Evidencia 2 | Proyecto integrador

Daniel Barreras, Yair Beltrán, Víctor Symonds

2023-05-07

Uso de paquetes de R

library(viridis)

## Warning: package 'viridis' was built under R version 4.2.3

## Loading required package: viridisLite

## Warning: package 'viridisLite' was built under R version 4.2.3

library(Biostrings)

## 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, aperm, 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: S4Vectors

## Loading required package: stats4

##   
## Attaching package: 'S4Vectors'

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

## Loading required package: IRanges

##   
## Attaching package: 'IRanges'

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

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##   
## Attaching package: 'Biostrings'

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

library(DECIPHER)

## Loading required package: RSQLite

## Warning: package 'RSQLite' was built under R version 4.2.3

## Loading required package: parallel

library(ade4)

## Warning: package 'ade4' was built under R version 4.2.3

##   
## Attaching package: 'ade4'

## The following object is masked from 'package:Biostrings':  
##   
## score

## The following object is masked from 'package:BiocGenerics':  
##   
## score

library(seqinr)

## Warning: package 'seqinr' was built under R version 4.2.3

##   
## Attaching package: 'seqinr'

## The following object is masked from 'package:Biostrings':  
##   
## translate

library(adegenet)

## Warning: package 'adegenet' was built under R version 4.2.3

##   
## /// adegenet 2.1.10 is loaded ////////////  
##   
## > overview: '?adegenet'  
## > tutorials/doc/questions: 'adegenetWeb()'   
## > bug reports/feature requests: adegenetIssues()

library(ape)

## Warning: package 'ape' was built under R version 4.2.3

##   
## Attaching package: 'ape'

## The following objects are masked from 'package:seqinr':  
##   
## as.alignment, consensus

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

library(ggtree)

## ggtree v3.6.2 For help: https://yulab-smu.top/treedata-book/  
##   
## If you use the ggtree package suite in published research, please cite  
## the appropriate paper(s):  
##   
## Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam.  
## ggtree: an R package for visualization and annotation of phylogenetic  
## trees with their covariates and other associated data. Methods in  
## Ecology and Evolution. 2017, 8(1):28-36. doi:10.1111/2041-210X.12628  
##   
## S Xu, Z Dai, P Guo, X Fu, S Liu, L Zhou, W Tang, T Feng, M Chen, L  
## Zhan, T Wu, E Hu, Y Jiang, X Bo, G Yu. ggtreeExtra: Compact  
## visualization of richly annotated phylogenetic data. Molecular Biology  
## and Evolution. 2021, 38(9):4039-4042. doi: 10.1093/molbev/msab166  
##   
## Guangchuang Yu, Tommy Tsan-Yuk Lam, Huachen Zhu, Yi Guan. Two methods  
## for mapping and visualizing associated data on phylogeny using ggtree.  
## Molecular Biology and Evolution. 2018, 35(12):3041-3043.  
## doi:10.1093/molbev/msy194

##   
## Attaching package: 'ggtree'

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

## The following object is masked from 'package:Biostrings':  
##   
## collapse

## The following object is masked from 'package:IRanges':  
##   
## collapse

## The following object is masked from 'package:S4Vectors':  
##   
## expand

library(viridis)  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

Declaración de variables de las variantes de Coronavirus detectadas en animales

names\_sars\_cov\_2 <- c("SARS-CoV-2", "RaTG13", "RmYN02",  
 "Pangolin-CoV", "HKU2", "MERS-CoV",  
 "HCoV-NL63", "HCoV-OC43", "HCoV-HKU1", "SARS-CoV")  
  
  
variants\_coronavirus <- c("MN908947", "MT121215", "NC\_045512",   
 "MT121216", "NC\_004718", "NC\_019843",  
 "AY567487", "NC\_006213", "NC\_006577", "NC\_004718")  
  
names(variants\_coronavirus) <- names\_sars\_cov\_2

Obtén las secuencias de los genomas de los virus elegidos según la investigación que hayas decidido realizar. Para ello utiliza la función read.GenBank para obtener los genomas directamente del NCBI desde R Studio. Muestra el código empleado para obtenerlos.

coronavirus\_sequences <- read.GenBank(variants\_coronavirus)

Realiza el alineamiento de los genomas virales y visualiza el resultado de tu alineamiento en tu navegador web. Muestra el código empleado para realizar lo anterior e incluye dos imágenes con el resultado del alineamiento, una de los primeros 150 nucleótidos y otra de los nucleótidos 500 al 650.

# Escribir las secuencias en formato fasta  
write.dna(coronavirus\_sequences, file = "coronavirus\_sequences.fasta", format = "fasta")  
  
# Leer las secuencias en formato fasta  
coronavirus\_string\_set <- readDNAStringSet("coronavirus\_sequences.fasta", format = "fasta")

## Warning in .Call2("fasta\_index", filexp\_list, nrec, skip, seek.first.rec, :  
## reading FASTA file coronavirus\_sequences.fasta: ignored 24761 invalid  
## one-letter sequence codes

# Orientar las secuencias  
oriented\_coronavirus\_string\_set <- OrientNucleotides(coronavirus\_string\_set)

## ========================================================================================================================================================================================================  
##   
## Time difference of 0.19 secs

# Alinear las secuencias   
aligned\_coronavirus\_string\_set <- AlignSeqs(oriented\_coronavirus\_string\_set)

## Determining distance matrix based on shared 12-mers:  
## ================================================================================  
##   
## Time difference of 0.75 secs  
##   
## Clustering into groups by similarity:  
## ================================================================================  
##   
## Time difference of 0 secs  
##   
## Aligning Sequences:  
## ================================================================================  
##   
## Time difference of 39.02 secs  
##   
## Iteration 1 of 2:  
##   
## Determining distance matrix based on alignment:  
## ================================================================================  
##   
## Time difference of 0.01 secs  
##   
## Reclustering into groups by similarity:  
## ================================================================================  
##   
## Time difference of 0 secs  
##   
## Realigning Sequences:  
## ================================================================================  
##   
## Time difference of 0.01 secs  
##   
## Alignment converged - skipping remaining iteration.  
##   
## Refining the alignment:  
## ================================================================================  
##   
## Time difference of 53.82 secs

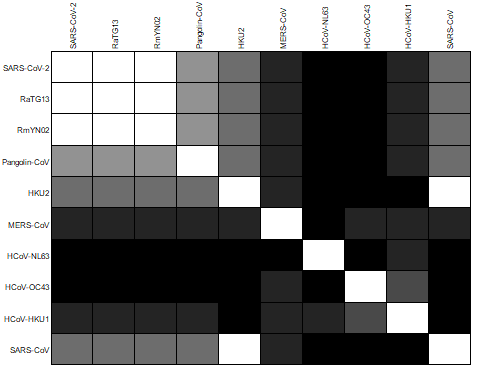
# Se eligen nucleotidos determinados  
first\_150\_nucleotides <- subseq(coronavirus\_string\_set, start = 1, end = 150)  
  
names(first\_150\_nucleotides) <- names\_sars\_cov\_2  
  
nucleotides\_500\_650 <- subseq(coronavirus\_string\_set, start = 500, end = 650)  
  
names(nucleotides\_500\_650) <- names\_sars\_cov\_2  
  
  
BrowseSeqs(first\_150\_nucleotides)  
BrowseSeqs(nucleotides\_500\_650)

Agrega una interpretación escrita, desde el punto de vista biológico, para esta gráfica

En el árbol filogenético, cada rama representa una conexión evolutiva entre las variantes del virus. Las ramas más cercanas entre sí indican una mayor similitud genética, lo que sugiere una relación más estrecha en términos de descendencia común. Por otro lado, las ramas más separadas representan una mayor divergencia genética y pueden indicar eventos de evolución independiente. Las etiquetas en los extremos de las ramas representan las variantes del virus y están asociadas con los nombres comunes asignados a cada una de ellas. Estos nombres proporcionan una forma más fácil de identificar y referirse a las diferentes variantes del SARS-CoV-2 encontradas en animales. Al observar el árbol filogenético, podemos identificar agrupaciones o clados de variantes del virus que comparten una mayor similitud genética. Estos clados pueden indicar la existencia de linajes virales específicos que se han propagado en poblaciones animales y pueden tener implicaciones en la transmisión y la adaptación del virus.

Genera una matriz de distancia a partir de los genomas alineados. Crea una tabla en escala de grises en la que observes de manera visual el resultado de la matriz de distancia e inclúyela en tu reporte. Muestra el código empleado para obtener lo anterior e incluye la matriz de distancia y la tabla que obtuviste.

# Escribe las secuencias alineadas en formato fasta  
writeXStringSet(aligned\_coronavirus\_string\_set, "coronavirus\_sequences\_aligned.fasta", format = "fasta")  
  
# Lee las secuencias alineadas desde el archivo fasta  
alignment\_coronavirus <- read.alignment(file = "coronavirus\_sequences\_aligned.fasta", format = "fasta")  
  
# Calcula la matriz de distancia a partir del alineamiento  
distance\_coronavirus <- dist.alignment(alignment\_coronavirus, matrix = "similarity")  
  
# Convierte la matriz de distancia en un data frame  
temp <- as.data.frame(as.matrix(distance\_coronavirus))  
  
# Establece los nombres de las filas y columnas del data frame como los nombres de los virus  
rownames(temp) <- names\_sars\_cov\_2  
colnames(temp) <- names\_sars\_cov\_2  
  
# Crea una tabla en escala de grises para visualizar la matriz de distancia  
grayscale\_table <- table.paint(temp, cleg = 0, clabel.row = 0.5, clabel.col = 0.5) + scale\_color\_viridis()

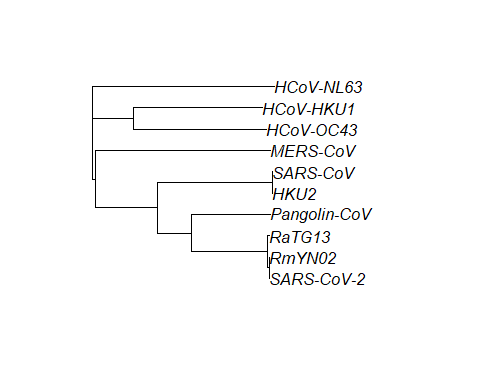


Agrega una interpretación escrita, desde el punto de vista biológico, para esta gráfica

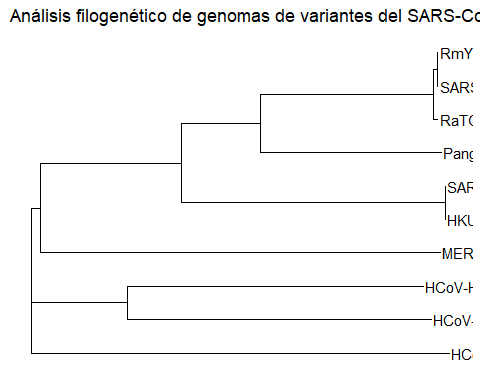
La gráfica generada muestra una representación visual de la matriz de distancias entre los genomas virales alineados. Cada casilla de la tabla representa la similitud entre dos genomas, donde los colores más oscuros indican una mayor similitud y los colores más claros indican una menor similitud. Al observar la gráfica, podemos identificar patrones de similitud y diferencias entre los genomas. Las regiones con colores oscuros indican una alta similitud, lo que sugiere que esos genomas comparten secuencias conservadas. Por otro lado, las regiones con colores claros indican diferencias entre los genomas, lo que puede corresponder a regiones únicas o variantes genéticas. El análisis de la matriz de distancias nos permite identificar relaciones evolutivas entre los genomas virales y comprender su diversidad genética. Esto es relevante para estudiar la transmisión y adaptación de los virus entre especies, así como para comprender la posible emergencia de nuevas variantes. Una comprensión más detallada de la matriz de distancias requeriría análisis adicionales y comparaciones con genomas de referencia.

“Construye un árbol filogenético a partir de la matriz de distancia obtenida e incluye en el árbol los números de accesión de los genomas utilizados, sus nombres comunes o cualquier otra leyenda que te permita indicar la ubicación de cada virus en el árbol. Muestra el código empleado para realizar lo anterior e incluye la imagen del árbol filogenético.

# Genera un árbol filogenético utilizando el método nj   
phylogenetic\_tree\_coronavirus <- nj(distance\_coronavirus)  
phylogenetic\_tree\_coronavirus <- ladderize(phylogenetic\_tree\_coronavirus)  
  
# Obtiene los nombres originales de las puntas (hojas) del árbol y asigna nuevos  
original\_tip\_labels <- rownames(phylogenetic\_tree\_coronavirus$tip.label)  
tip\_labels <- setNames(names\_sars\_cov\_2, original\_tip\_labels)  
  
# Crea una copia modificada del árbol filogenético y la actualiza  
modified\_phylogenetic\_tree <- phylogenetic\_tree\_coronavirus  
modified\_phylogenetic\_tree$tip.label <- tip\_labels  
  
# Visualiza el árbol filogenético con las etiquetas de las puntas  
plot.phylo(modified\_phylogenetic\_tree, show.tip.label = TRUE)



# Muestra el arbol filognético  
plot\_virus\_phylogeny <- ggtree(modified\_phylogenetic\_tree) + geom\_tiplab() + ggtitle("Análisis filogenético de genomas de variantes del SARS-CoV-2 encontrados en animales")  
  
plot\_virus\_phylogeny



Agrega una interpretación escrita, desde el punto de vista biológico, para esta gráfica

En el árbol filogenético, cada rama representa una conexión evolutiva entre las variantes del virus. Las ramas más cercanas entre sí indican una mayor similitud genética, lo que sugiere una relación más estrecha en términos de descendencia común. Por otro lado, las ramas más separadas representan una mayor divergencia genética y pueden indicar eventos de evolución independiente. Las etiquetas en los extremos de las ramas representan las variantes del virus y están asociadas con los nombres comunes asignados a cada una de ellas. Estos nombres proporcionan una forma más fácil de identificar y referirse a las diferentes variantes del SARS-CoV-2 encontradas en animales. Al observar el árbol filogenético, podemos identificar agrupaciones o clados de variantes del virus que comparten una mayor similitud genética. Estos clados pueden indicar la existencia de linajes virales específicos que se han propagado en poblaciones animales y pueden tener implicaciones en la transmisión y la adaptación del virus.