Alignments

Logan Wisteard

2022-03-03

#Project Info ##Github repository:https://github.com/LoganWisteard/DNA-Alignments ####Date: 2022-03-02

```
devtools::install_github("YuLab-SMU/ggtree")
```

```
## Skipping install of 'ggtree' from a github remote, the SHA1 (60be2d12) has not changed since last ins
##   Use `force = TRUE` to force installation
```

DNA Alignment

Load required packages

```
library(annotate)
```

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

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

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

```
## Warning: package 'BiocGenerics' was built under R version 4.0.5
```

```
## Loading required package: parallel
```

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

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## 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 object is masked from 'package:base':
##
##     expand.grid

## Loading required package: XML

## Warning: package 'XML' was built under R version 4.0.5
```

```
library(ape)
```

```
## Warning: package 'ape' was built under R version 4.0.5
```

```
library(muscle)
```

```
## Loading required package: Biostrings

## Loading required package: XVector

##
## 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)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5

##
## Attaching package: 'dplyr'

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

## 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 v3.3.1  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 vi

##
## 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 35 seconds

## elapsed time 35 seconds

## elapsed time 46 seconds
```

```
## elapsed time 57 seconds

## elapsed time 68 seconds

## elapsed time 79 seconds

## elapsed time 90 seconds

## elapsed time 101 seconds

## elapsed time 112 seconds

## elapsed time 123 seconds

## elapsed time 134 seconds

## elapsed time 145 seconds

## elapsed time 156 seconds

## elapsed time 167 seconds

## elapsed time 178 seconds

## elapsed time 189 seconds

## elapsed time 200 seconds

## elapsed time 211 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)
```

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`.
```

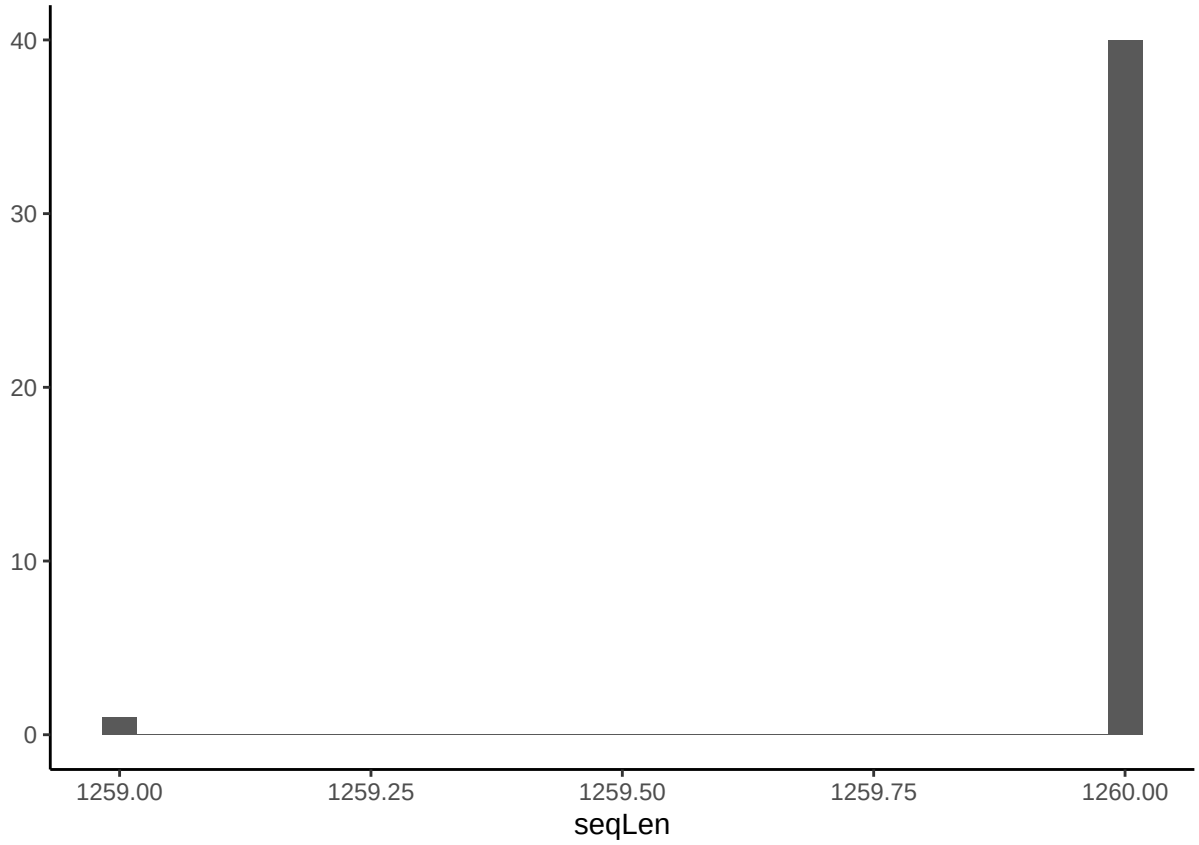


Figure 1. Histogram shows that the lengths of sequences that match the original sequence

The lengths of the sequences are very similar, so there will be no alignment or adjustments needed.

Distance matrix

Use 'dist.dna()' function to estimate 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))
```

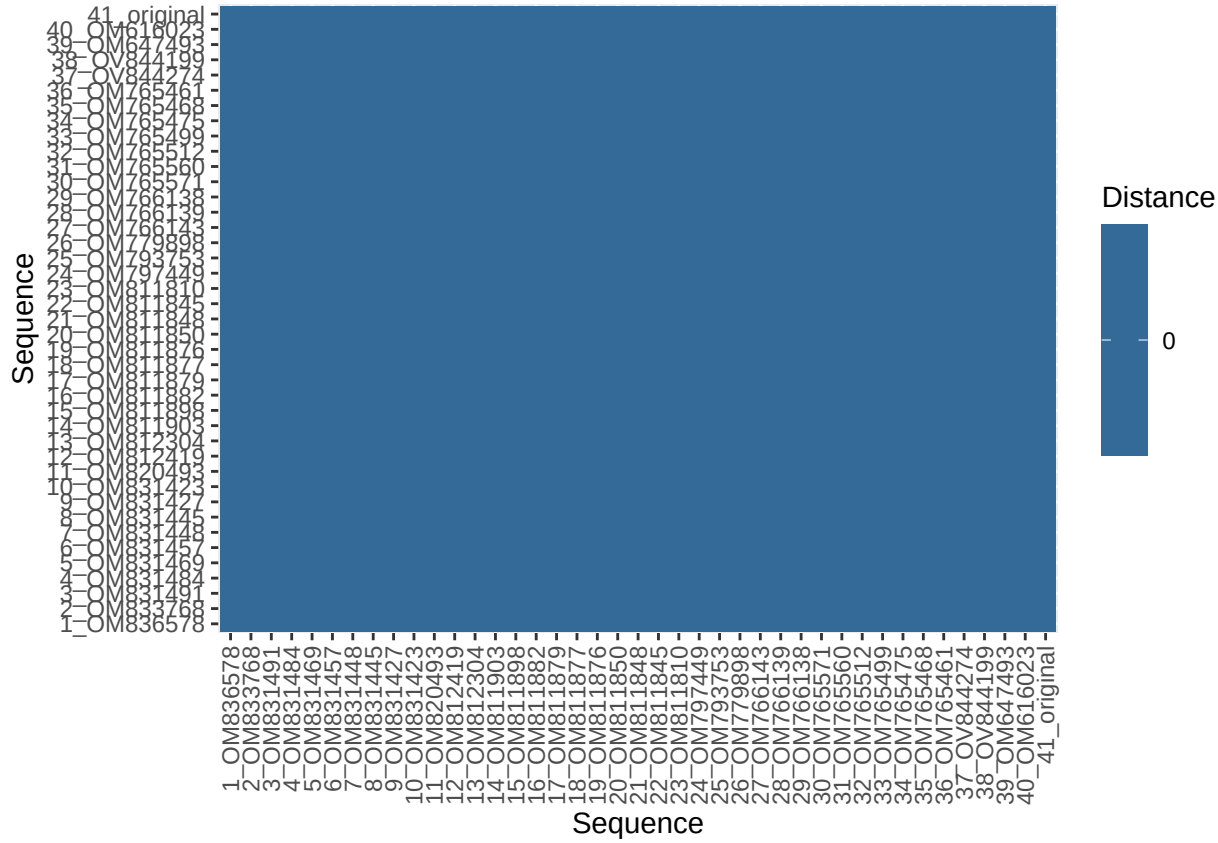


Figure 2. Pairwise distance matrix of the original sequence displays the 40 results from the BLAST search

All of the 40 hits are identical to the original (sequence 41). 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"
```

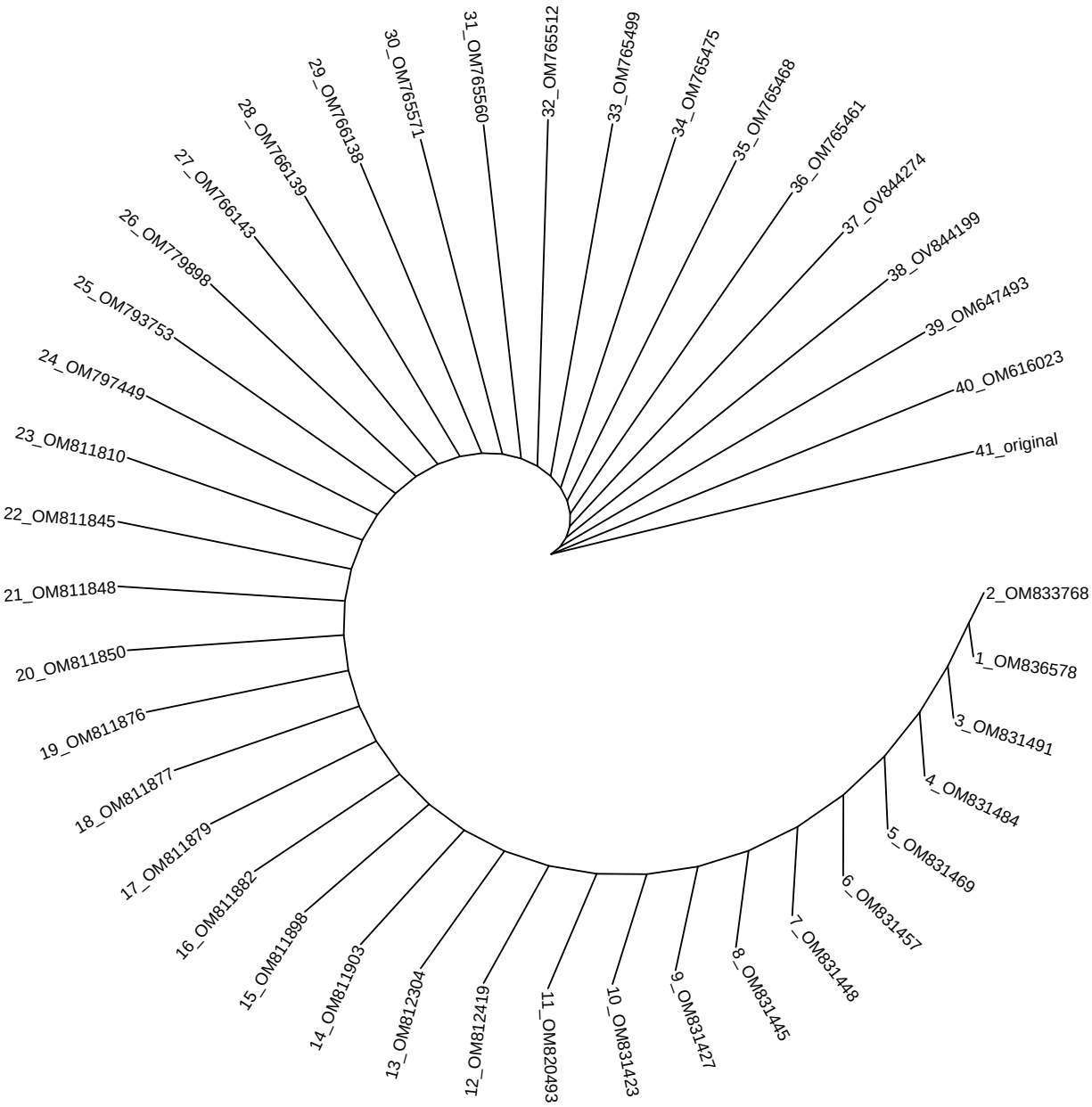
```
## [14] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [15] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [16] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [17] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [18] "Severe_acute_respiratory_syndrome_coronavirus_2"
## [19] "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 sequences are found in SARS-Cov-2,novel coronavirus. This unknown DNA sequence that was present in the patient is likely to be from the Covid-19 virus, it now must be tested to reveal if any new mutations of concern are present. A phylogeny can be created to investigate any new mutations, this unknown sequence seemed to entirely match the others.

Phylogeny

Using the neighbour-joining method,a Phylogenetic tree will be created.

```
seqTree <- nj(blastDM)
ggtree(seqTree, branch.length = "none", layout = "radial") +
  geom_tiplab()
```


2_OM833768
1_OM836578
3_OM831491
4_OM831484
5_OM831469
6_OM831457
7_OM831448
8_OM831445
9_OM831427
10_OM831423
11_OM820493
12_OM812419
13_OM812304
14_OM811903
15_OM811898
16_OM811882
17_OM811879
18_OM811877
19_OM811876
20_OM811850
21_OM811848
22_OM811845
23_OM811810
24_OM797449
25_OM793753
26_OM779898
27_OM766143
28_OM766139
29_OM766138
30_OM765571
31_OM765560
32_OM765512
33_OM765499
34_OM765475
35_OM765468
36_OM765461
37_OV844274
38_OV844199
39_OM647493
40_OM616023
41_original