DNA Alignments Assignment

Chenyang Wu

2022/3/2

Project Info

GitHub user name: Wuris

Date: 2022/3/2

GitHub Link: https://github.com/Wuris/Biol432_A6.git (https://github.com/Wuris/Biol432_A6.git)

```
# Load the packages we need
library(dplyr)
```

```
##
## 载入程辑包：'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(BiocManager)
library(genbankr)
library(annotate)
```

```
## 载入需要的程辑包：AnnotationDbi
```

```
## 载入需要的程辑包：stats4
```

```
## 载入需要的程辑包：BiocGenerics
```

```
##
## 载入程辑包：'BiocGenerics'
```

```
## The following objects are masked from 'package:dplyr':
##
##     combine, intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min
```

```
## 载入需要的程辑包：Biobase
```

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

```
## 载入需要的程辑包：IRanges
```

```
## 载入需要的程辑包：S4Vectors
```

```
##
## 载入程辑包：'S4Vectors'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, rename
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
##
## 载入程辑包：'IRanges'
```

```
## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice
```

```
## The following object is masked from 'package:grDevices':
##
##     windows
```

```
##
## 载入程辑包：'AnnotationDbi'
```

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

```
## 载入需要的程辑包：XML
```

```
library(ape)
library(Biostrings)
```

```
## 载入需要的程辑包：XVector
```

```
## 载入需要的程辑包：GenomeInfoDb
```

```
##
## 载入程辑包：'Biostrings'
```

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

```
## The following object is masked from 'package:base':
##
##     strsplit
```

```
library(muscle)
```

```
##
## 载入程辑包：'muscle'
```

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

```
library(ggplot2)
library(reshape2)
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 Bioinformatics. 2020, 69:e96. doi:10.1002/cpbi.96
## 2. Guangchuang Yu, Tommy Tsan-Yuk Lam, Huachen Zhu, Yi Guan. Two methods for mapping and visualizing associated data on phylogeny using ggtree. Molecular Biology and Evolution. 2018, 35(12):3041-3043. doi:10.1093/molbev/msy194
## 3. Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017, 8(1):28-36. doi:10.1111/2041-210X.12628
```

```
##
## 载入程辑包：'ggtree'
```

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

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

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

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

Input the unknown sequence

```
# The Human isolate, unknown sequence
UnSeq <- "ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAACCAGA
ATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCCAAGGTTTACCCAATAATACTGCGTCTTGGTTCACCGCTCTCACTCAACATGGC
AAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTTCCAATTAACACCAATAGCAGTCCAGATGACCAAATTGGCTACTACCGAAGAGCTACCAG
ACGAATTCGTGGTGGTGACGGTAAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTACCTAGGAACTGGGCCAGAAGCTGGACTTCCCTATG
GTGCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAATACACCAAAAGATCACATTGGCACCCGCAATCCTGCTAACAATGCT
GCAATCGTGCTACAACTTCCTCAAGGAACAACATTGCCAAAAGGCTTCTACGCAGAAGGGAGCAGAGGCGGCAGTCAAGCCTCTTCTCGTTCCTC
ATCACGTAGTCGCAACAGTTCAAGAAATTCAACTCCAGGCAGCAGTAGGGGAACTTCTCCTGCTAGAATGGCTGGCAATGGCGGTGATGCTGCTC
TTGCTTTGCTGCTGCTTGACAGATTGAACCAGCTTGAGAGCAAAATGTCTGGTAAAGGCCAACAACAACAAGGCCAAACTGTCACTAAGAAATCT
GCTGCTGAGGCTTCTAAGAAGCCTCGGCAAAAACGTACTGCCACTAAAGCATACAATGTAACACAAGCTTTCGGCAGACGTGGTCCAGAACAAAC
CCAAGGAAATTTTGGGGACCAGGAACTAATCAGACAAGGAACTGATTACAAACATTGGCCGCAAATTGCACAATTTGCCCCCAGCGCTTCAGCGT
TCTTCGGAATGTCGCGCATTGGCATGGAAGTCACACCTTCGGGAACGTGGTTGACCTACACAGGTGCCATCAAATTGGATGACAAAGATCCAAAT
TTCAAAGATCAAGTCATTTTGCTGAATAAGCATATTGACGCATACAAAACATTCCCACCAACAGAGCCTAAAAAGGACAAAAAGAAGAAGGCTGA
TGAAACTCAAGCCTTACCGCAGAGACAGAAGAAACAGCAAACTGTGACTCTTCTTCCTGCTGCAGATTTGGATGATTTCTCCAAACAATTGCAAC
AATCCATGAGCAGTGCTGACTCAACTCAGGCCTAA"
```

Generate the alignment

```
# Pairwise Alignments
UnSeqBLAST <- blastSequences(paste(UnSeq),
                             as = 'data.frame',
                             hitListSize = 20,
                             timeout = 600)
```

```
## estimated response time 34 seconds
```

```
## elapsed time 35 seconds
```

```
## elapsed time 46 seconds
```

```
## elapsed time 56 seconds
```

```
## elapsed time 67 seconds
```

```
## elapsed time 78 seconds
```

```
## elapsed time 89 seconds
```

```
## elapsed time 100 seconds
```

```
## elapsed time 111 seconds
```

```
## elapsed time 122 seconds
```

```
## elapsed time 133 seconds
```

```
## elapsed time 144 seconds
```

```
## elapsed time 154 seconds
```

```
## elapsed time 165 seconds
```

```
## elapsed time 176 seconds
```

```
## elapsed time 187 seconds
```

```
## elapsed time 198 seconds
```

```
## elapsed time 209 seconds
```

```
## elapsed time 220 seconds
```

```
# Multiple Alignments
USHitsDF <- data.frame(ID = UnSeqBLAST$Hit_accession,
                       Seq = UnSeqBLAST$Hsp_hseq,
                       stringsAsFactors = FALSE)
UnSeqBLAST$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"
```

All these 20 base pairs have similar length.

```
USHitSeqs <- read.GenBank(UnSeqBLAST$Hit_accession)
# Take a look at the species
attr(USHitSeqs, "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"
```

From this output we could confirm that the separated unknown sequence are related to coronavirus 2, which could cause severe acute respiratory syndrome.

```
# Convert DNASbin to DNAStringSet
Cov2DNAstring <- USHitsDF$Seq %>%
  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
```

```
# Give each sequence a unique names
names(Cov2DNAstring) <- paste(1:nrow(USHitsDF), USHitsDF$ID, sep = "_")

# Use MUSCLE to align the sequences
Cov2Align <- muscle::muscle(stringset = Cov2DNAstring, quiet = T)
```

```
## Warning in file.remove(tempIn, tempOut): 无法删除文件'C:
## \Users\huawei\AppData\Local\Temp\RtmpYDFWB8\file4164166b6c0a.afa', 原因
## 是'Permission denied'
```

```
# Check the result
Cov2Align
```

```
## 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 our DNAMultipleAlignment object to a DNABin object
Cov2AlignBin <- as.DNAbin(Cov2Align)
```

```
SeqLen <- as.numeric(lapply(Cov2DNAstring, length))
# Show the distribution of sequence length
qplot(SeqLen) + theme_bw()
```

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

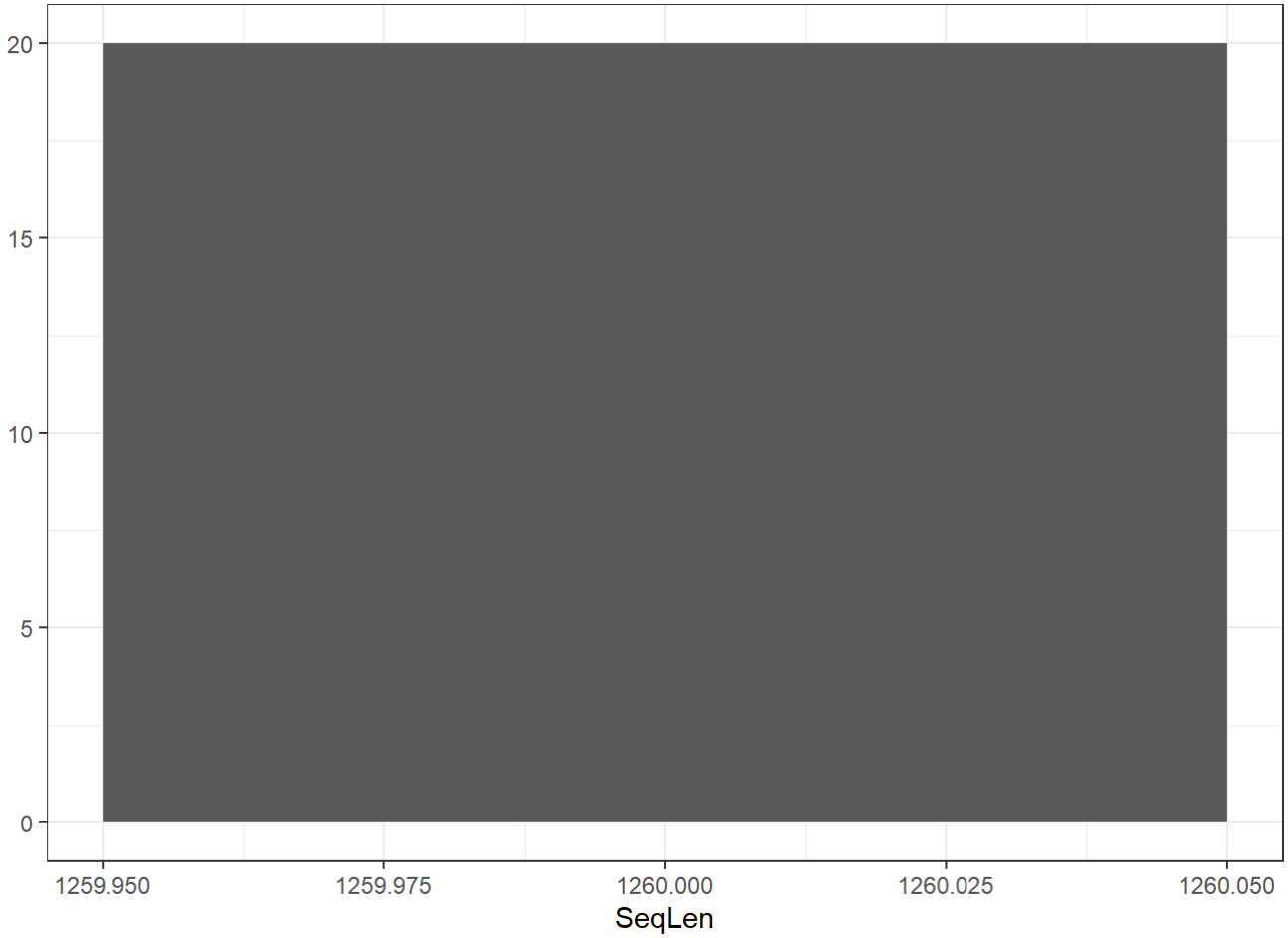

Fig.1 Distribution of the matches of sequence length. #### According to the alignment and distribution graph, there is no obvious gap within the sequence. Thus, we don't need to remove any part of them.

Visualizing the distance matrix

```
CovDM <- dist.dna(Cov2AlignBin, model = "K80")
CovDMmat <- as.matrix(CovDM)

# Plot the distance matrix
PDat <- melt(CovDMmat)
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 = 45, hjust = 1))
```

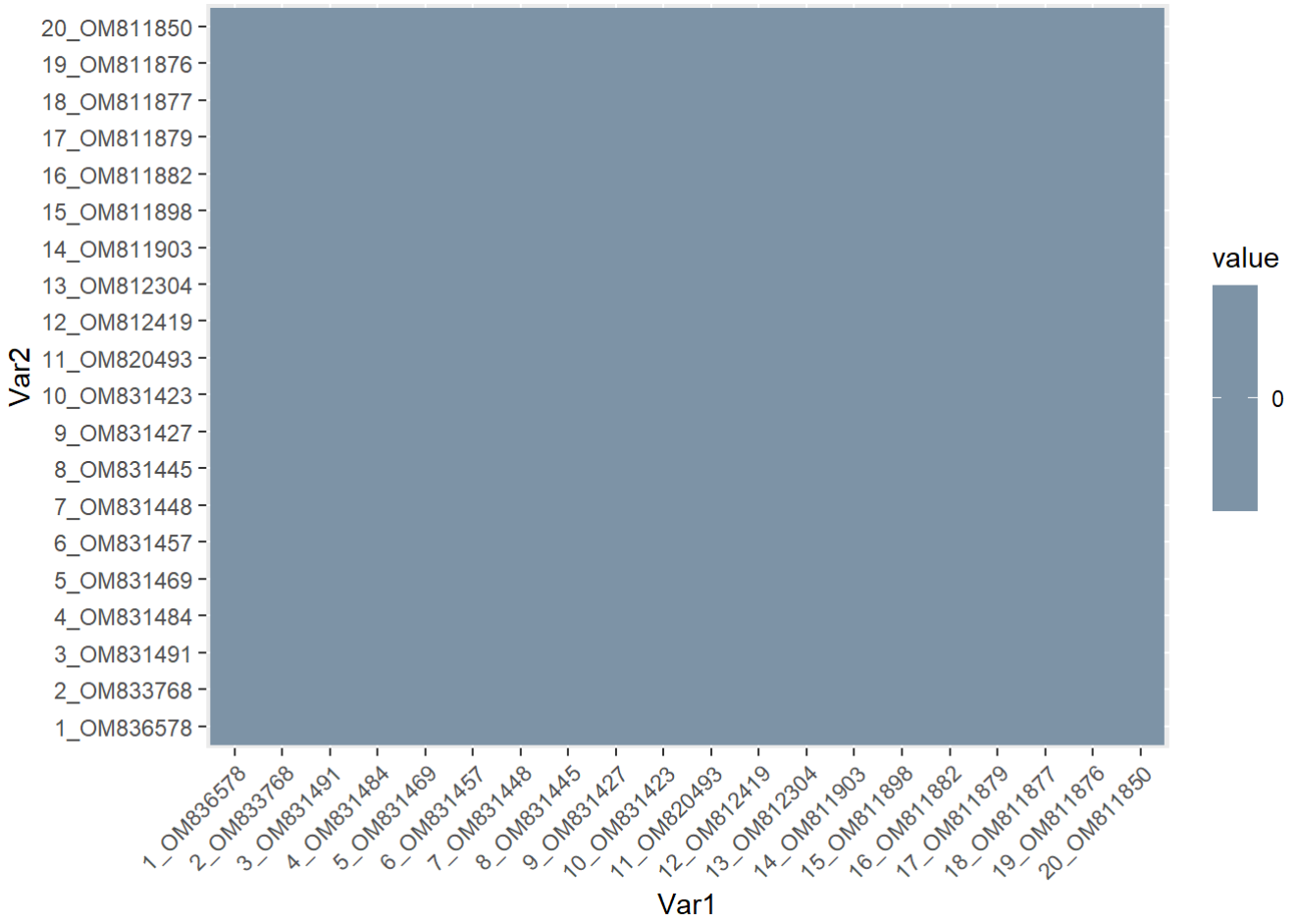

Fig.2 Figure of the distance matrix. #### The figure shows that all the sequences has a really small distance to each other. Since they are all form same species, coronavirus 2, it is possible that they have 0 distance to each other.

```
Cov2Tree <- nj(CovDM)

# Visualizing the phylogenetic tree
ggtree(Cov2Tree)
```

Fig.3 Phylogenetic tree of the 20 selected sequences. #### There is no branch in this phylogenetic tree, since the distances among sequences are similar and small (as shown in Fig.2). Thus, all the 20 sequences are close related to each other and probabily came from same species. #### The branch lengths in the above graph are based on the pairwise distance matrix, thus we could remove the branch length info to focus on the relationships among our 20 sequences.

```
# Adjust the tree
ggtree(Cov2Tree, branch.length = 'none', layout = "rectangular") + geom_tiplab()
```

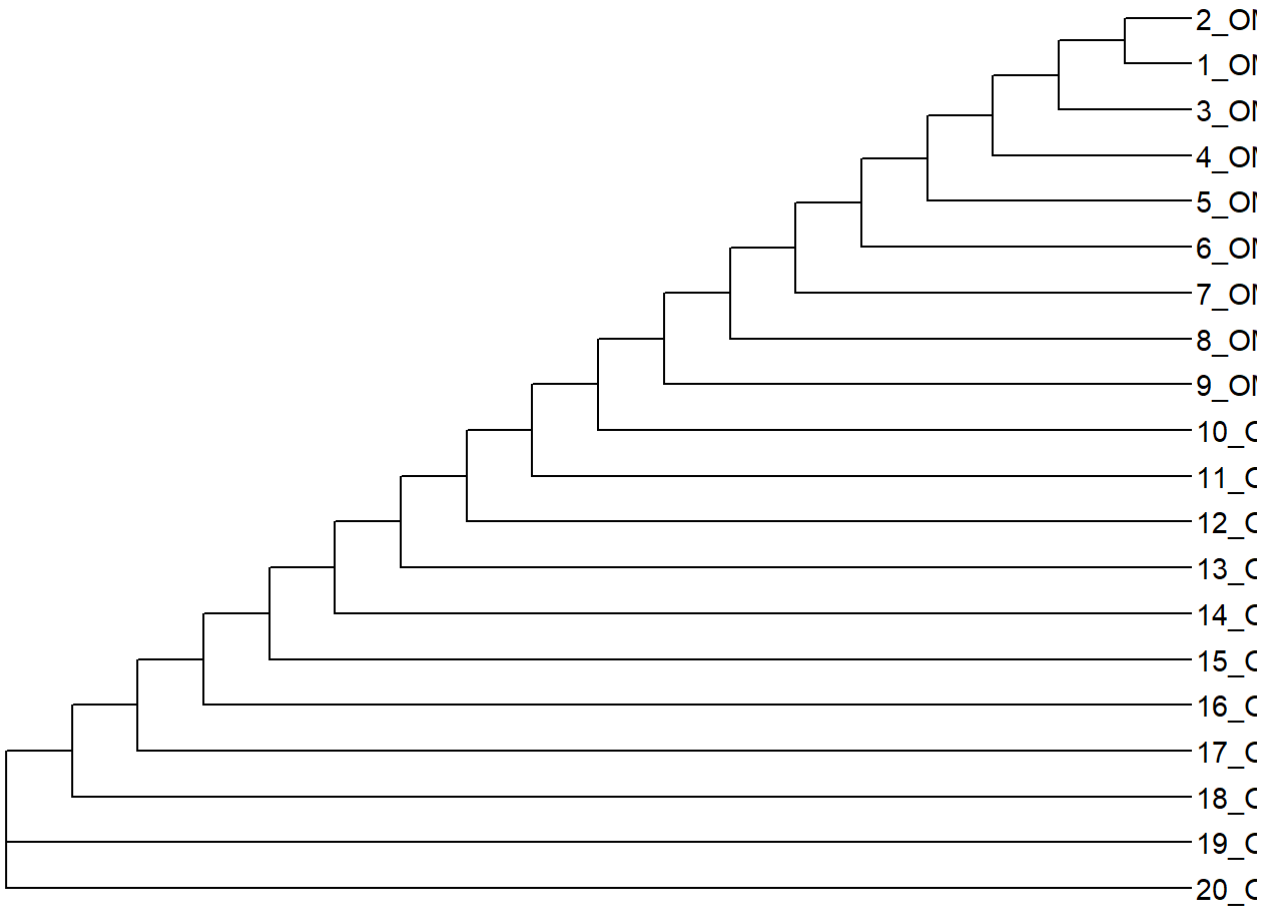

Fig.4 Phylogenetic tree of the 20 selected sequences without consider the branch length.

Save the tree

```
write.tree(Cov2Tree, "Coronavirus_2_tree.tre")
```

Report

For this unknown sequence, we performed a series of tests including alignment and production of phylogenetic tree (Fig.1-4). Based on the results of our analysis of the patient's blood samples, we found that this is a sequence from coronavirus 2. This is a situation to be concerned about because it may cause severe acute respiratory syndrome, affecting the patient's treatment or recovery.