Assignment 6 - Caroline Tang 20115082

Project Info

Github repository
Github username: carolinetang77
Date: 2022-03-01

DNA Alignment

Load required packages

```
library(annotate)
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: stats4
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## 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
```

```
## 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")'.
## Loading required package: IRanges

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

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

##
## Attaching package: 'IRanges'

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

## Loading required package: XML
```

```
library(ape)
library(muscle)
```

```
## Loading required package: Biostrings

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

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

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

##
## Attaching package: 'muscle'

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

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:Biostrings':
##
##     collapse, intersect, setdiff, setequal, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following object is masked from 'package:XVector':
##
##     slice

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

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following object is masked from 'package:Biobase':
##
##     combine

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

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

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

```
library(Biostrings)
library(ggplot2)
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 Bioi
## 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 v
```

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

Save sequence as an object

```
newSeq <- "ATGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGCATTACGTTTGGTGGACCCTCAGATTCAACTGGCAGTAACCAGAATGGAGA
print(newSeq)
```

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

BLAST search for similar sequences

```
seqBlast <- blastSequences(newSeq, as = "data.frame", hitListSize = 40, timeout = 600)
```

```
## estimated response time 28 seconds
```

```
## elapsed time 28 seconds
```

```
## elapsed time 39 seconds
```

```
## elapsed time 50 seconds
```

```
## elapsed time 61 seconds
## elapsed time 71 seconds

## elapsed time 82 seconds

## elapsed time 93 seconds

## elapsed time 104 seconds

## elapsed time 115 seconds

## elapsed time 126 seconds

## elapsed time 137 seconds

## elapsed time 147 seconds

## elapsed time 158 seconds

## elapsed time 169 seconds

## elapsed time 179 seconds

## elapsed time 190 seconds

## elapsed time 201 seconds

## elapsed time 212 seconds

## elapsed time 222 seconds
```

Alignments

Create dataframe of just hit accession IDs and the matching sequences

```
blastDF <- data.frame(ID = seqBlast$Hit_accession,
                      Seq = seqBlast$Hsp_hseq,
                      stringsAsFactors = FALSE)
#append the original sequence
blastDF <- rbind(blastDF, data.frame(ID = "original", Seq = newSeq))
```

Convert the sequences to a DNAStringSet object

```
blastString <- blastDF$Seq %>%
  as.character() %>%
  lapply(., paste0, collapse = "") %>%
  unlist() %>%
  DNAStringSet()
names(blastString) <- paste0(1:nrow(blastDF), "_", blastDF$ID)
```

Align the sequences

```
blastAlign <- muscle::muscle(stringset = blastString, quiet = T)
```

```
## Warning in file.remove(tempIn, tempOut): cannot remove file 'C:
## \Users\carol\AppData\Local\Temp\Rtmp4iukXg\filed383f5d5697.afa', reason
## 'Permission denied'
```

Check for gaps in the sequences

```
seqLen <- as.numeric(lapply(blastString, length))
qplot(seqLen) + theme_classic()
```

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

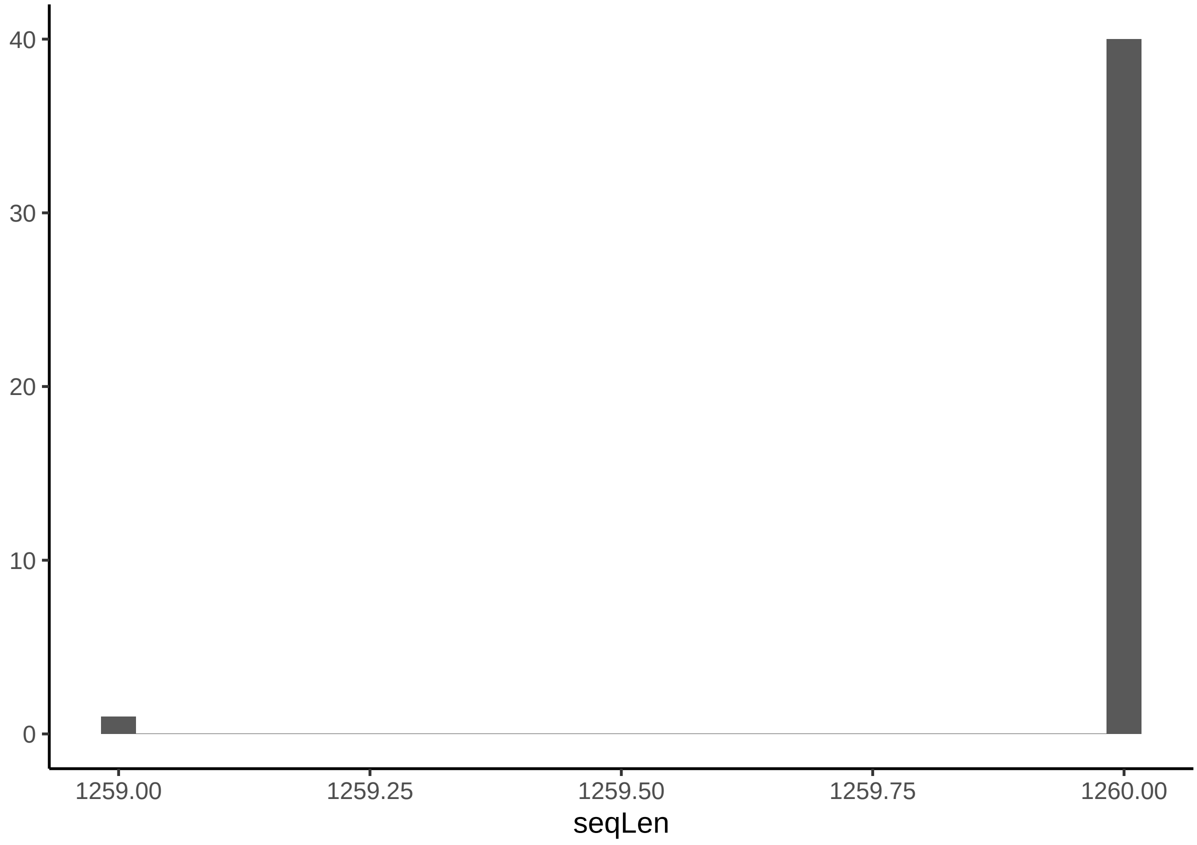

Figure 1. Histogram showing the lengths of sequences that match the original sequence

The lengths of the sequences are approximately the same, so the alignment will not need to be re-adjusted.

Distance matrix

Convert blastString to a DNAbin and create a pairwise distance matrix

```
blastBin <- as.DNAbin(blastAlign)
blastDM <- dist.dna(blastBin, model = "K80")

#Convert to a matrix format
blastDM <- as.matrix(blastDM)

#Reshape the matrix
blastReshape <- reshape2::melt(blastDM)

#Plot the matrix
ggplot(data = blastReshape, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  labs(x = "Sequence", y = "Sequence", fill = "Distance") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5))
```

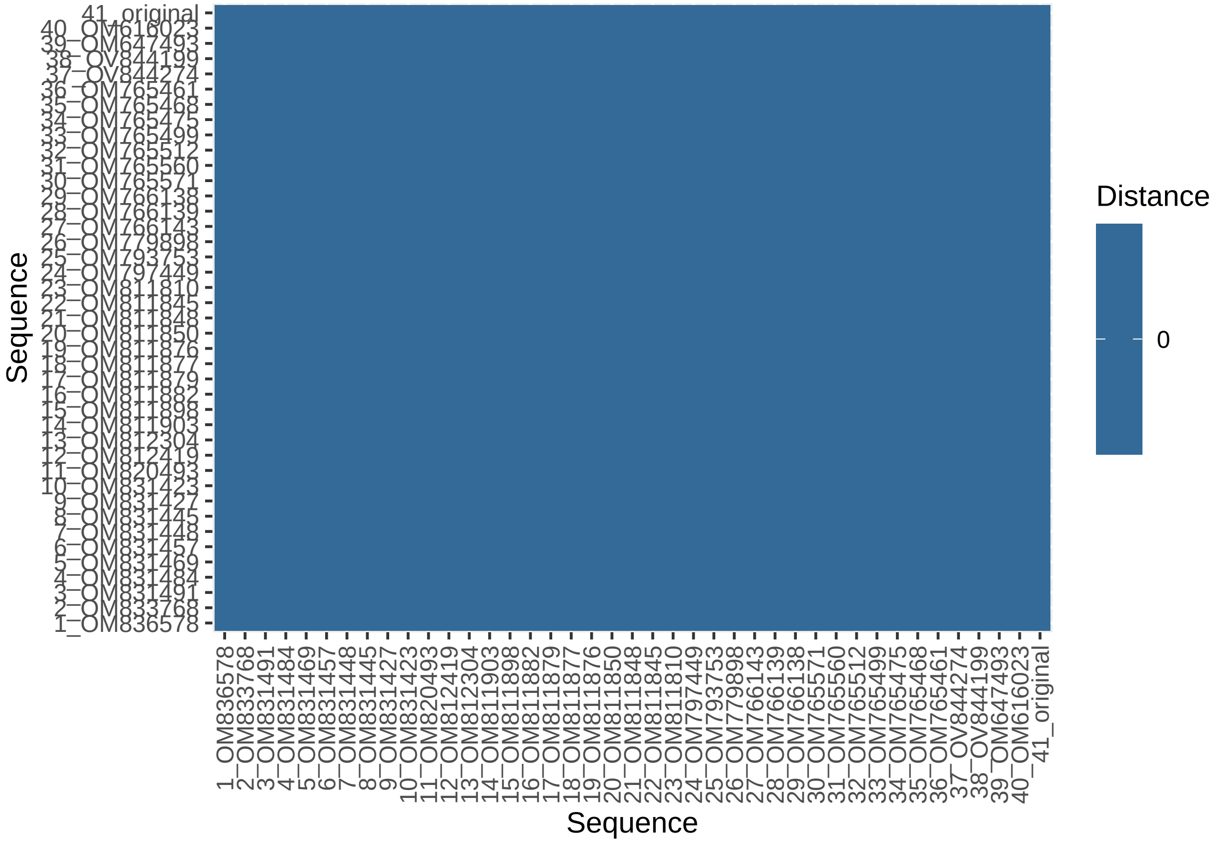

Figure 2. A visualization of the pairwise distance matrix of the original sequence and the 40 results from the BLAST search

Of the 40 hits that the BLAST search returned, all of them are identical to the original (sequence 41). Thus, finding the species identity of these sequences will likely provide an identity to the original DNA sequence.

```
blastHitSeqs <- read.GenBank(seqBlast$Hit_accession)
attr(blastHitSeqs, "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"
## [21] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [22] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [23] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [24] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [25] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [26] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [27] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [28] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [29] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [30] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [31] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [32] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [33] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [34] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [35] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [36] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [37] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [38] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [39] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [40] "Severe_acute_respiratory_syndrome_coronavirus_2"
```

All the sequences seem to be from SARS-Cov-2, i.e. Covid-19. Thus the unknown DNA sequence found in the patient is likely to be from the Covid-19 virus, which would be cause for concern. We can also create a phylogeny to determine if there were any new mutations, though the unknown sequence seemed to entirely match the others.

Phylogeny

Using the neighbour-joining method, we can create the phylogenetic tree.

```
#calculate distances with the neighbour-joining method
seqTree <- nj(blastDM)

#plot the tree
```

```r
ggtree(seqTree, branch.length = "none", layout = "radial") +
  geom_tiplab()
```

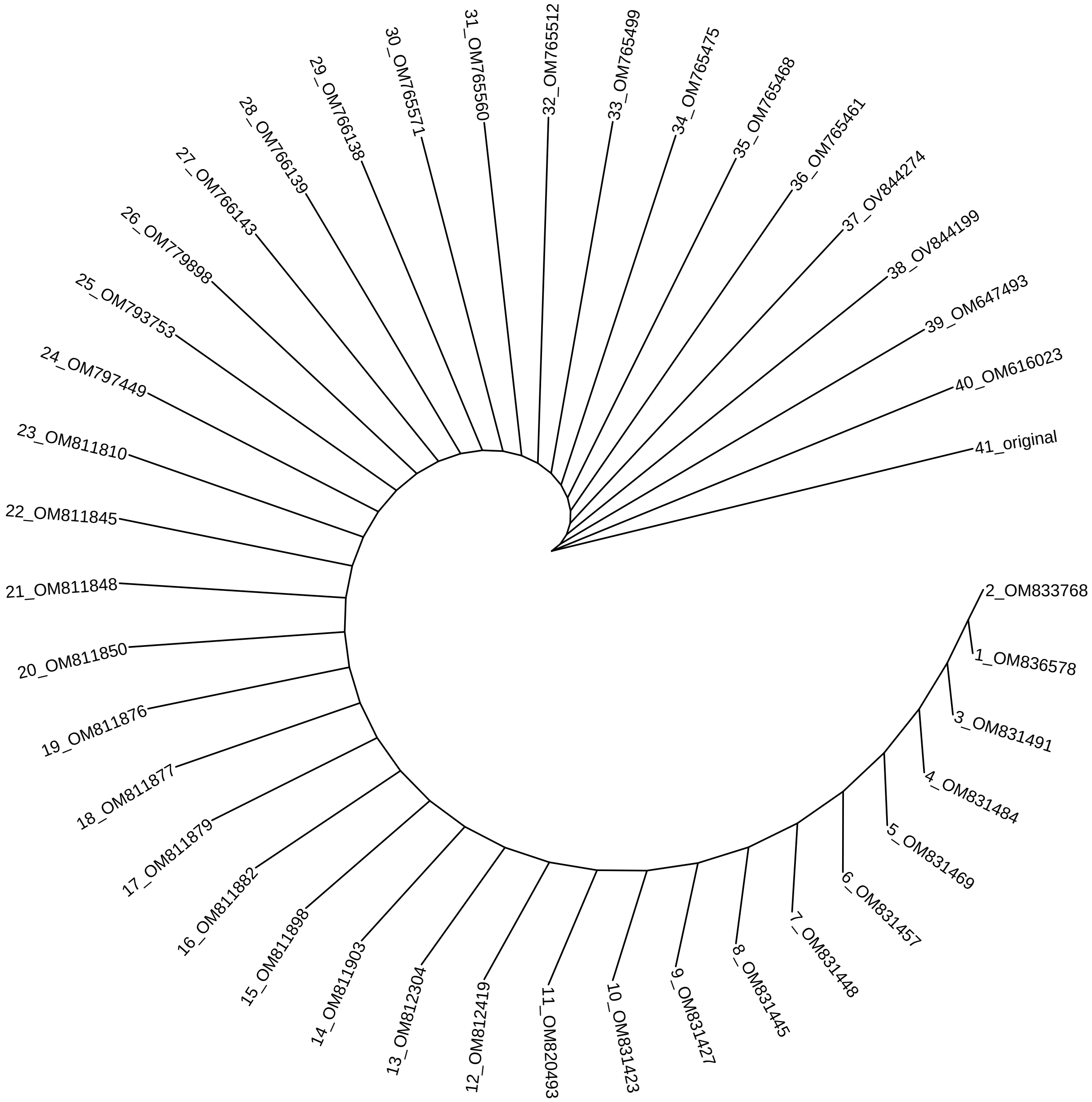

Figure 3. A cladogram depicting the evolutionary relationships between the different sequences

As mentioned above, the unknown sequence was identical to the known sequences, and this cladogram shows that all the sequences seem to be equally related. Thus the virus from the patient likely does not contain any novel mutations.