782" "29782"
```

# Determine the species of the sequence

```
#check the species of each sequence with their hit accession from Genebank.
UKHitSeqs<-read.GenBank(UKGbkBLAST$Hit_accession)
attr(UKHitSeqs,"species")
```

```
##  [1] "Severe_acute_respiratory_syndrome_coronavirus_2"
##  [2] "Severe_acute_respiratory_syndrome_coronavirus_2"
##  [3] "Severe_acute_respiratory_syndrome_coronavirus_2"
##  [4] "Severe_acute_respiratory_syndrome_coronavirus_2"
##  [5] "Severe_acute_respiratory_syndrome_coronavirus_2"
##  [6] "Severe_acute_respiratory_syndrome_coronavirus_2"
##  [7] "Severe_acute_respiratory_syndrome_coronavirus_2"
##  [8] "Severe_acute_respiratory_syndrome_coronavirus_2"
##  [9] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [10] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [11] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [12] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [13] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [14] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [15] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [16] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [17] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [18] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [19] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [20] "Severe_acute_respiratory_syndrome_coronavirus_2"
```

According to the length info and gene bank species result, we can identified this isolated sequence is identified as coronavirus 2 strain which relative with severe acute respiratory syndrome.

```
# Conduct the DNA mutiple Alignment
CVHitsDNAstring <- UKHitsDF$Seq %>% # Start with the sequences
  as.character %>% # Convert to strings
  lapply(.,paste0,collapse="") %>% # Collapse each sequence to a single string
  unlist %>% # Flatten list to a vector
  DNAStringSet # Convert vector to DNAStringSet object

names(CVHitsDNAstring)<-paste(1:nrow(UKHitsDF),UKHitsDF$ID,sep="_") #Give each sequen
ce a unique names

CVAlign<-muscle::muscle(stringset=CVHitsDNAstring, quiet=T)

CVAlign
```

```
## DNAMultipleAlignment with 20 rows and 1260 columns
##       aln                                                  names
##  [1] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 1_OM797449
##  [2] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 2_OM793753
##  [3] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 3_OM779898
##  [4] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 4_OM766143
##  [5] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 5_OM766139
##  [6] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 6_OM766138
##  [7] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 7_OM765571
##  [8] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 8_OM765560
##  [9] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 9_OM765512
##  ... ...
## [12] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 12_OM765468
## [13] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 13_OM765461
## [14] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 14_OV888164
## [15] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 15_OV886263
## [16] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 16_OM757013
## [17] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 17_OM756986
## [18] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 18_OM756962
## [19] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 19_OM756961
## [20] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 20_OM756959
```

# Check the Aligment

```
# According to the Aligment graph, there is no large gaps exist. Here check the seque
nce again to make sure there is no big gaps.
SeqLen<-as.numeric(lapply(CVHitsDNAstring,length))
qplot(SeqLen)+theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

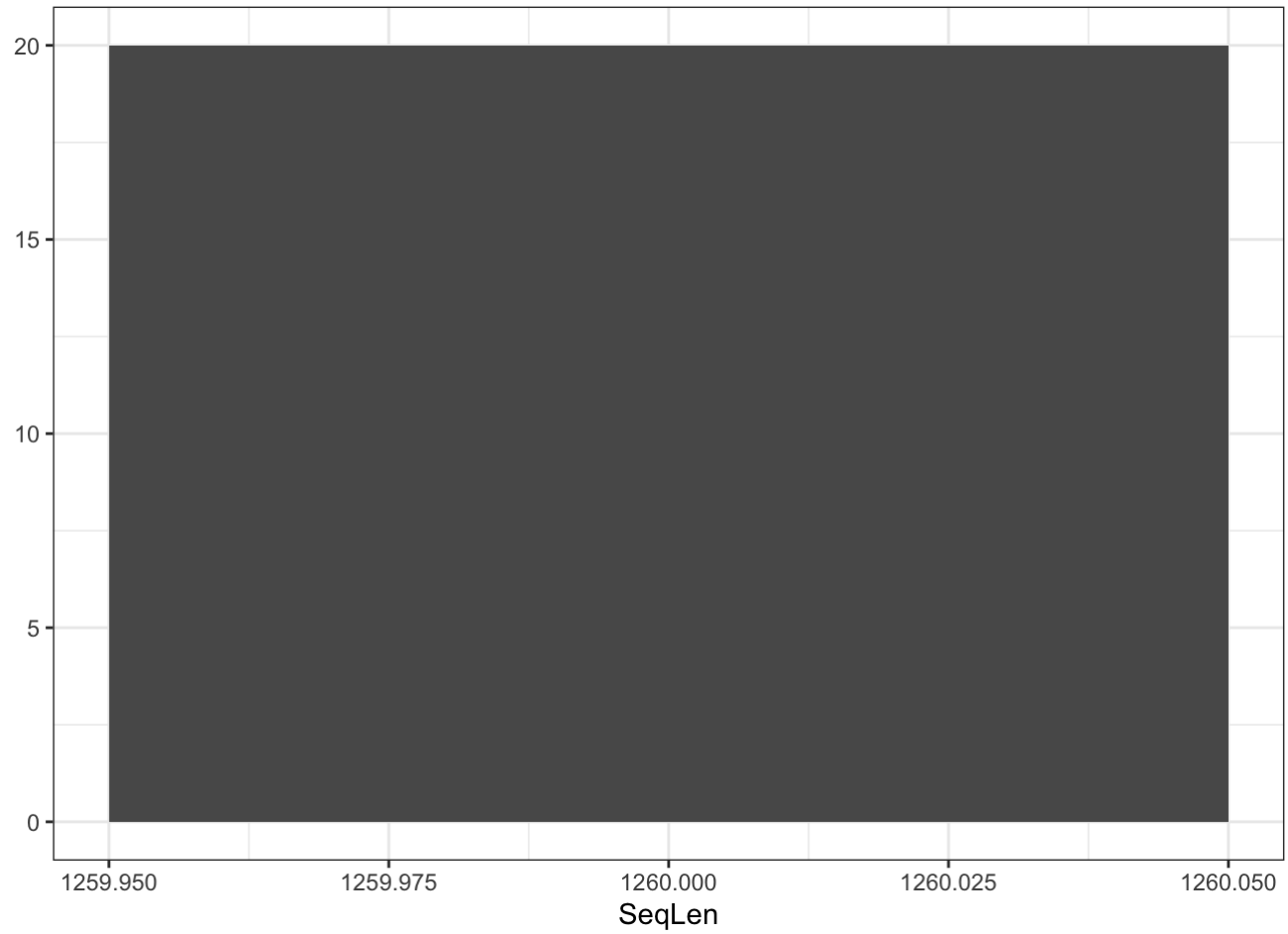


Figure 1. The bar graph shows the the distribution of the Sequences length.

From the alignment result and the distribution, it looks like there is neither large gap nor any new sequence insertions over all the 20 subject sequences. This step shows that there is no need to remove any sequence fragment.

# Distance Matrix

```
# Convert the DNAMultipleAlignment object into DNAbin
CVAlign <- as.DNAbin(CVAlign)
CVDM<-dist.dna(CVAlign, model="K80")

CVDMmat<-as.matrix(CVDM)

PDat<-melt(CVDMmat)


ggplot(data = PDat, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()+scale_fill_gradientn(colours=c("white","blue","green","red"))+
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```

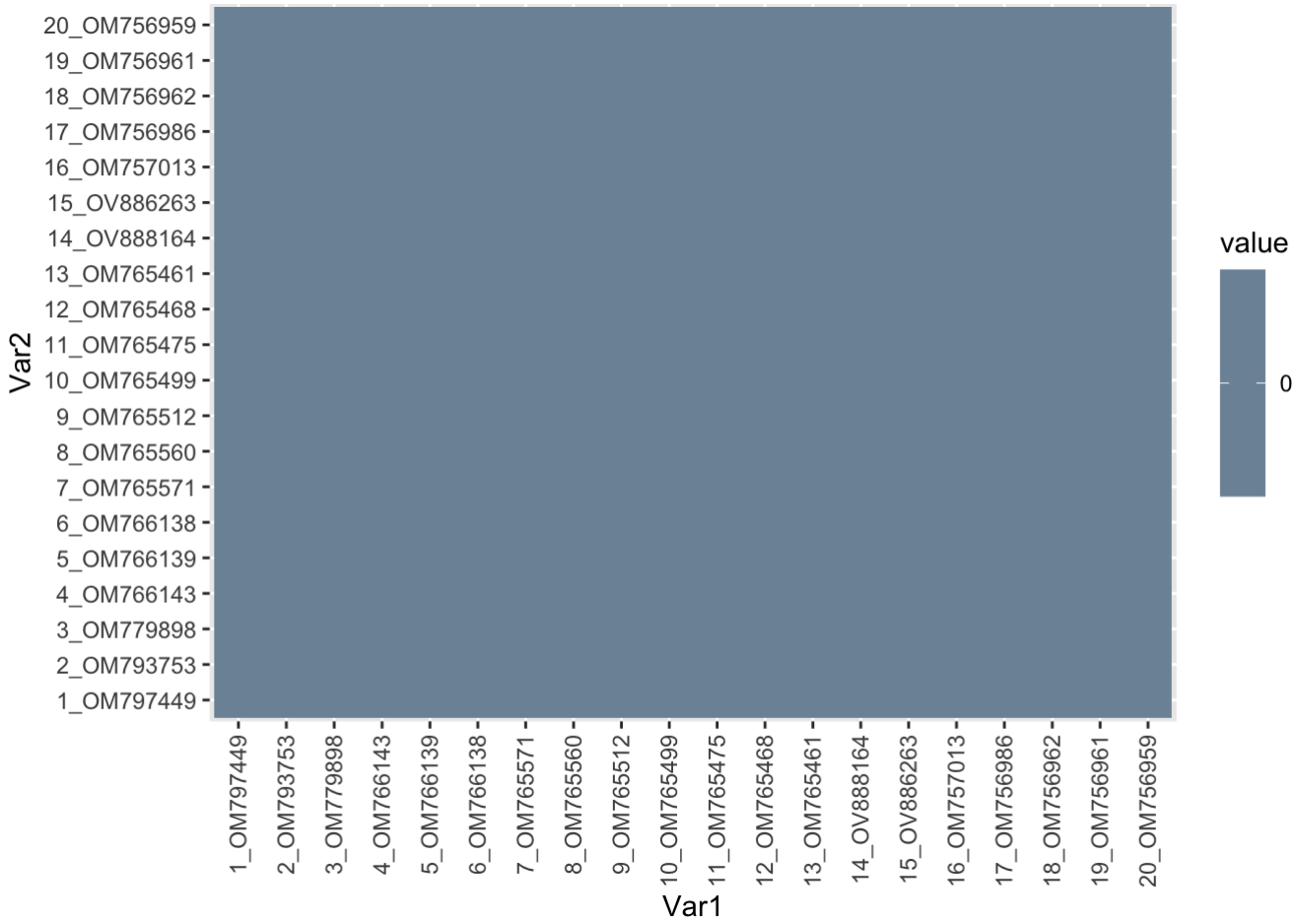


Figure 2. Pairwise distance matrices for 20 genes using data from the GeneBank.X-axis and Y-axis represent gene accession, and colour represent the different distance.

From the graph, we can find that there is no any distance among the 20 subject sequence. That means all of these sequence from same species and share same distance with each other.

# Build Phylogeny

```
CVTree<-nj(CVDM)
ggtree(CVTree)
```

Figure 3. Phylogenetic tree of 20 viruse sequences filter out human DNA constructed by a neighbor-joining method.

There not exist any branch in Phylogenetic tree, because the branch lengths in the above graph are based on the pairwise distance matrix. According to the Phylogenetic tree and distance matrix, it indicates that these sequences are closely related and fall into same taxon.

```
#remove the branch length info to focus on the relationships
ggtree(CVTree,branch.length='none')+ geom_tiplab()
```

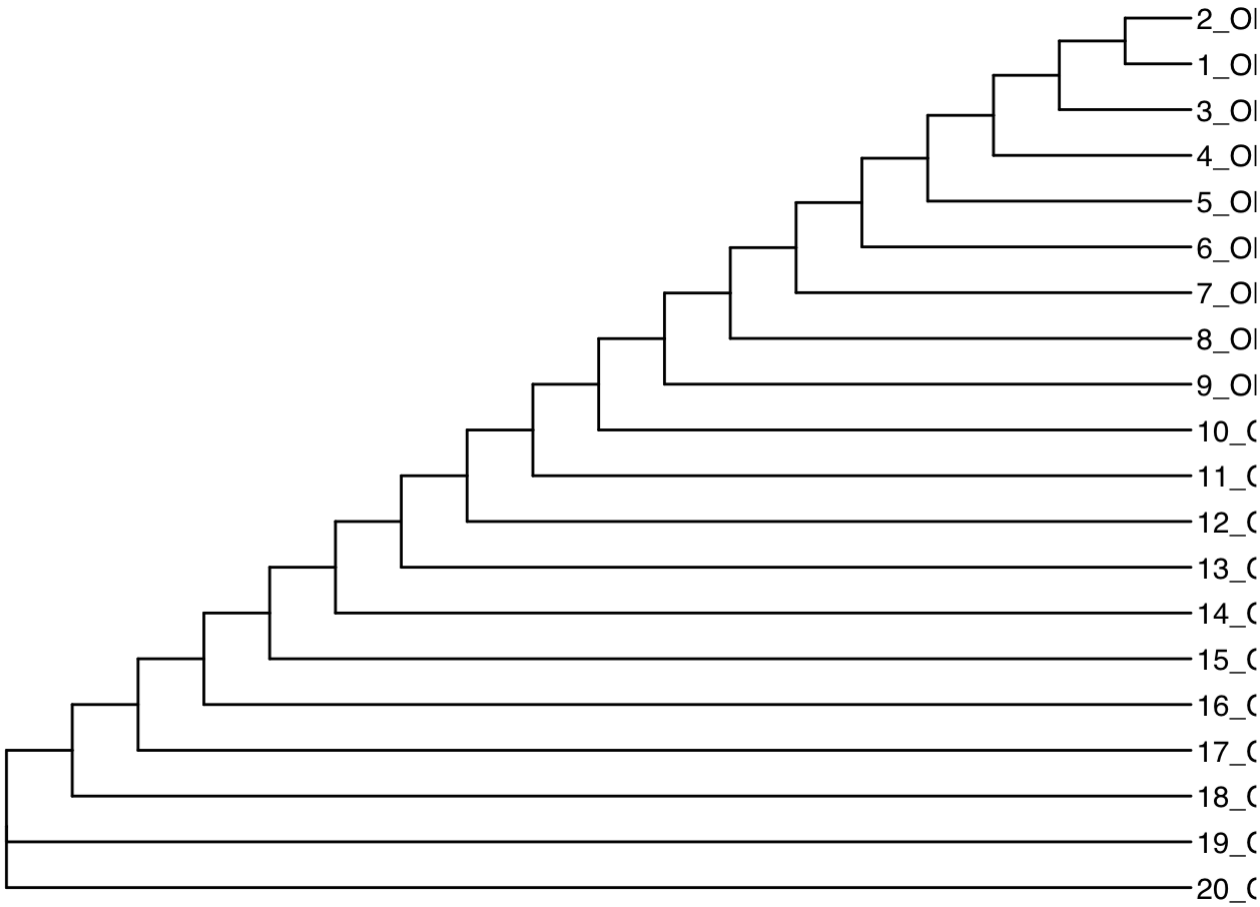


Figure 4. Phylogenetic tree of Viruse filters out human DNA constructed by a neighbor-joining method without brach length.

Figure 4 remove the branch length info to focus on the relationships. It suggested clear relationship between these 20 sequence, which fall into same strain with a number of differences.

```
# Because having trouble reading the labels, here exporting to a pdf file
pdf("A6_Tianyi_Zuo_Cov2_Virus_tree.pdf",width=8,height=4)
ggtree(CVTree,branch.length='none',layout="circular") + geom_tiplab()
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
# save the tree
write.tree(CVTree,"A6_Tianyi_Zuo_Cov2_Virus_tree.tre")
```