In [1]:
```
library("igraph")
library("ggtree")
library("phangorn")
library("treeio")
library("Biostrings")
library("msa")
library("ape")
library("insect")
```

```
Warning message:
"package 'igraph' was built under R version 3.6.3"
Attaching package: 'igraph'

The following objects are masked from 'package:stats':

    decompose, spectrum

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

    union

Warning message:
"package 'ggtree' was built under R version 3.6.3"Registered S3 methods overwr
itten by 'ggplot2':
  method         from
  [.quosures     rlang
  c.quosures     rlang
  print.quosures rlang
Registered S3 method overwritten by 'treeio':
  method     from
  root.phylo ape
ggtree v2.0.2  For help: https://yulab-smu.github.io/treedata-book/

If you use ggtree in published research, please cite the most appropriate pape
r(s):

- Guangchuang Yu, Tommy Tsan-Yuk Lam, Huachen Zhu, Yi Guan. Two methods for ma
pping and visualizing associated data on phylogeny using ggtree. Molecular Bio
logy and Evolution 2018, 35(12):3041-3043. doi: 10.1093/molbev/msy194
- Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtre
e: an R package for visualization and annotation of phylogenetic trees with th
eir covariates and other associated data. Methods in Ecology and Evolution 201
7, 8(1):28-36, doi:10.1111/2041-210X.12628


Warning message:
"package 'phangorn' was built under R version 3.6.3"Loading required package:
ape
Warning message:
"package 'ape' was built under R version 3.6.3"
Attaching package: 'ape'

The following object is masked from 'package:ggtree':

    rotate

The following objects are masked from 'package:igraph':

    edges, mst, ring

Attaching package: 'phangorn'

The following object is masked from 'package:igraph':

    diversity

treeio v1.10.0  For help: https://yulab-smu.github.io/treedata-book/

If you use treeio in published research, please cite:

LG Wang, TTY Lam, S Xu, Z Dai, L Zhou, T Feng, P Guo, CW Dunn, BR Jones, T Bra
dley, H Zhu, Y Guan, Y Jiang, G Yu. treeio: an R package for phylogenetic tree
input and output with richly annotated and associated data. Molecular Biology
and Evolution 2019, accepted. doi: 10.1093/molbev/msz240


Attaching package: 'treeio'

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

    drop.tip

The following object is masked from 'package:igraph':

    parent

Loading required package: BiocGenerics
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:igraph':

    normalize, path, 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, which.max, which.min
Loading required package: S4Vectors
Warning message:
"package 'S4Vectors' was built under R version 3.6.2"Loading required package:
stats4

Attaching package: 'S4Vectors'

The following object is masked from 'package:ggtree':

    expand

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

    expand.grid

Loading required package: IRanges
Warning message:
"package 'IRanges' was built under R version 3.6.2"
Attaching package: 'IRanges'

The following object is masked from 'package:ggtree':

    collapse

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

    windows

Loading required package: XVector

Attaching package: 'Biostrings'

The following object is masked from 'package:treeio':

    mask

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

    complement

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

    strsplit

Warning message:
"package 'insect' was built under R version 3.6.3"Registered S3 method overwri
tten by 'openssl':
  method      from
```

Построем филогенетическое дерево по геномам различных штаммов коронавируса. Данные загружены с сайта NCBI (https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/#nucleotide-sequences) и собраны в один файл `common.fasta` в формате `fasta` . Длины ДНК-последовательностей отличаются, поэтому предварительно требуется выровнять их. Это можно сделать с помощью следующего кода:

In [ ]:
```
fasta_data <- read.fasta("common.fasta")
l = c(dna2char(fasta_data[1]))
for (i in 2:16) {
    l = rbind(l, c(dna2char(fasta_data[i])))
}
string.set <- DNAStringSet(l)
string.set <- msa(DNAStringSet(l))
fasta_data <- as.DNAbin(string.set)
```

Демонстрация работы на маленьких данных:

In [2]:
```
Q1 <- as.DNAbin(c("T","C","C","G","A","A","T","A","A","G","T","A","A","A"))
Q2 <- as.DNAbin(c("C","C","G","A","A","T","C","A","G","T","A"))
Q3 <- as.DNAbin(c("T","C","T","A","A","A","T","A","A","G","C","A","C"))
print(msa(DNAStringSet(c(dna2char(Q1),dna2char(Q2),dna2char(Q3)))))
```

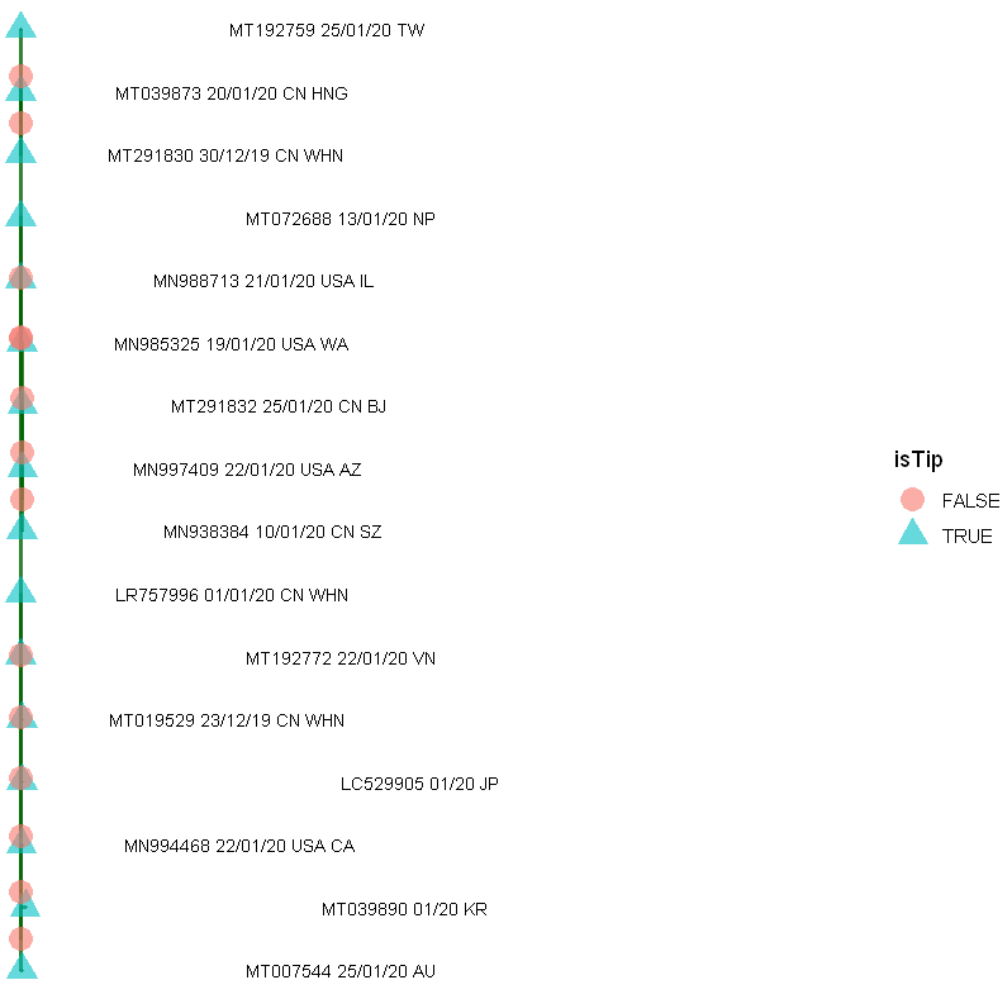
```
use default substitution matrix
CLUSTAL 2.1

Call:
   msa(DNAStringSet(c(dna2char(Q1), dna2char(Q2), dna2char(Q3))))

MsaDNAMultipleAlignment with 3 rows and 14 columns
    aln
[1] TCCGAATAAGTAAA
[2] -CCGAATCAGTA--
[3] TCTAAATAAGCAC-
Con TCCGAATAAGTA?-
```

Однако для выравнивания нескольких десятков последовательностей длины порядка 30000 нуклеотидов у меня не достаточно вычислительной мощности (это займет слишком много времени), поэтому для получения того же результата воспользуемся онлайн-ресурсом Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). Полученный в результате файл с выровненными ДНК-последовательностями назовем `common_msa.fasta` .

Теперь построем филогенетического дерева. Для удобства в подписи вершин вынесем только **id** штамма, страну и дату обнаружения.

In [3]:
```
rnames <- c("MT007544 25/01/20 AU",
            "MT039890 01/20 KR",
            "MN988713 21/01/20 USA IL",
            "MT291832 25/01/20 CN BJ",
            "MN985325 19/01/20 USA WA",
            "MN938384 10/01/20 CN SZ",
            "MN997409 22/01/20 USA AZ",
            "MT291830 30/12/19 CN WHN",
            "LC529905 01/20 JP",
            "MT019529 23/12/19 CN WHN",
            "MN994468 22/01/20 USA CA",
            "MT072688 13/01/20 NP",
            "MT039873 20/01/20 CN HNG",
            "LR757996 01/01/20 CN WHN",
            "MT192772 22/01/20 VN",
            "MT192759 25/01/20 TW")
```

In [7]:
```
fasta_data <- as.matrix(read.fasta("common_msa.fasta"))
rownames(fasta_data) <- rnames
phy.data <- as.phyDat(as.matrix(fasta_data))
tree <- nj(dist.ml(phy.data))
ggtree(tree, lwd = 1, color = "darkgreen", alpha = 0.8, right = TRUE) +
  geom_tiplab(size = 3, angle = 0, offset = 0.05, hjust = 3) +
  geom_point(aes(shape = isTip, color = isTip), size = 5, alpha = 0.6)
```



Реализуем функцию, позволяющую сравнивать поэлементно нуклеотидные последовательности одной длины и возвращающую сходство в виде числа от 0 до 1.

In [49]:
```
element_wise.compare <- function(seq1,seq2){
    mean(strsplit(as.character(DNAseq1),"")[[1]] ==
       strsplit(as.character(DNAseq2),"")[[1]])
}
```

In [50]:
```
element_wise.compare(DNAseq1, DNAseq2)
```

```
0.99933117078554
```