

Evidencia 2

Contents

Preparación	2
Librerías	2
Utilidades	3
Metodología	4
Video	4
Caso seleccionado	4
Longitud de las secuencias	6
Comparación de las secuencias	7
Árbol filogenético	9
Árbol filogenético por continentes	15
Resultado final	16
Conclusión	16
Referencias	17

Preparación

Librerías

```
suppressMessages(library(seqinr))  
suppressMessages(library(ape))  
suppressMessages(library(ggplot2))  
suppressMessages(library(ggtree))
```

Utilidades

```
seq.length <- function(dna.seq) {
  getLength(dna.seq)
}

seq.composition <- function(dna.seq) {
  total <- seq.length(dna.seq)
  bases <- count(dna.seq, 1)
  bases["n"] <- total - sum(bases)

  sapply(bases, function(x) {
    round(x / total * 100, 2)
  })
}

base.colors <- c(
  "-" = "#FD8D0E",
  "t" = "#106BFF",
  "a" = "#FC2B2D",
  "g" = "#30D33B",
  "c" = "#FECF0F",
  "n" = "#3D3D3D",

  "T" = "#106BFF",
  "A" = "#FC2B2D",
  "G" = "#30D33B",
  "C" = "#FECF0F",
  "N" = "#3D3D3D"
)

title.theme <- theme(
  plot.title = element_text(size = 30, face = "bold", hjust = 0.5),
  plot.subtitle = element_text(size = 20, hjust = 0.5),
  plot.caption = element_text(size = 10)
)

legend.theme <- theme(
  legend.title = element_text(size = 25),
  legend.key.size = unit(1, "cm"),
  legend.text = element_text(size = 20)
)

caption <- labs(caption = "Data source: NCBI")
```

Metodología

Video

<http://www.youtube.com/watch?v=XurCL2ZvR5c>

Caso seleccionado

La investigación que se vamos a realizar es sobre las variantes del virus en los países con mayor cantidad de casos. Al finalizar el estudio, se busca determinar qué tan diferentes son las variantes entre estos países.

También se demostrará si hay diferencias significativas en el virus entre las poblaciones Asiáticas, Hispanas, Europeas, y Africanas, o si son similares.

Muestras de los primeros 20 países con mayor número de casos (descendiente, April 26, 2021 at 3:45 p.m. ET)

```
accessions <- c(
  "MZ021779.1" = "USA",
  "MW828655.1" = "India",
  "MW592707.1" = "Brazil",
  "HG993789.1" = "France",
  "MW305251.1" = "Russia",
  "MW320691.1" = "Turkey",
  "OD906787.1" = "United Kingdom",
  "MW854297.1" = "Italy",
  "MW976780.1" = "Spain",
  "MW633324.1" = "Germany",
  "MW633898.1" = "Argentina",
  "MT470219.1" = "Colombia",
  "HG994166.1" = "Poland",
  "MW737421.1" = "Iran",
  "MW595909.1" = "Mexico",
  ##""      = "Ukraine",
  "MT263074.1" = "Peru",
  "MZ026854.1" = "Indonesia",
  "MT517423.1" = "Czech Republic",
  "MW981442.1" = "South Africa",
  "MW309426.1" = "Canada"
)

continents <- c(
  "MZ021779.1" = "North America",
  "MW828655.1" = "Asia",
  "MW592707.1" = "South America",
  "HG993789.1" = "Europe",
  "MW305251.1" = "Europe",
  "MW320691.1" = "Asia",
  "OD906787.1" = "Europe",
  "MW854297.1" = "Europe",
  "MW976780.1" = "Europe",
  "MW633324.1" = "Europe",
  "MW633898.1" = "South America",
  "MT470219.1" = "South America",
  "HG994166.1" = "Europe",
  "MW737421.1" = "Asia",
  "MW595909.1" = "North America",
  "MT263074.1" = "South America",
  "MZ026854.1" = "Asia",
  "MT517423.1" = "Europe",
  "MW981442.1" = "Africa",
  "MW309426.1" = "North America"
)

if (!file.exists("./data/MERGED.fasta")) { ## only search once
  sequences <- read.GenBank(names(accessions))
  write.dna(sequences, file = "./data/MERGED.fasta", colsep = "", format = "fasta")
}
```

Longitud de las secuencias

```
all.seq <- read.fasta("./data/MERGED.fasta")
all.seq.bin <- read.dna("./data/MERGED.fasta", format = "fasta")
all.seq.bin
```

```
## 20 DNA sequences in binary format stored in a list.
##
## Mean sequence length: 29833.8
##   Shortest sequence: 29717
##   Longest sequence: 29903
##
## Labels:
## MZ021779.1
## MW828655.1
## MW592707.1
## HG993789.1
## MW305251.1
## MW320691.1
## ...
##
## Base composition:
##   a   c   g   t
## 0.299 0.184 0.196 0.321
## (Total: 596.68 kb)
```

```
lengths <- data.frame(Accession = character(), Country = character(), Length = integer())

for (i in 1:length(all.seq)) {
  ac <- labels(all.seq)[i]
  lengths[i, ] <- list(ac, accessions[ac], seq.length(all.seq[[ac]]))
}

lengths
```

```
##      Accession      Country Length
## 1 MZ021779.1      USA    29739
## 2 MW828655.1      India   29903
## 3 MW592707.1      Brazil  29862
## 4 HG993789.1      France  29885
## 5 MW305251.1      Russia  29841
## 6 MW320691.1      Turkey  29813
## 7 OD906787.1 United Kingdom 29903
## 8 MW854297.1      Italy   29849
## 9 MW976780.1      Spain   29763
## 10 MW633324.1     Germany  29779
## 11 MW633898.1     Argentina 29717
## 12 MT470219.1     Colombia 29903
## 13 HG994166.1     Poland   29903
## 14 MW737421.1     Iran     29816
## 15 MW595909.1     Mexico   29866
## 16 MT263074.1     Peru     29856
```

```
## 17 MZ026854.1      Indonesia 29782
## 18 MT517423.1 Czech Republic 29866
## 19 MW981442.1      South Africa 29848
## 20 MW309426.1      Canada 29782
```

Todas las secuencias tienen una longitud de alrededor de 29800 bases. Estas variaciones de longitudes se pueden deber a los cambios que hay en el genoma por mutaciones, o simplemente por la probabilidad de que ocurran errores al medir.

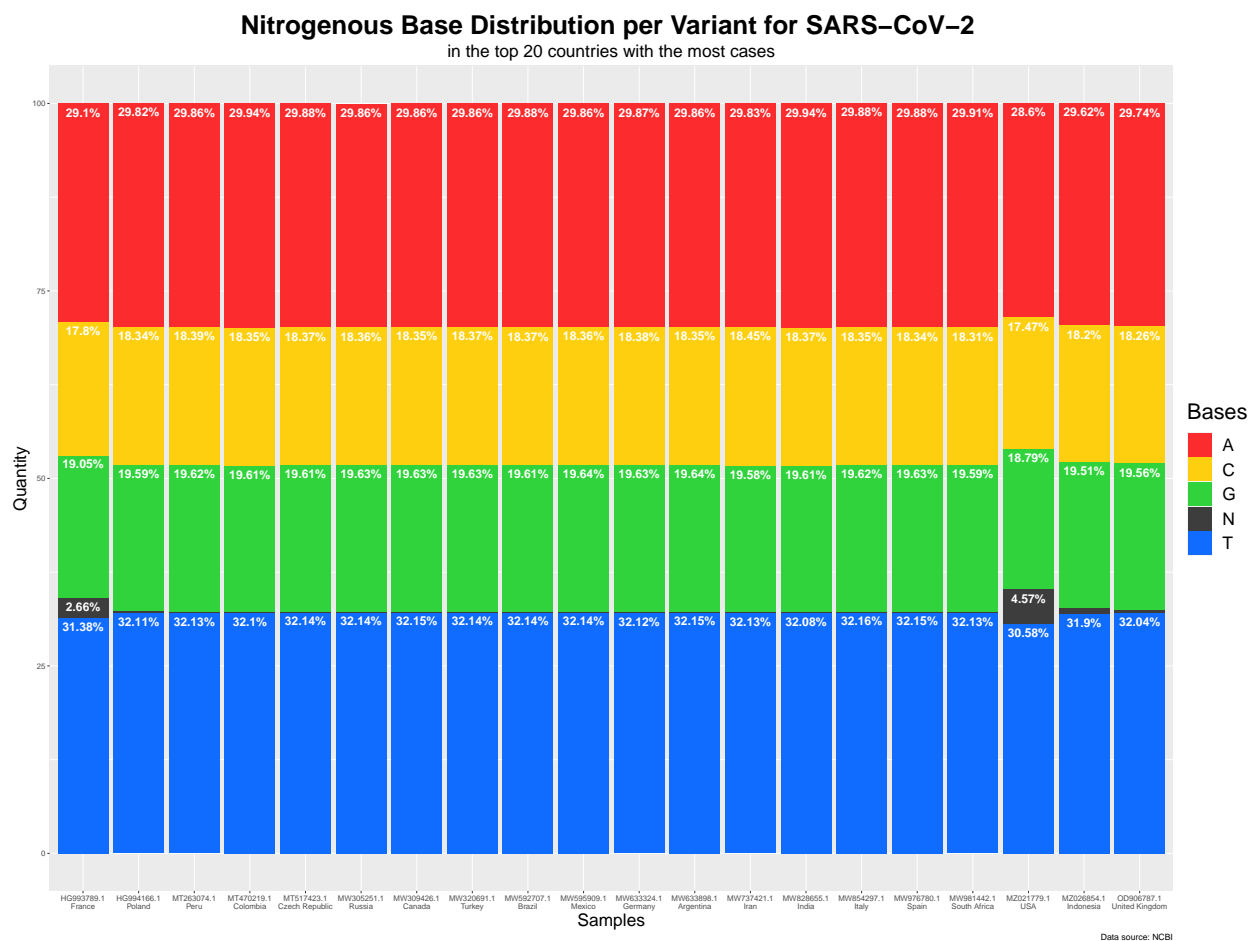
Comparación de las secuencias

```
data <- data.frame(Variant = character(), Base = character(), Value = integer())

n <- 0
for (i in labels(all.seq)) {
  var.name <- paste(i, accessions[i], sep = "\n")
  comps <- seq.composition(all.seq[[i]])
  data[n + 1,] <- list(Variant = var.name, Base = "A", Value = comps["a"])
  data[n + 2,] <- list(Variant = var.name, Base = "T", Value = comps["t"])
  data[n + 3,] <- list(Variant = var.name, Base = "G", Value = comps["g"])
  data[n + 4,] <- list(Variant = var.name, Base = "C", Value = comps["c"])
  data[n + 5,] <- list(Variant = var.name, Base = "N", Value = comps["n"])
  n <- n + 5
}

data.labs <- sapply(data$Value, function(v) { paste(v, "%", sep = "") })

ggplot(data, aes(fill = Base, y = Value, x = Variant)) +
  geom_bar(position = "stack", stat = "identity") +
  scale_fill_manual(values = base.colors) +
  geom_text(
    aes(label = data.labs),
    position = position_stack(),
    vjust = 1.5,
    colour = "white",
    fontface = "bold",
    check_overlap = TRUE,
    size = 5
  ) +
  labs(
    title = "Nitrogenous Base Distribution per Variant for SARS-CoV-2 ",
    subtitle = "in the top 20 countries with the most cases",
    x = "Samples",
    y = "Quantity",
    fill = "Bases"
  ) + caption +
  theme(axis.title = element_text(size = 20)) + title.theme + legend.theme
```



Con esta gráfica de las distribuciones se puede observar que, al menos en su composición porcentual de bases nitrogenadas, las muestras se podrían dividir en 2 categorías. Las muestras de Estados Unidos y Francia tienen una composición similar. Las muestras de los otros 18 países también se parecen entre sí. La única diferencia significativa entre estos dos grupos es la cantidad de bases que no pudieron ser identificadas, marcadas como N.


```
MT263074.1      gttctctaaacgaacttttaaaatctgtgtggctgtcactcggctgcatgcttagtgact      111
                  *****
```

...

```
clust <- read.alignment(
  "./data/clustalo-I20210427-013920-0382-41949213-p2m.clustal_num",
  format = "clustal", forceToLower = TRUE,
)
dna <- as.DNABin(clust)
dna
```

```
## 20 DNA sequences in binary format stored in a matrix.
##
## All sequences of same length: 30406
##
## Labels:
## MZ021779.1
## HG993789.1
## MZ026854.1
## MW854297.1
## MW976780.1
## MW981442.1
## ...
##
## Base composition:
##      a      c      g      t
## 0.299 0.184 0.196 0.321
## (Total: 608.12 kb)
```

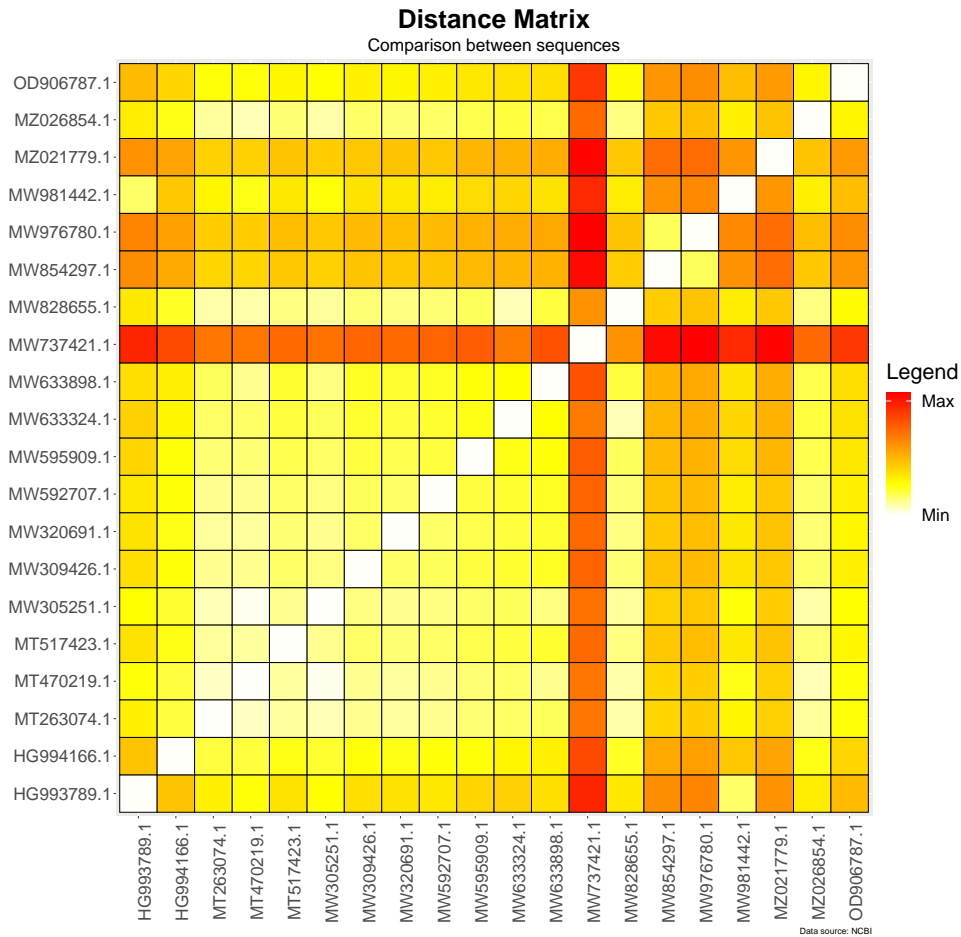
Matriz de distancia

```
D <- dist.dna(dna)
D.mat <- as.matrix(D)

heat.mat <- data.frame(X = character(), Y = character(), Val = double())

n <- 0
for (col in colnames(D.mat)) {
  for (r in rownames(D.mat)) {
    n <- n + 1
    heat.mat[n, ] <- list(col, r, D.mat[col, r])
  }
}

ggplot(heat.mat, aes(X, Y, fill = Val)) +
  geom_tile(color = "black") +
  coord_fixed() +
  scale_fill_gradientn(
    colours = heat.colors(100, rev = TRUE),
    n.breaks = 2,
    labels = c("Min", "Max"),
  ) +
  labs(
    title = "Distance Matrix",
    subtitle = "Comparison between sequences",
    fill = "Legend",
    x = NULL,
    y = NULL
  ) + caption +
  theme(
    axis.text.x = element_text(size = 20, angle = 90),
    axis.text.y = element_text(size = 20)
  ) + title.theme + legend.theme
```



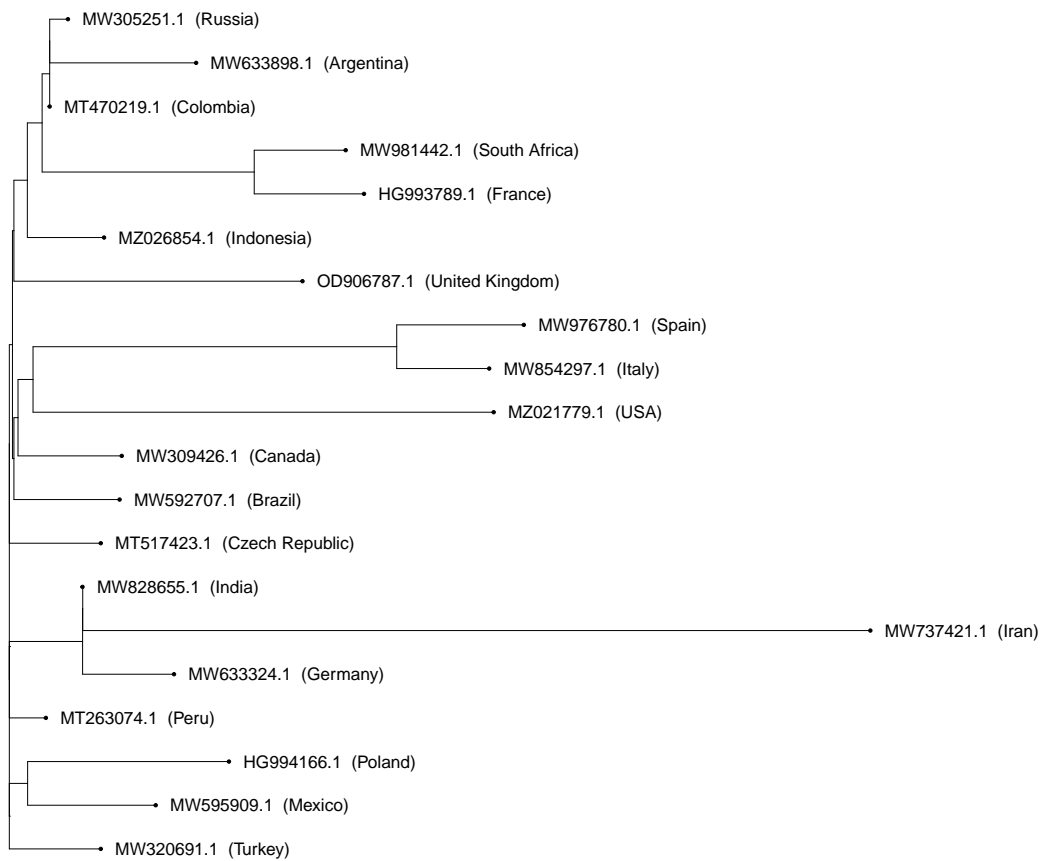
Las áreas de amarillo claro significan que hay poca diferencia, mientras que los rojos oscuros significan mayor diferencia entre secuencias. Se aprecia una diagonal, que es la comparación de las secuencias con ellas mismas. Como se espera, la distancia entre ellas es nula. Visualizando este resultado, se pudiera inferir que la secuencia MW737421.1, de Irán, estará más alejada de las demás en el árbol filogenético.

Árboles

```
tree <- nj(D)

ggtree(tree) +
  xlim(0, 0.0022) +
  geom_tippoint() +
  geom_tiplab(
    aes(label = paste(label, "  (", accessions[label], ")", sep = "")),
    offset = 0.00002,
    size = 7
  ) +
  labs(title = "SARS-CoV-2 Phylogenetic Tree") + caption + title.theme
```

SARS-CoV-2 Phylogenetic Tree

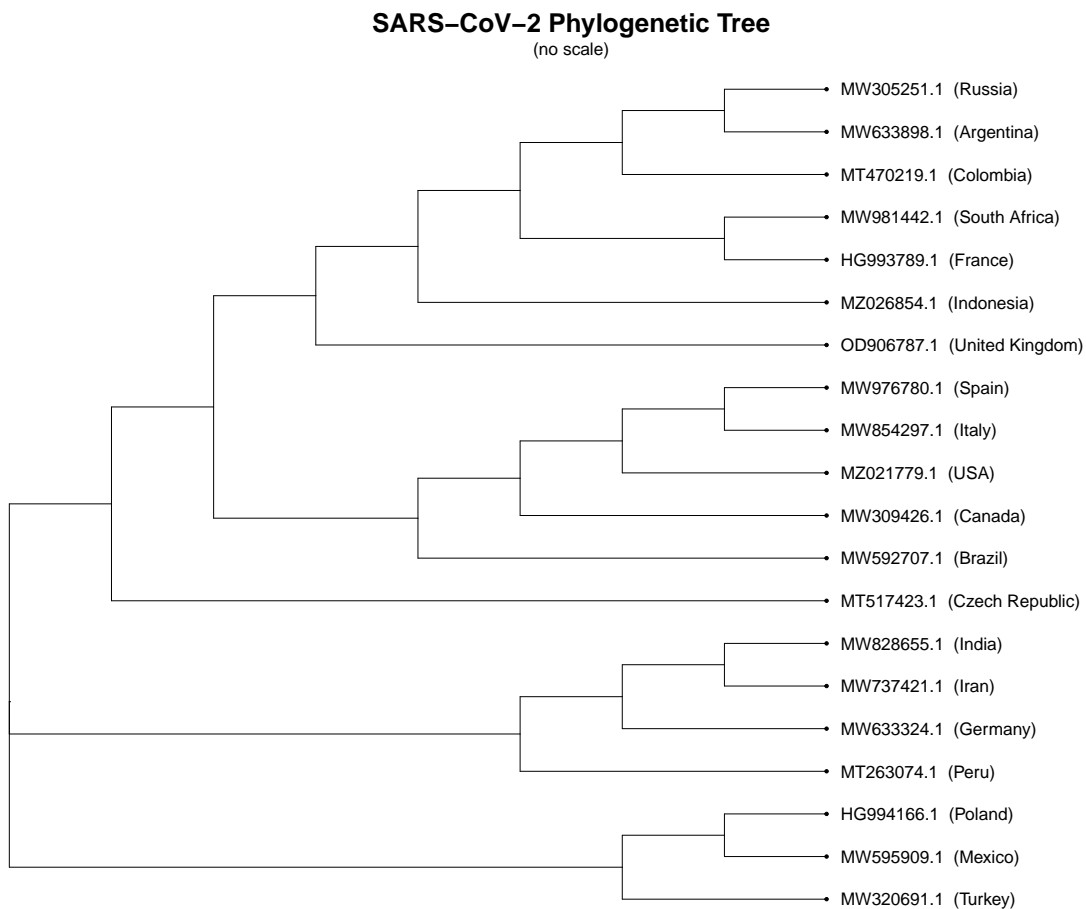


Data source: NCBI

```

ggtree(tree, branch.length = "none") +
  xlim(0, 11) +
  geom_tippoint() +
  geom_tiplab(
    aes(label = paste(label, " (", accessions[label], ")", sep = "")),
    offset = 0.1,
    size = 7
  ) +
  labs(
    title = "SARS-CoV-2 Phylogenetic Tree",
    subtitle = "(no scale)"
  ) + caption + title.theme

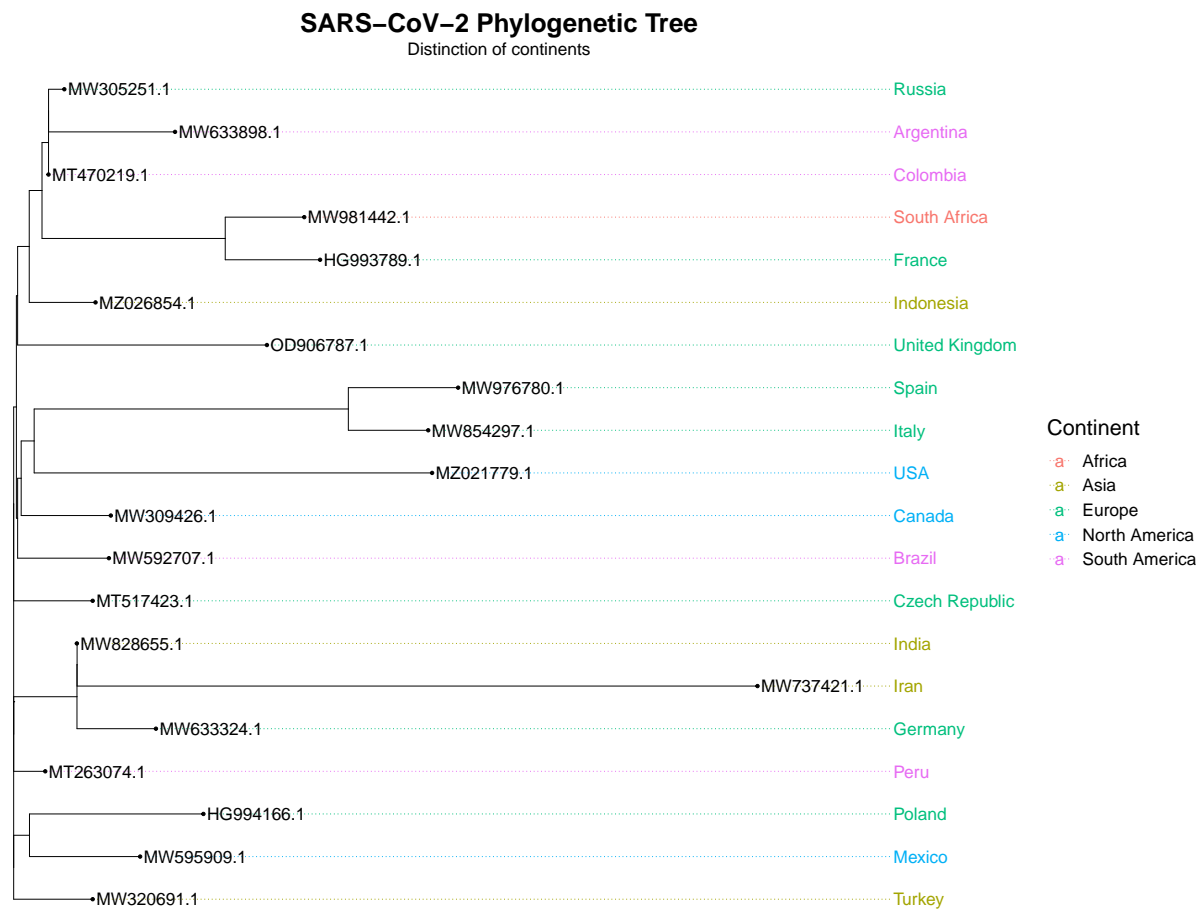
```



Data source: NCBI

Árbol filogenético por continentes

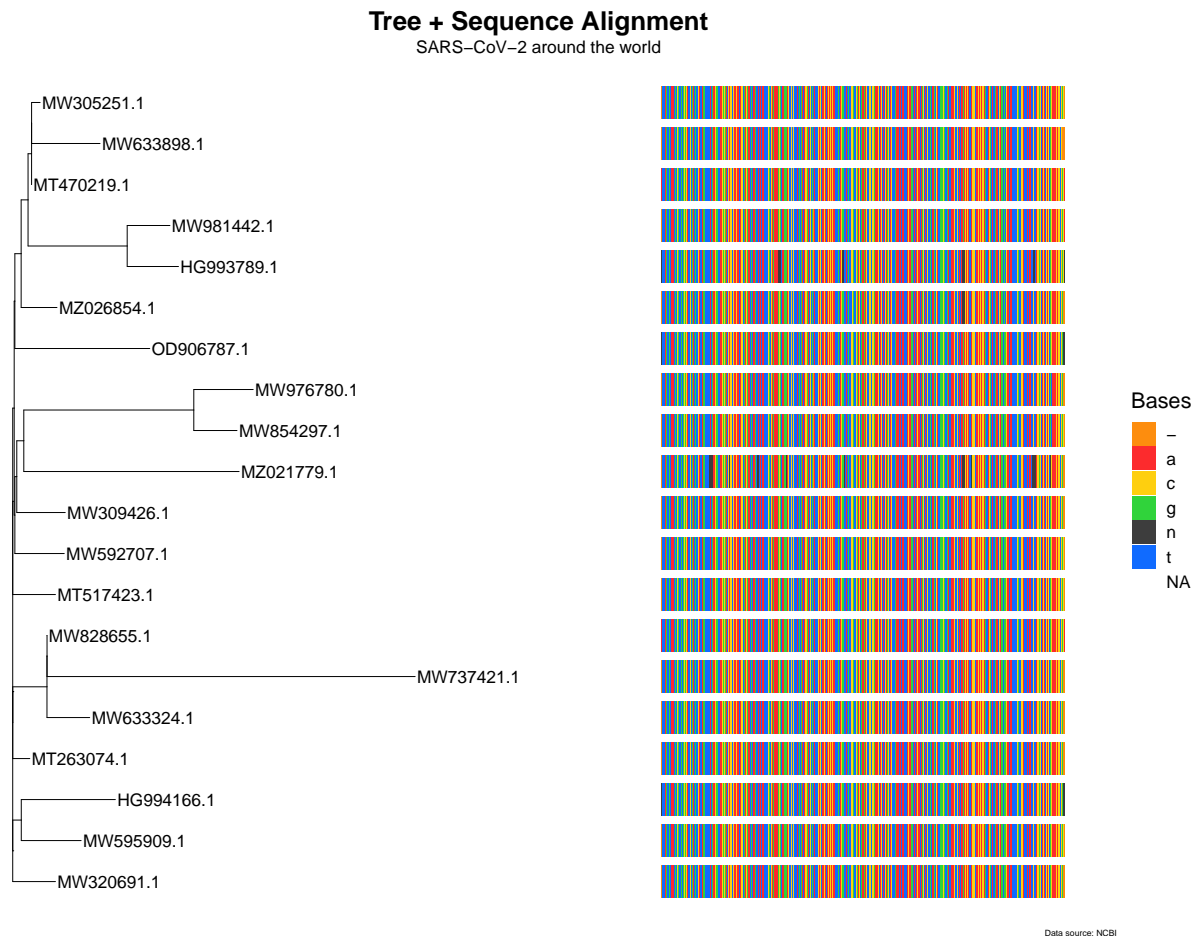
```
ggtree(tree) +
  xlim(0, 0.0022) +
  geom_tippoint() +
  geom_tiplab(
    aes(label = accessions[label], color = continents[label]),
    offset = 0.0003,
    align = TRUE,
    size = 7
  ) +
  geom_tiplab(size = 7) +
  labs(
    title = "SARS-CoV-2 Phylogenetic Tree",
    subtitle = "Distinction of continents",
    color = "Continent"
  ) + caption + title.theme + legend.theme
```



Data source: NCBI

Resultado final

```
msaplot(p = ggtree(tree), fasta = dna, color = base.colors, offset = 0.001) +  
  geom_tiplab(size = 7) +  
  labs(  
    title = "Tree + Sequence Alignment",  
    subtitle = "SARS-CoV-2 around the world",  
    fill = "Bases"  
  ) + caption + title.theme + legend.theme
```



Conclusión

Visualizando los resultados del análisis, se encuentra que, al menos para las muestras que fueron seleccionadas, **no hay una correlación grande entre la variante del virus y su localización geográfica**, pues en el árbol filogenético los países están mezclados entre ramas.

Aún así, las secuencias que se tomaron para cada país no son una muestra representativa, pues sólo es una por país, y fueron elegidas de manera aleatoria. Se podría hacer un análisis complementario que tomara en cuenta más muestras del virus, para comparar la distribución de los nodos con las variantes del virus y así encontrar con mayor certeza si las variantes del SARS-CoV-2 que se encuentran en los diferentes continentes son en realidad la misma variante o una diferente.

Referencias

- *Severe acute respiratory syndrome coronavirus 2 data hub*. Recuperado el 26 de abril del 2021, de https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Severe%20acute%20respiratory%20syndrome%20coronavirus%202,%20taxid:2697049
- *Tracking Covid-19's global spread*. Recuperado el 26 de abril del 2021, de <https://edition.cnn.com/interactive/2020/health/coronavirus-maps-and-cases/>