A6_LI_ZHIJUN_Alignments

Zhijun Li

2022-03-01

GitHub Link: https://github.com/zazauwu/A6_Alignments

```
# load the required packages
library(dplyr)
library(BiocManager)
library(Biostrings)
library(genbankr)
library(rentrez)
```

Sequence Analysis

Input the sequence

```
#load the human isolate, unknown sequence
unknseq <- "ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAA
CTGGCAGTAACCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAA
TAATACTGCGTCTTGGTTCACCGCTCTCACTCAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGACAA
GGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAATTGGCTACTACCGAAGAGCTACCAGACGAA
TTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTACCTAGGAACTGGGCC
AGAAGCTGGACTTCCCTATGGTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAAT
ACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCTGCAATCGTGCTACAACTTCCTCAAG
GAACAACATTGCCAAAAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCCTC
ATCACGTAGTCGCAACAGTTCAAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATG
GCTGGCAATGGCGGTGATGCTGCTCTTGCTTTGCTGCTGCTTGACAGATTGAACCAGCTTGAGAGCAAAA
TGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCTGCTGCTGAGGCTTCTAAGAA
GCCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAA
CAAACCCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAA
TTGCACAATTTGCCCCCAGCGCTTCAGCGTTCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTC
GGGAACGTGGTTGACCTACACAGGTGCCATCAAATTGGATGACAAAGATCCAAATTTCAAAGATCAAGTC
ATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGA
AGAAGGCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTTCCTGC
TGCAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTGCTGACTCAACTCAGGCCTAA"

#use regular expression 'gsub' to remove the 'carriage return' and 'newline' special character
unknseq <- gsub("[\r\n]", "", unknseq)

unknseq
```

```
## [1] "ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAACCAGAATGGAGAACG
```

Generate alignments

```r
# pairwise alignments
library(annotate)
```

```
## Loading required package: AnnotationDbi

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Attaching package: 'AnnotationDbi'

## The following object is masked from 'package:dplyr':
##
##     select

## Loading required package: XML
```

```r
useqBLAST <- blastSequences(paste(unknseq),as = 'data.frame',
                            hitListSize = 20, timeout = 600)
```

```
## estimated response time 46 seconds
```

```r
# multiple alignments
library(ape)
```

```
##
## Attaching package: 'ape'

## The following object is masked from 'package:Biostrings':
##
##     complement
```

```r
# create a DNAbin object
useqHitsDF <- data.frame(ID = useqBLAST$Hit_accession, # specifying an ID column
                         Seq = useqBLAST$Hsp_hseq,
                         stringsAsFactors = FALSE)
```

```r
# length of each sequence
useqBLAST$Hit_len
```

```
##  [1] "29831" "29800" "29782" "29782" "29782" "29782" "29782" "29782" "29782"
## [10] "29782" "29801" "29816" "29793" "29903" "29903" "29903" "29903" "29903"
## [19] "29903" "29903"
```

The 20 sequences have similar number of base pairs.

Determine if it is human or other organism

```r
# read in the obtained 20 sequences from GenBank using the read.Genbank()
useqHitSeqs <- read.GenBank(useqBLAST$Hit_accession)

# take a look at the species
attr(useqHitSeqs,"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"
```

The isolated sequence is identified as Severe acute respiratory syndrome-related coronavirus 2 instead of human.

```r
# convert DNAbin to a DNAStringSet for an alignment
library(Biostrings)
CovDNAstring <- useqHitsDF$Seq %>%
  as.character %>% # convert to strings
  lapply(., paste0, collapse = "") %>%  # collaspe each sequence to a single string
  unlist %>% # flatten list to a vector
  DNAStringSet # convert the vector to the required DNAStringSet object
```

```r
# give each sequence a unique names
names(CovDNAstring) <- paste(1:nrow(useqHitsDF),useqHitsDF$ID,sep="_")
```

```r
# use MUSCLE (MUltiple Sequence Comparison by Log-Expectation) to align the sequences
library(muscle)
```

```
##
## Attaching package: 'muscle'
```

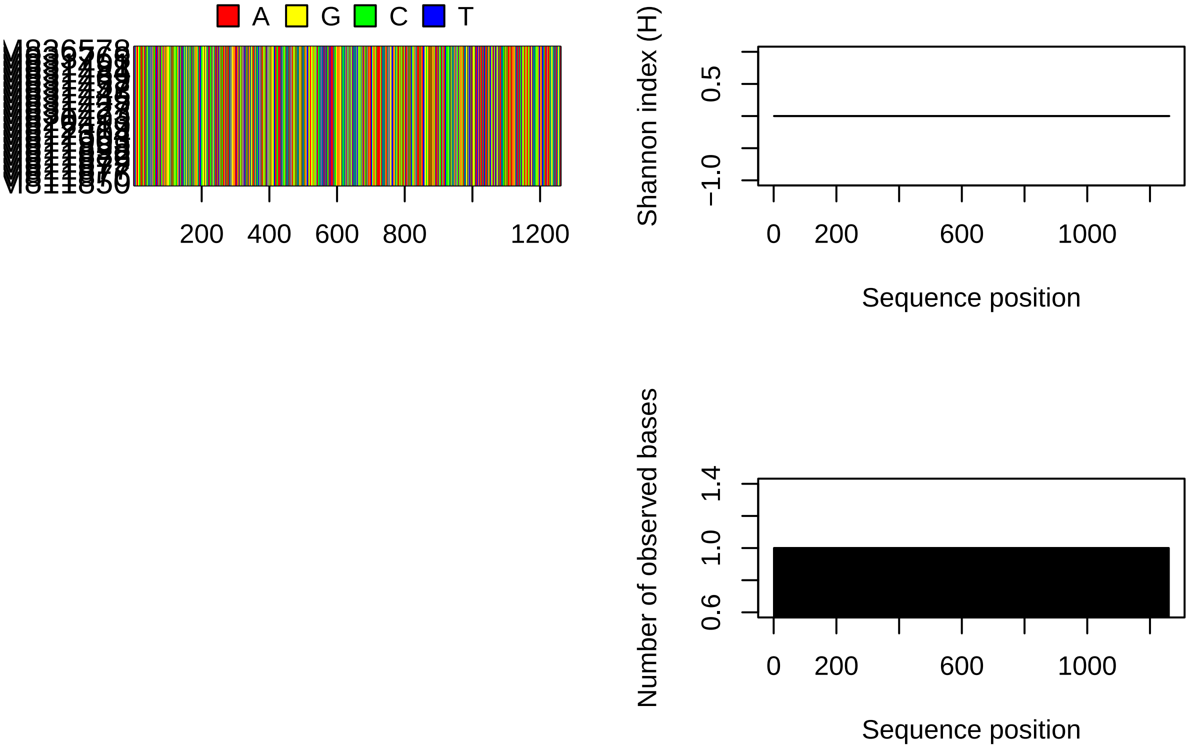
```
## The following object is masked from 'package:ape':
##
##     muscle
```

```r
# create a DNAMultipleAlignment object
CovAlign <- muscle::muscle(stringset = CovDNAstring, quiet = T)

CovAlign
```

```
## DNAMultipleAlignment with 20 rows and 1260 columns
##       aln                                                  names
##  [1] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 1_OM836578
##  [2] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 2_OM833768
##  [3] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 3_OM831491
##  [4] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 4_OM831484
##  [5] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 5_OM831469
##  [6] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 6_OM831457
##  [7] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 7_OM831448
##  [8] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 8_OM831445
##  [9] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 9_OM831427
##  ... ...
## [12] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 12_OM812419
## [13] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 13_OM812304
## [14] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 14_OM811903
## [15] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 15_OM811898
## [16] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 16_OM811882
## [17] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 17_OM811879
## [18] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 18_OM811877
## [19] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 19_OM811876
## [20] ATGTCTGATAATGGACCCCAAAATC...CAGTGCTGACTCAACTCAGGCCTAA 20_OM811850
```

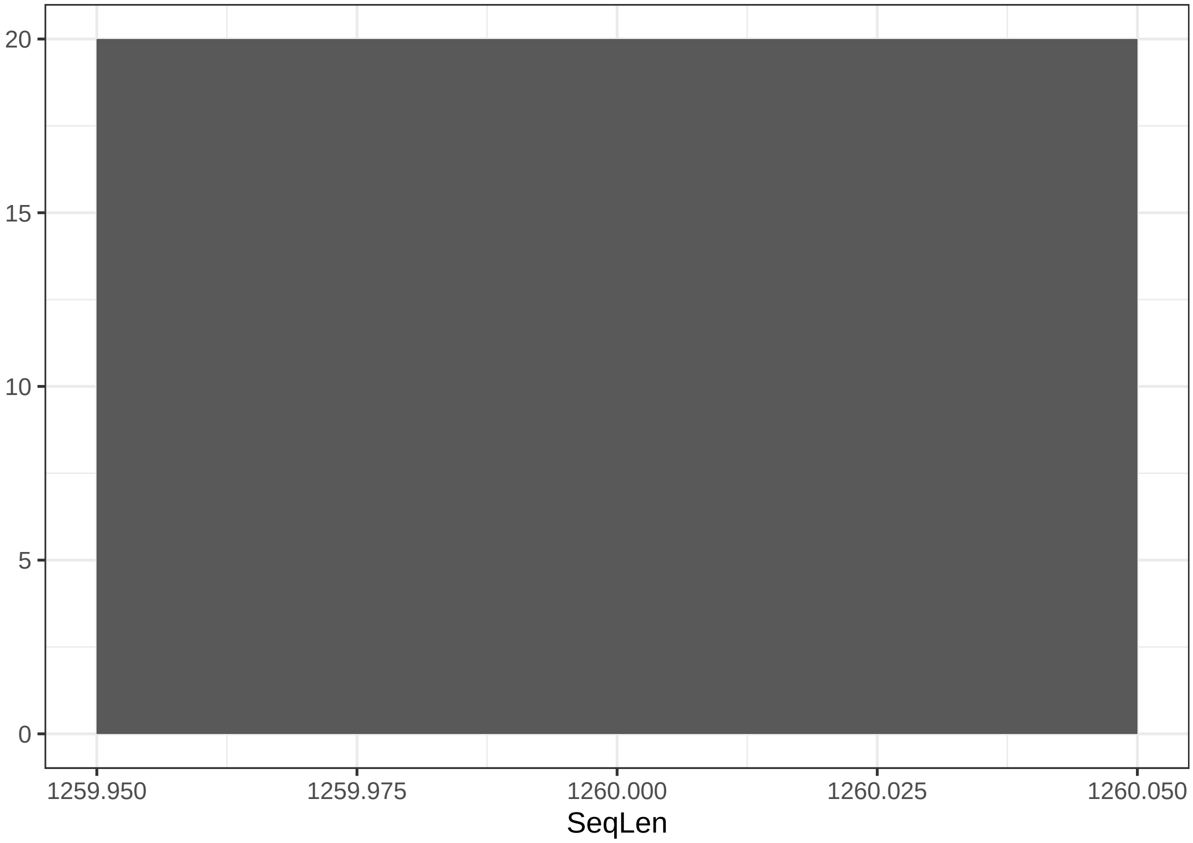
```
# convert the DNAMultipleAlignment object to a DNAbin
CovAlignBin <- as.DNAbin(CovAlign)

#perform a series of diagnostics on a DNA alignement
checkAlignment(CovAlignBin)
```

```
##
## Number of sequences: 20
## Number of sites: 1260
##
## No gap in alignment.
##
## Number of segregating sites (including gaps): 0
## Number of sites with at least one substitution: 0
## Number of sites with 1, 2, 3 or 4 observed bases:
##    1    2    3    4
## 1260    0    0    0
```



```r
SeqLen <- as.numeric(lapply(CovDNAstring, length))

# plot the distribution of sequence length
library(ggplot2)
qplot(SeqLen) + theme_bw()
```

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



Based on the alignment and the distribution, it looks like there is neither distinct gap nor substitution across all the 20 subject sequences, hence, there is no need to remove any sequence.

Build a phylogeny

```
# make distance matrix for tree
CDM <- dist.dna(CovAlignBin, model = "K80")
CDMmat <- as.matrix(CDM)

# rearrange CDMmat to a 'linear' matrix
library(reshape2)
PDat <- melt(CDMmat)

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))
```

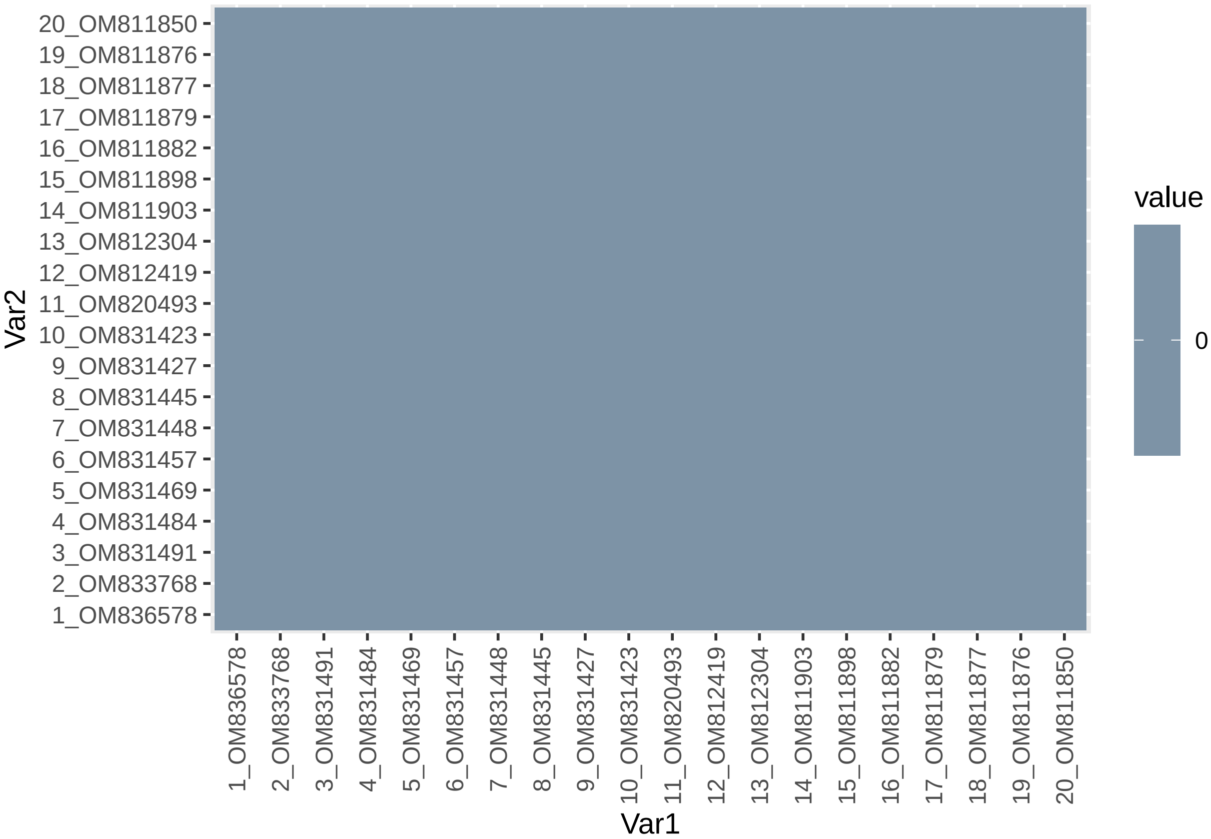


Figure 1. The variation among the 20 subject sequences based on distance matrix. All of them share complete similarity with each other.

Report

Given the results from BLAST and alignment, it should be concerned that the human isolated sequence is from the coronavirus SARS-CoV-2 which can trigger serve respiratory symptoms.

```
# create a phylogenetic tree using the Neighbour-Joining (NJ) approach
CovTree <- nj(CDMmat)

# plot the phylogenetic tree
library(ggtree)
```

```
## ggtree v3.2.1  For help: https://yulab-smu.top/treedata-book/
##
## If you use ggtree in published research, please cite the most appropriate paper(s):
##
## 1. Guangchuang Yu. Using ggtree to visualize data on tree-like structures. Current Protocols in Bioin
## 2. Guangchuang Yu, Tommy Tsan-Yuk Lam, Huachen Zhu, Yi Guan. Two methods for mapping and visualizing
## 3. Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtree: an R package for vi
```

```
##
## Attaching package: 'ggtree'
```

```
## The following object is masked from 'package:ape':
##
##     rotate
## The following object is masked from 'package:Biostrings':
##
##     collapse
```

```
## The following object is masked from 'package:IRanges':
##
##     collapse
```

```
## The following object is masked from 'package:S4Vectors':
##
##     expand
```

```
ggtree(CovTree)
```

Figure 2. Phylogenetic tree of the 20 sequences. It suggests that these sequences are closely related and fall into the same strain.

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

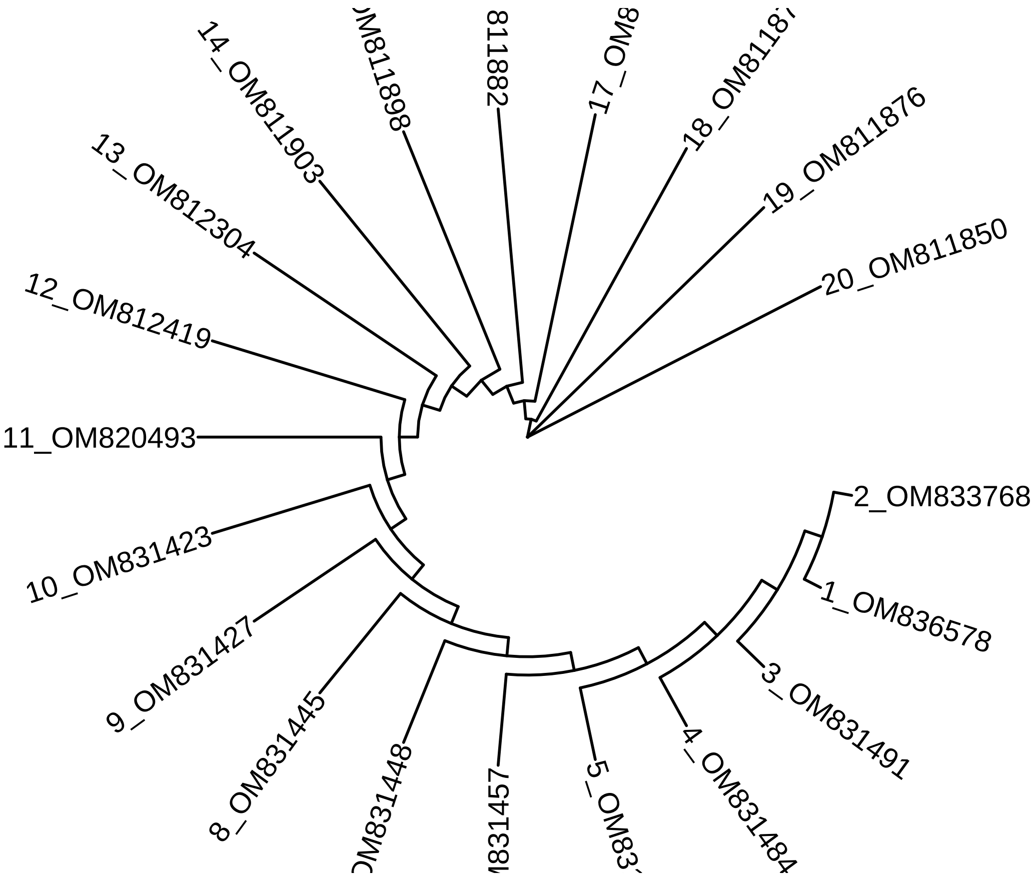


Figure 3. Relationship among the 20 sequences. It suggests that each sequence, though of the same SARS-CoV-2 virus strain, contains a number of differences.

```
# save the tree
write.tree(CovTree,"A6_LI_ZHIJUN_Cov2_tree.tre")
```