Evidencia2

Santiago Alducin Villaseñor - A01707122

2024-04-29

Diferencias en el código genético del SARS-CoV2 entre los continentes

El SARS-CoV2, mayormente conocido como Covid-19, fue una enfermedad que se propagó de manera exponencial alrededor del mundo, llegando a todos los continentes y causando una pandemia mundial, varios miles de muertes, y también un gran esfuerzo por miles de científicos para descubrir una vacuna contra esta enfermedad. Algunas de las complicaciones que se presentaron fueron debido a la facilidad de este patógeno viral a mutar, generando diferentes versiones de sí mismo con un código genético diferente entre las variantes pero con elementos en común. Por ello, el tema de estudio en esta ocación es cómo varían las variantes que más se propagaron en los 10 primeros países con más con más casos reportados. La lista de países fue extraída de la página de World Health Organization: 1. Estados Unidos: 103m 2. China: 99.4m 3. India: 45m 4. Francia: 39m 5. Alemania: 38.4m 6. Brasil: 37.5m 7. Korea: 34.6m 8. Japón: 33.8m 9. Italia: 26.7m 10. Reino Unido: 24.9m

Las variantes que se utilizarán para el análisis serán: (NCBI, s.f.) Asia: China, India, Korea América: Brasil, Estados Unidos Europa: Francia, Alemania, Italia África: Sudáfrica, Nigeria

Importar librerías y funciones

Las librerías que se utilizarán para este proyecto serán las siguientes. Estas ayudan a analizar la información, ordenarla y graficarla de la mejor manera posible. A continuación, también se crean las funciones necesarias que se utilizarán en otros momentos del código.

```
# BiocManager::install("ggtree")
# BiocManager::install("DECIPHER")
# install.packages("viridis")

suppressMessages(library(seqinr))
suppressMessages(library(adegenet))
suppressMessages(library(ape))
suppressMessages(library(ggtree))
suppressMessages(library(DECIPHER))
suppressMessages(library(viridis))
suppressMessages(library(ggplot2))


porcentajes_nucleotidos <- function(secuencia)
{
  library(stringr)
  cant_A <- sum(str_count(secuencia, "A"))
  cant_T <- sum(str_count(secuencia, "T"))
```

```
  cant_C <- sum(str_count(secuencia, "C"))
  cant_G <- sum(str_count(secuencia, "G"))
  total <- nchar(secuencia)

  porcentaje_A <- cant_A / total * 100
  porcentaje_T <- cant_T / total * 100
  porcentaje_C <- cant_C / total * 100
  porcentaje_G <- cant_G / total * 100

  return(c(cant_A,cant_T,cant_C,cant_G))
}
```

Código

Para empezar, tenemos que importar la información de un archivo fasta o texto que alamcene la información de los códigos genéticos a analizar. En este caso he creado un archivo con el conjunto de un virus de los países anteriormente mencionados para poder compararlos y analizarlos. Esta información tiene que ser procesada antes de poder graficarse, por ello orientamos y alineamos para posteriormente crear un nuevo archivo que tenga la información en una matriz completa y no cause ningún error al conseguir el distanciamiento o similitud de las versiones.

```
covid <- readDNAStringSet("covid_file.fasta", format = "fasta")

covid
```

```
## DNAStringSet object of length 10:
##      width seq                                                names
##  [1] 21291 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA NC_045512_China
##  [2] 21282 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA PP584641_Alemania
##  [3] 21291 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA PP693357_Brasil
##  [4] 21281 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA PP734341_EUA
##  [5] 21282 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA PP405601_Francia
##  [6] 21291 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA PP434597_India
##  [7] 21279 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA OR841414_Italia
##  [8] 21282 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA PP584656_Korea
##  [9] 21291 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA OR260866_Nigeria
## [10] 21291 ATGGAGAGCCTTGTCCCTGGTTT...TGATGTTCTTGTTAACAACTAA PP518586_Sud_Africa
```

```
suppressMessages(covid <- OrientNucleotides(covid))
```

```
## ================================================================================================================
##
## Time difference of 0.13 secs
```

```
suppressMessages(alineada <- AlignSeqs(covid))
```

```
## Determining distance matrix based on shared 11-mers:
## ================================================================================
##
## Time difference of 0.01 secs
```

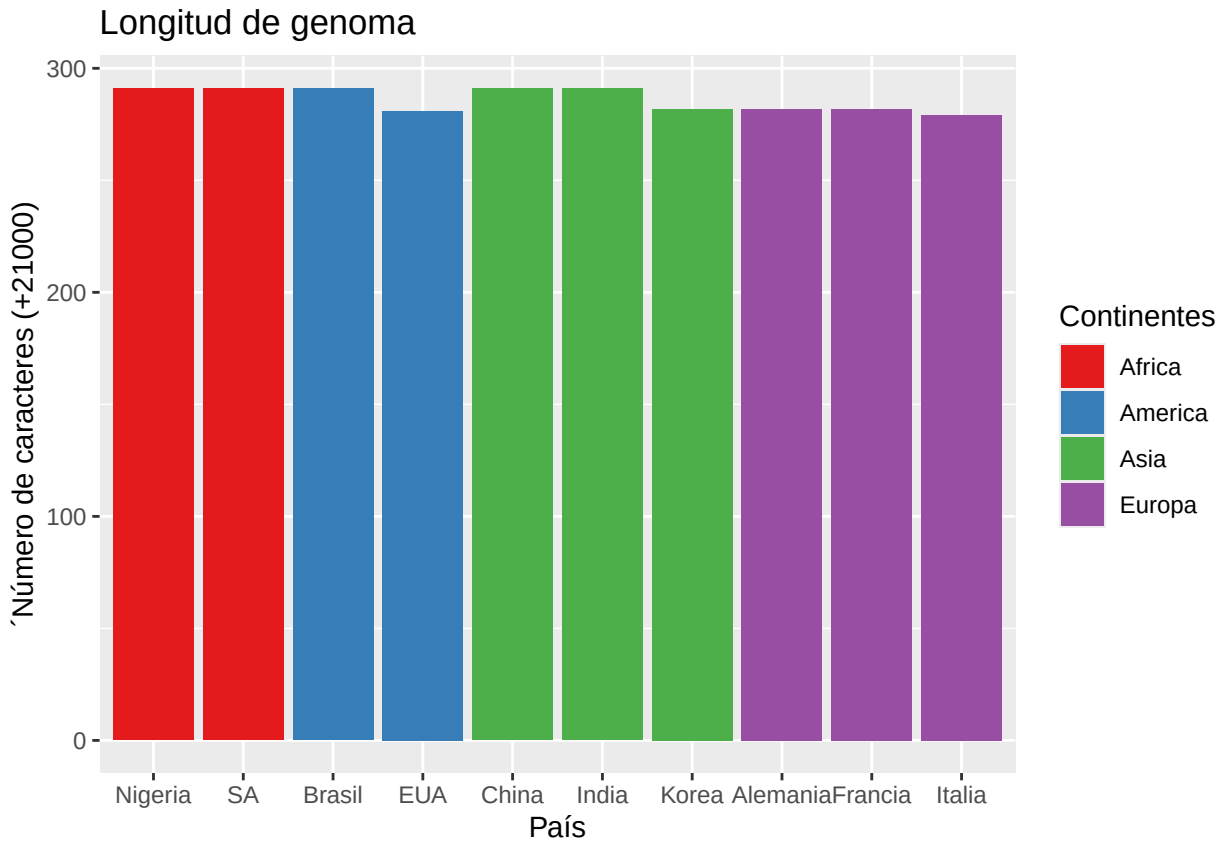
```
##
## Clustering into groups by similarity:
## ================================================================================
##
## Time difference of 0 secs
##
## Aligning Sequences:
## ================================================================================
##
## Time difference of 0.55 secs
##
## Iteration 1 of 2:
##
## Determining distance matrix based on alignment:
## ================================================================================
##
## Time difference of 0 secs
##
## Reclustering into groups by similarity:
## ================================================================================
##
## Time difference of 0 secs
##
## Realigning Sequences:
## ================================================================================
##
## Time difference of 0.11 secs
##
## Iteration 2 of 2:
##
## Determining distance matrix based on alignment:
## ================================================================================
##
## Time difference of 0 secs
##
## Reclustering into groups by similarity:
## ================================================================================
##
## Time difference of 0 secs
##
## Realigning Sequences:
## ================================================================================
##
## Time difference of 0 secs
```

```
writeXStringSet(alineada,
                # file="C:/Users/Santiago/Documents/R_Codes/Etapa2/covid.fasta")
                file="covid.fasta")
```

Longitud de las secuencias

Gracias a las secuencias de ADN obtenidas con anterioridad podemos conseguir transformarlo en un dataframe que contenga su longitud escalada para tener una visualización más agradable y poder mostrar la longitud de cada secuencia. De esta forma podemos ver la longiud estimada de cada continente.

```
secuencias <- as.character(covid)
long <- sapply(secuencias, nchar)
long <- long-21000

secuencias_df <- data.frame(
  Secuencias = c("China", "Alemania", "Brasil","EUA", "Francia", "India", "Italia", "Korea", "Nigeria",
  "SA"),
  # Secuencias <- names(long),
  Longitud = long,
  Continentes = c("Asia", "Europa", "America", "America", "Europa", "Asia", "Europa", "Asia", "Africa",
)

secuencia_df <- secuencias_df[order(secuencias_df$Continentes), ]

tabla <- ggplot(secuencia_df, aes(x = reorder(Secuencias, order(Continentes)), y = Longitud, fill = Con
  geom_bar(stat = "identity") +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "País", y = "´Número de caracteres (+21000)", title = "Longitud de genoma")
tabla
```

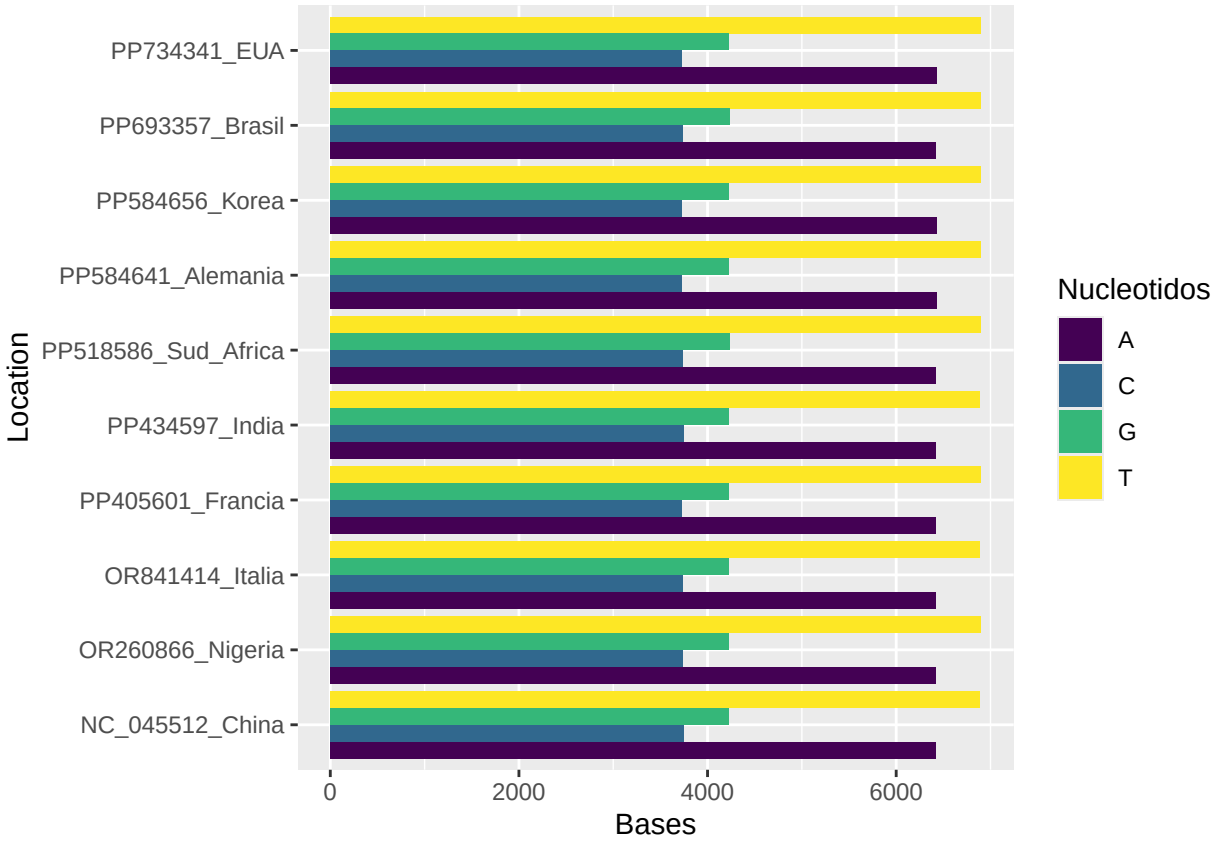


El resultado obtenido nos indica que las secuencias de ADN se mantiene constante a lo largo de países,

siendo los africanos los cuales tiene una secuencia un poco más larga, pero que no llega a ser nada fuera de lo esperado. Mientras que las europeas muestran secuencias un poco más cortas que las demás.

#Número de bases

```
covid_seq <- as.character(covid)
nucleotidos_df <- data.frame()
for(i in 1:length(covid_seq))
{
  nuc <- porcentajes_nucleotidos(covid_seq[[i]])
  nucleotidos <- data.frame("Location" = names(covid_seq[i]),
                            "Nucleotidos" = c("A", "T","C","G"),
                            "Bases" = c(nuc[1],nuc[2],nuc[3],nuc[4]))
  nucleotidos_df <- rbind(nucleotidos_df, nucleotidos)
}
num_bases <- ggplot(nucleotidos_df, aes(x = Bases, y = Location, fill = Nucleotidos)) +
  geom_bar(position = "dodge", stat = "identity") +
  scale_fill_viridis_d()

num_bases
```

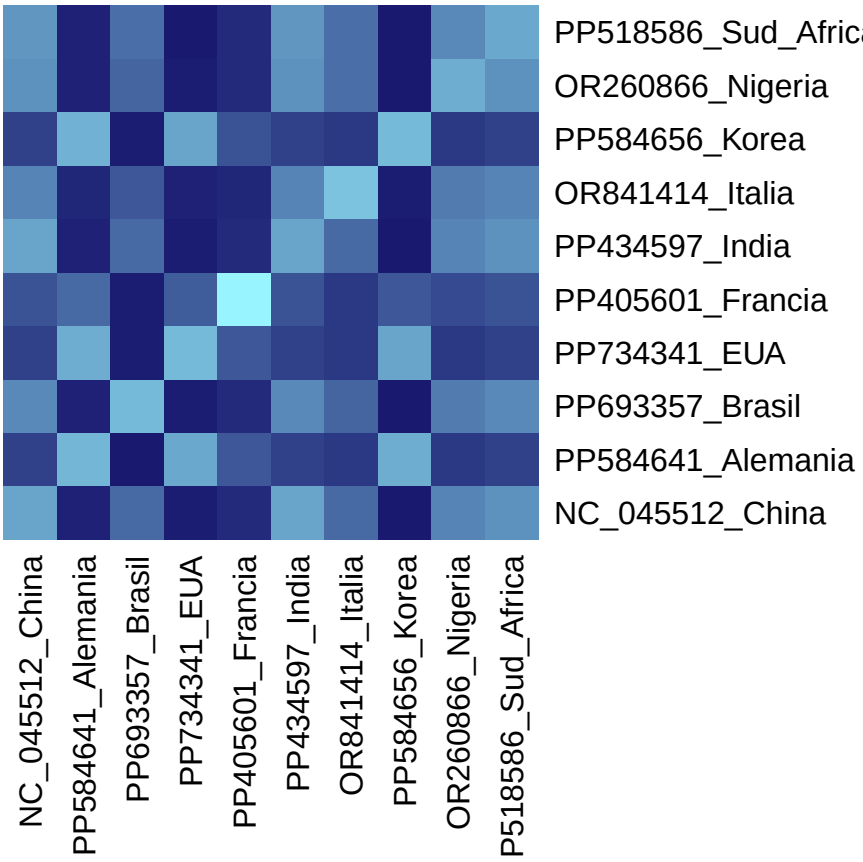


Como se puede observar en la gráfica, la cantidad de bases de cada tipo se mantiene casi constante en todas las variantes del virus alrededor del mundo. Esto tiene sentido considerando que se trata de un mismo virus y solo son variantes del mismo pero aún así merece la pena mencionar que los cambios parecen casi imperceptibles.

#Graficación mapa de similitudes A continuación se leerá el archivo creado con anterioridad para poder ser analizado con el modelo TN93, este es un modelo que toma en cuenta las diferencias de transición y

transversión dentro de la secuencia (TN93 model (Tamura-Nei, 93), (s.f.). Gracias a este modelo podemos identificar las secuencias que más se parecen y cuales son más diferentes.

```
dna <- read.dna("covid.fasta", format = "fasta")

D <- dist.dna(dna, model="TN93") #TN93 es un modelo en base a la probabilidad de cada nucleótido de mut

similitud <- as.matrix(D)

colores <- colorRampPalette(c("cadetblue1", "midnightblue"))
heatmap(similitud, Rowv = NA, Colv = NA, col =colores(50), margins =c(10, 10))
```



A pesar de que la cantidad de nucleótidos se mantiene casi igual entre todas las variantes, siguen siendo versiones diferentes del virus. En la gráfica superior se muestra la relación en similitudes que hay en estas variantes y cuales se parecen más entre ellas. Entre más oscuro significa que hay una menor similitud entre las variantes, mientras que entre más claro sea el color demuestra que se parecen en mayor proporción.

Árbol filogenético

A continuación podremos graficar un árbol filogenético utilizando el modelo de "nearest neighbour" el cual se puede formar con la secuencia de distanciamiento antes conseguida.

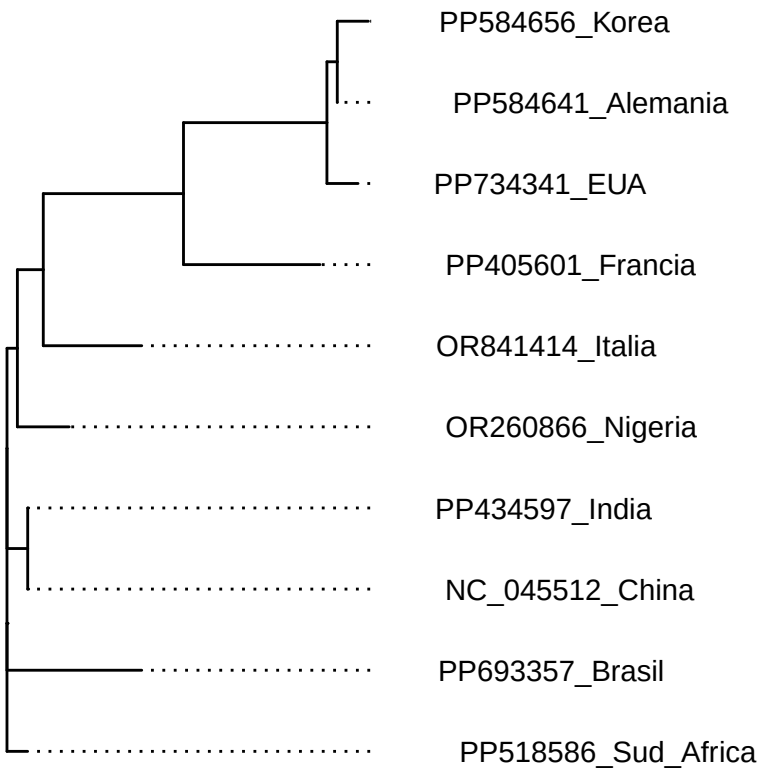
```
tree <- nj(D)

tree <- ladderize(tree)
```

```
myBoots <- boot.phylo(tree, dna, function(e) root(nj(dist.dna(e, model = "TN93")),1))
```

```
## Running bootstraps:       100 / 100
## Calculating bootstrap values... done.
```

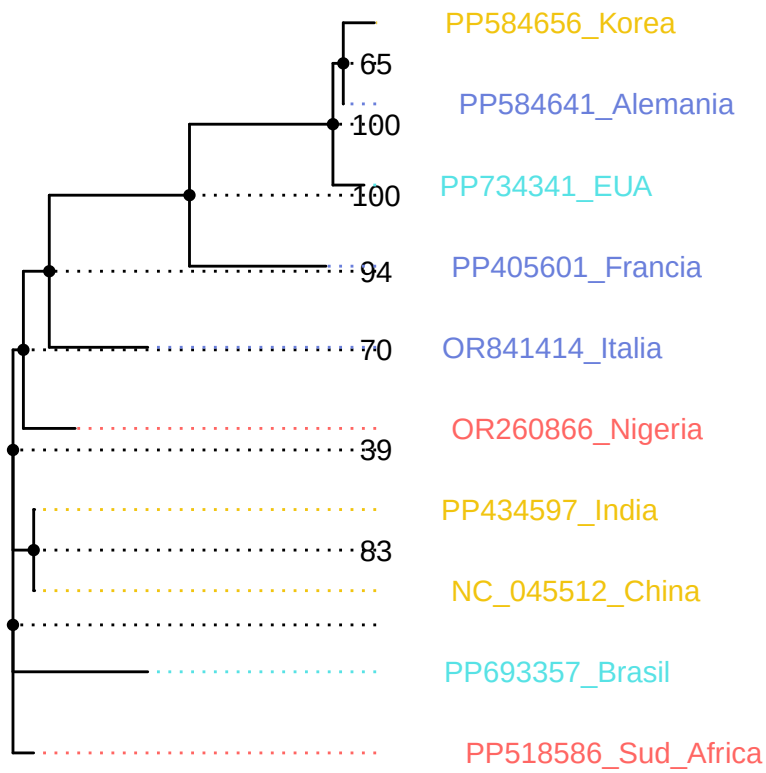
```
myBoots <- c(rep(0, times=10), myBoots)


ggtree(tree) +
  geom_tiplab(hjust = -0.3, size=4, align = TRUE)+
  xlim(0,.005)
```

```
## ! The tree contained negative edge lengths. If you want to ignore the edges,
## you can set `options(ignore.negative.edge=TRUE)`, then re-run ggtree.
```




```
tip_color_var = viridis(18)
ggtree(tree) +
  geom_nodepoint() +
  geom_nodelab(aes(label = myBoots), repel=TRUE, align=TRUE) +
  geom_tiplab(aes(color = tip_color_var), hjust = -0.3, size=4, align = TRUE) +
  scale_color_manual(values = c("#FF6663", "#FF6663", "#F1C40F", "#6a7fdb", "#F1C40F", "#6a7fdb", "#57e
  xlim(0, 0.005)
```

```
## ! The tree contained negative edge lengths. If you want to ignore the edges,
## you can set ‘options(ignore.negative.edge=TRUE)‘, then re-run ggtree.
```



Como se puede ver en la gráfica, se entiende que las versiones de china e india sean bastante parecidas, pues son bastante cercanos, en cambio también se puede ver que la versión de Korea cambia significativamente. En otra parte, podemos ver que las versiones de Europa se parecen demasiado, junto con la versión Americana de Estados Unidos. Esto puede tener bastante sentido siendo la mayoría países con bastante gente y donde sus habitantes tienen la facilidad de viajar entre los países diferentes países europeos, o Estados Unidos que siendo un país de primer mundo tiene más habitantes capaces de viajar a otros países, trayendo la enfermedad posiblemente de estos lugares europeos, siendo tambien estos los que menos variación tienen al hacer el bootstrapping. Los países africanos no tienen bastante relación al parecer, y Brasil también parece ser una cepa un poco alejada de las demás.

Video https://youtu.be/twulTWdJWhE

Referencias

World Health Organization. (2024) Number of COVID-19 cases reported to WHO (cumulative total). https://data.who.int/dashboards/covid19/cases NCBI Virus. (s.f.). Nih.Gov. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2697049 TN93 model (Tamura-Nei, 93). (s.f.). Man.Ac.Uk. Retrieved May 1, 2024, from http://www.bioinf.man.ac.uk/resources/phase/manual/node69.html