Clustering y Clasificación taxonómica de virus

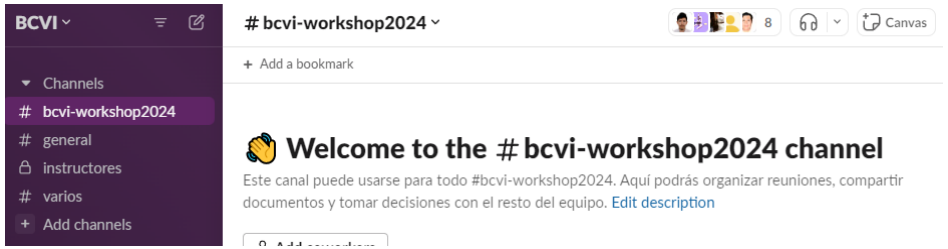
Laura Carolina Camelo Valera

McGill University (Maurice Lab)

Links a utilizar

- https://github.com/BCVI/Taller-Virtual-Bioinformatica-Genomica-Viral

- https://join.slack.com/t/bcviespacio/shared_invite/zt-2e2j9f6og-OjpP4I5QozHVMjOX8WIQzg

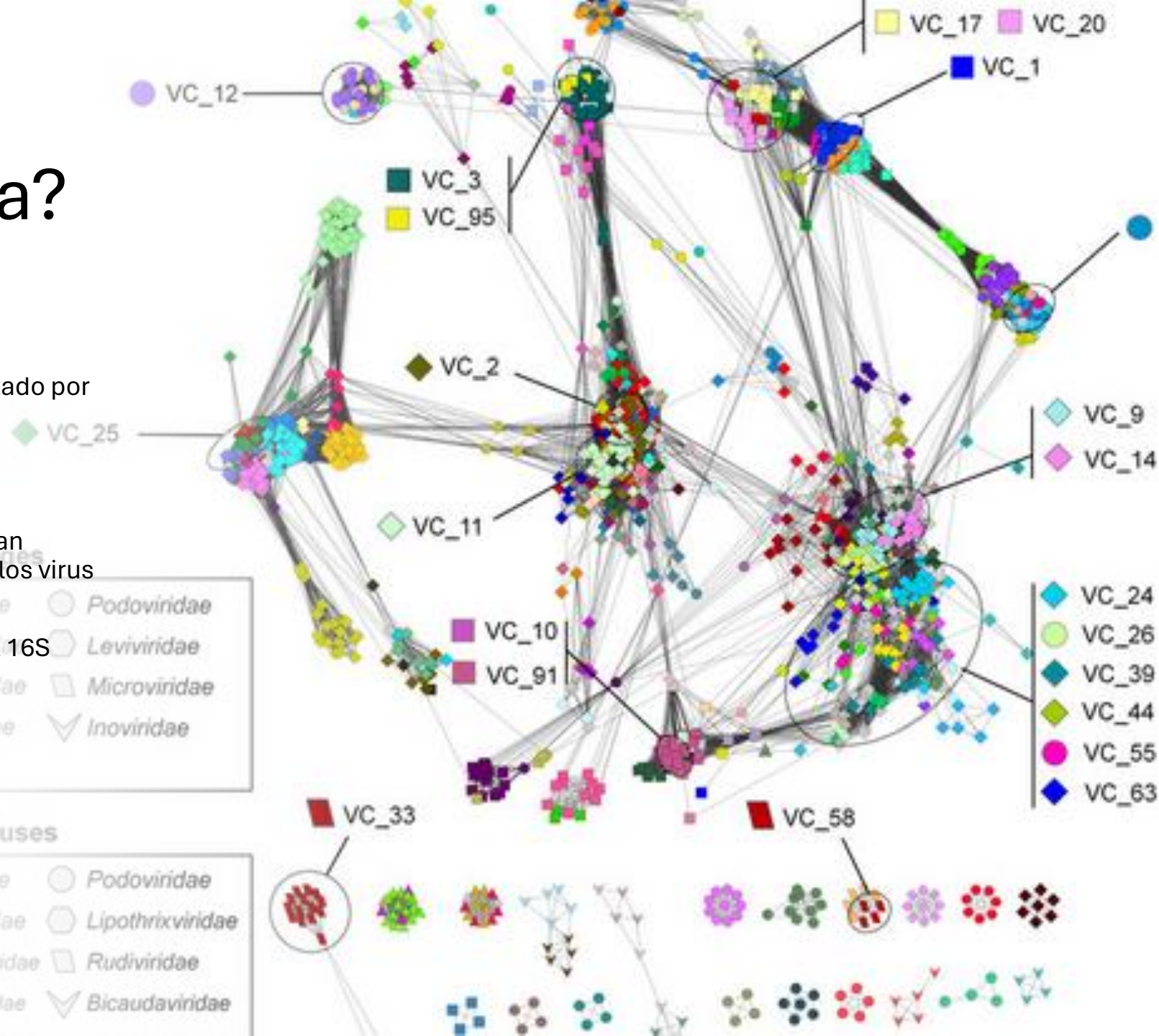


Por qué taxonomía?

Diversidad viral

- Cada dominio taxonómico puede ser infectado por virus
  - Archaea, Bacteria y Eucariota.

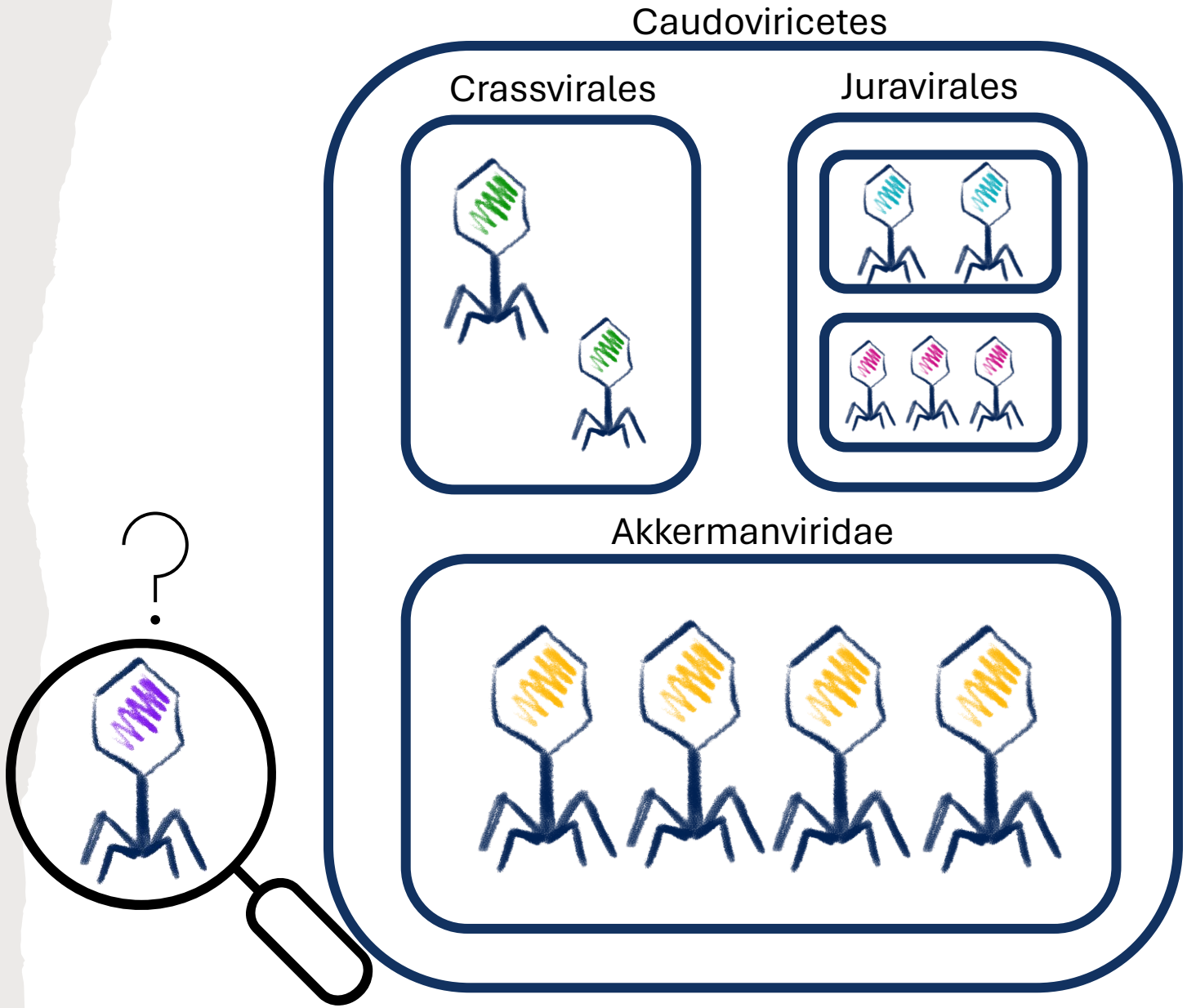
- Cada dominio contiene en sí mismo una gran diversidad de organismos, la diversidad de los virus ha de ser mayor.
- En virus no existe un gen marcador como el 16S
- Como se clasifican los virus?
  - Aislamiento de virus (antes)
  - Secuenciacion
  - Clustering
  - Anotacion



Bolduc et al., et al. 2017. PeerJ
Weiltz, J.S. 2016. Princeton University Press

Como se clasifican los virus?

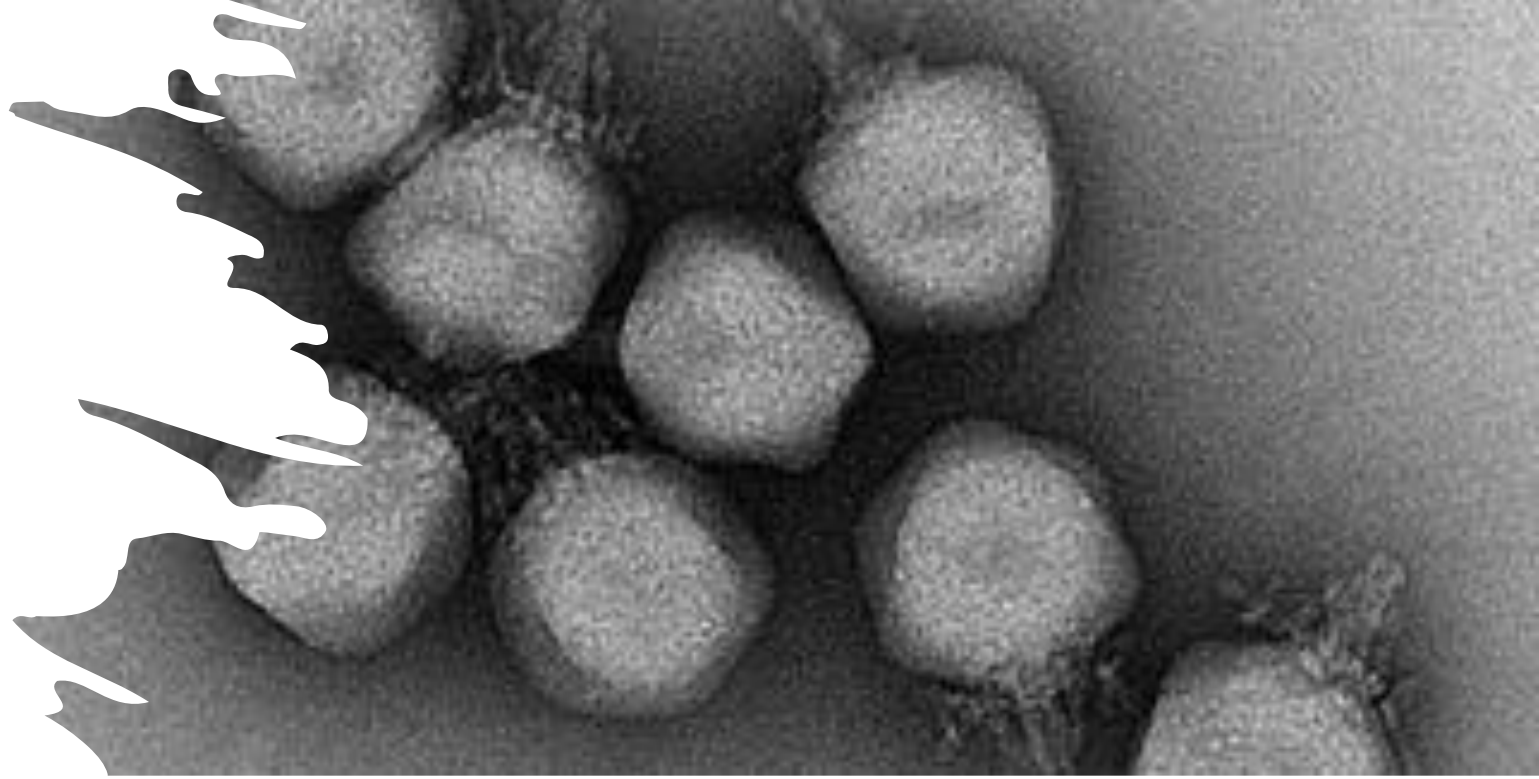
- Clasificar
  - Por similitud de rasgos.
- Organizar
  - Dependiendo de la similitud. Donde lo agrupo?
- Anotar
  - Asignarle un nombre.

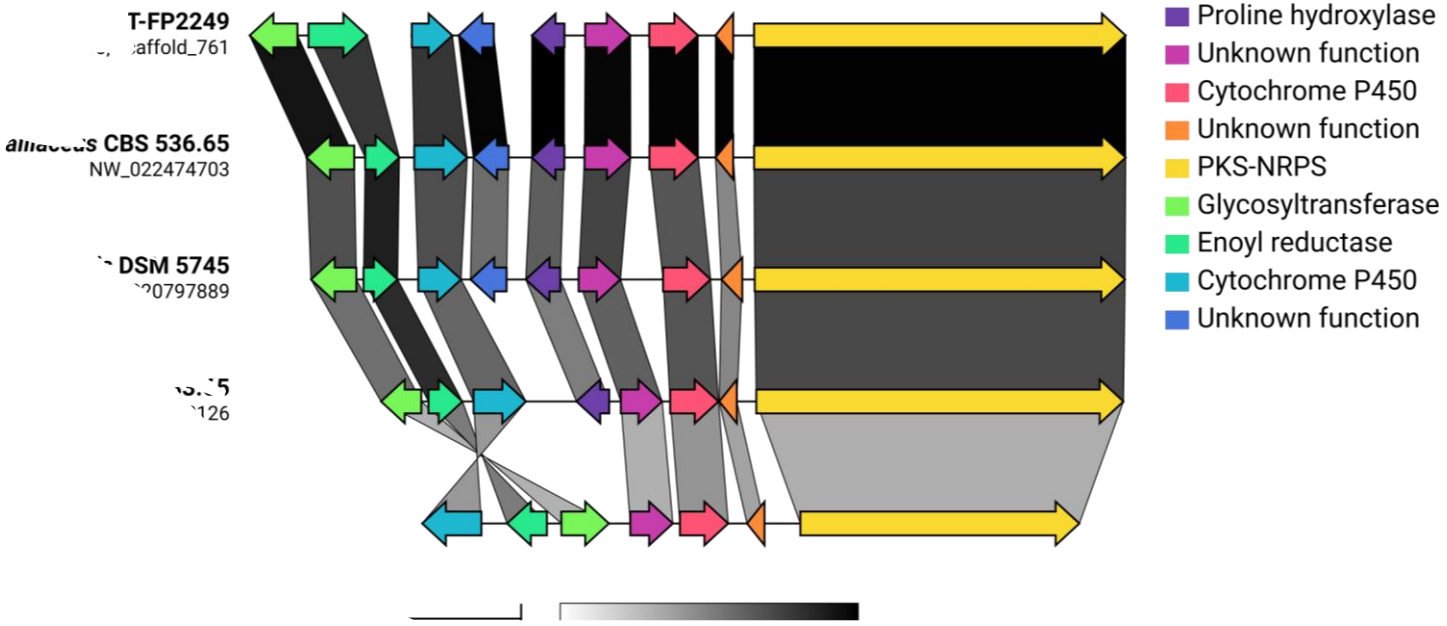


CLASIFICAR Y ORGANIZAR

- Morfología del virión
- Región específica del genoma (ex. gen).
- Totalidad del genoma (Recomendado por ICTV*)
  - A nivel de nucleótidos (a nivel de especie)
  - A nivel de proteínas (a nivel mayores que genero)
  - Organizacion del genoma (sintenia)

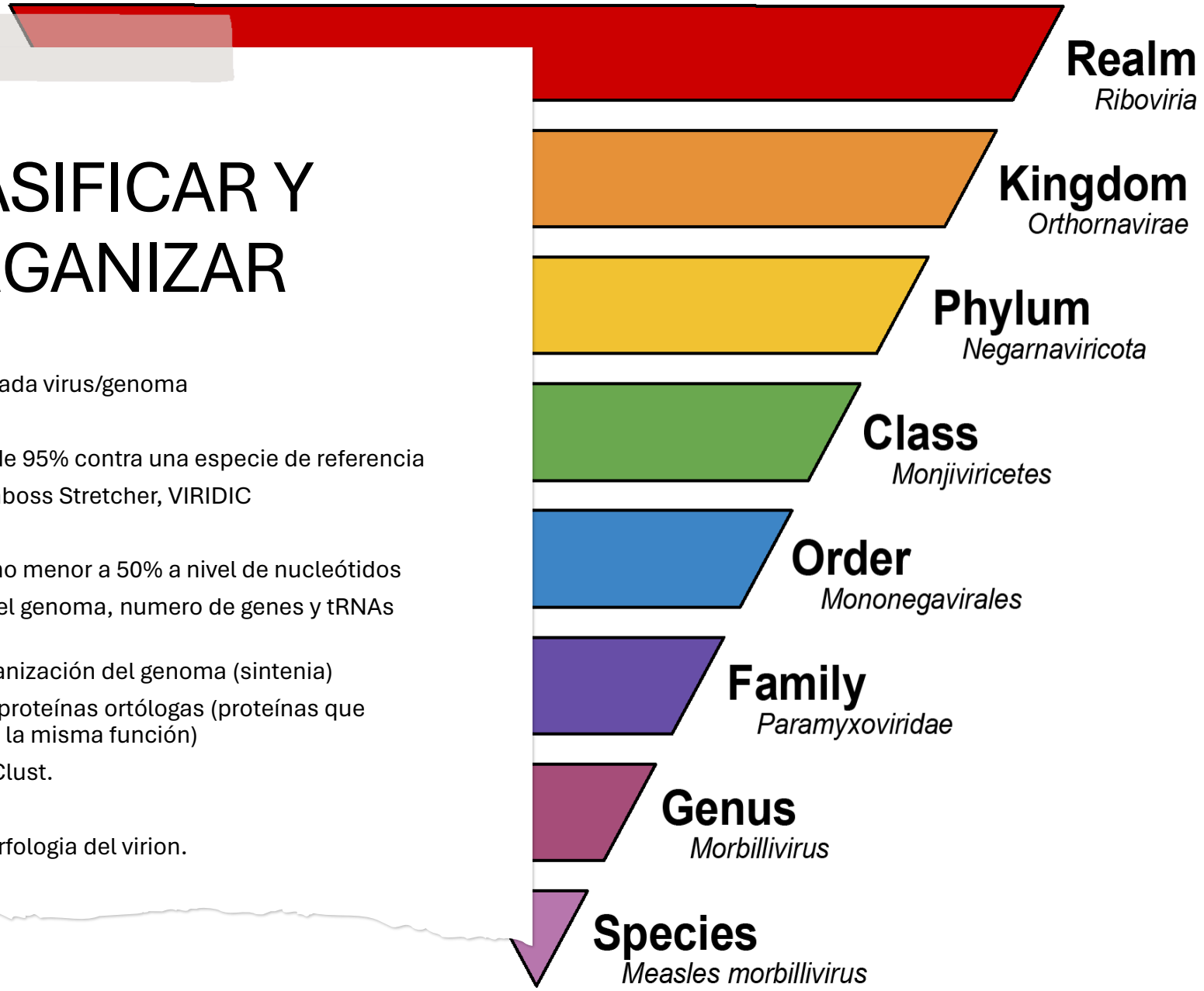
*ICTV. International Committee on Taxonomy of Viruses.





CLASIFICAR Y ORGANIZAR

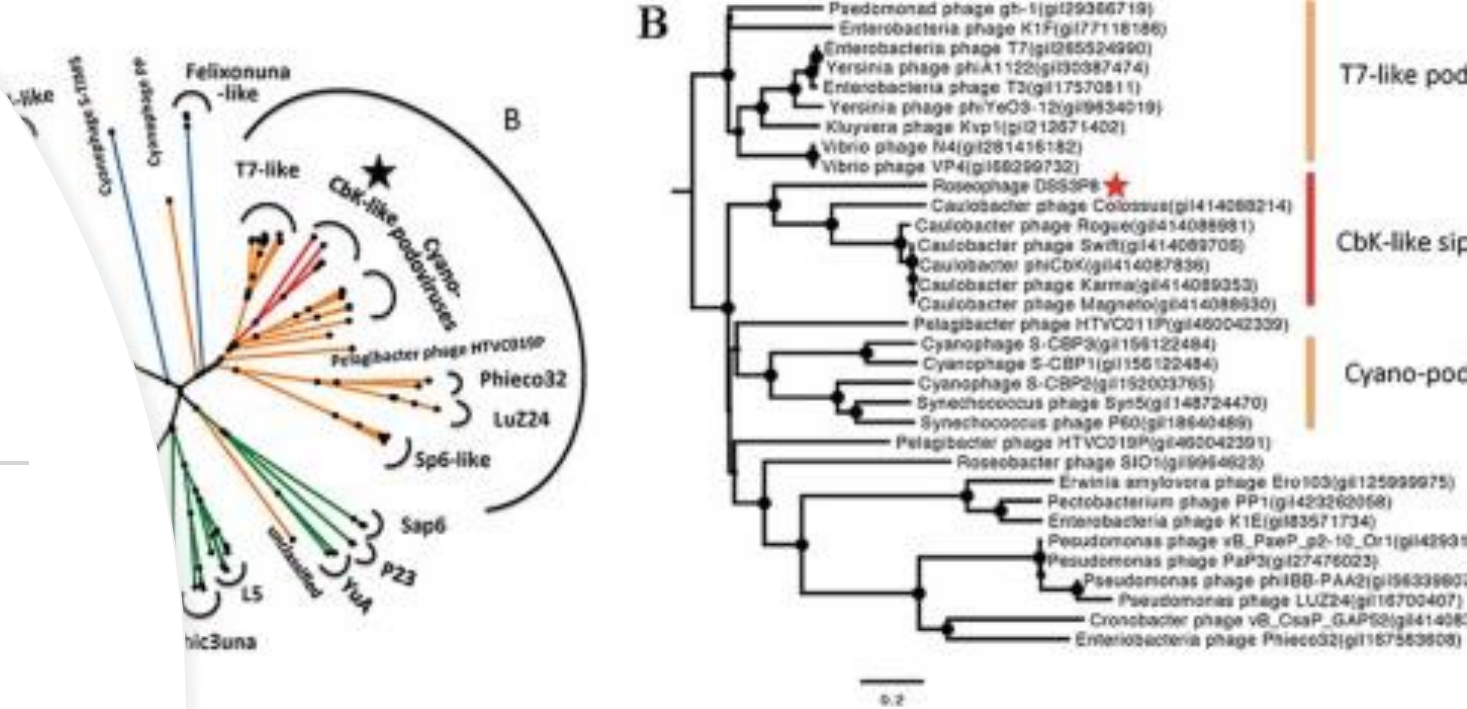
- Definir en que grupo va cada virus/genoma
  - Especie
    - Identidad de 95% contra una especie de referencia
    - BlastN, Emboss Stretcher, VIRIDIC
  - Genero
    - Identidad no menor a 50% a nivel de nucleótidos
    - Longitud del genoma, numero de genes y tRNAs similar
    - Misma organización del genoma (sintenia)
    - Compartir proteínas ortólogas (proteínas que comparten la misma función)
    - BlastP, VirClust.
  - Familia
    - Similar morfologia del virion.

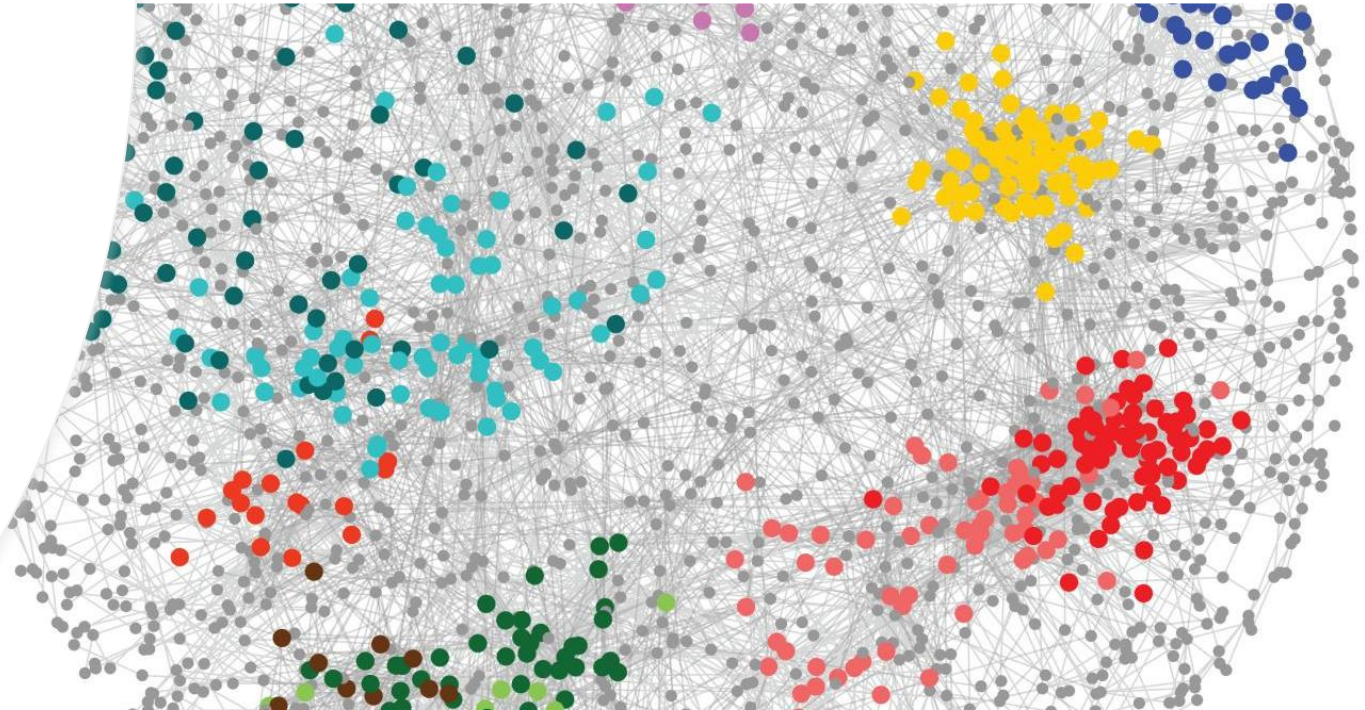


ORGANIZAR

- Una vez se tienen los datos de similitud se puede usar el agrupamiento (*clustering)* para ordenar los genomas/genes en arboles filogenéticos o dendrogramas

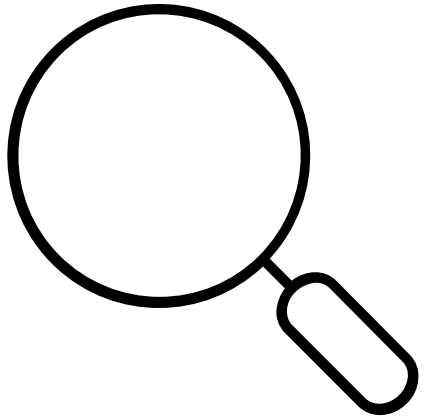
    - Clustering jerárquico (VIRIDIC, y librerías en R como ape y phangorn)
    - vConTACT2
    - Virclust



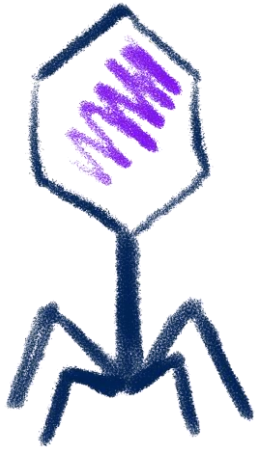


ANOTAR

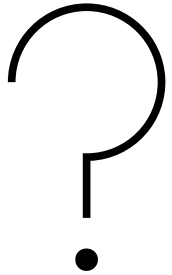
- Una vez se hayan clasificado y organizado las secuencias en grupo se debe asignar un nombre a la categoría. Generalmente se elige el nombre del grupo más cercano que cumpla con los requisitos de similitud.



Si usted tiene un fago que se parece al fago Pseudoalteromonas phage Cr39582. Cuyo linaje es: Viruses; Varidnaviria; Bamfordvirae; Preplasmiviricota; Tectiliviricetes; Vinavirales; Corticoviridae; Corticovirus; Corticovirus Cr39582

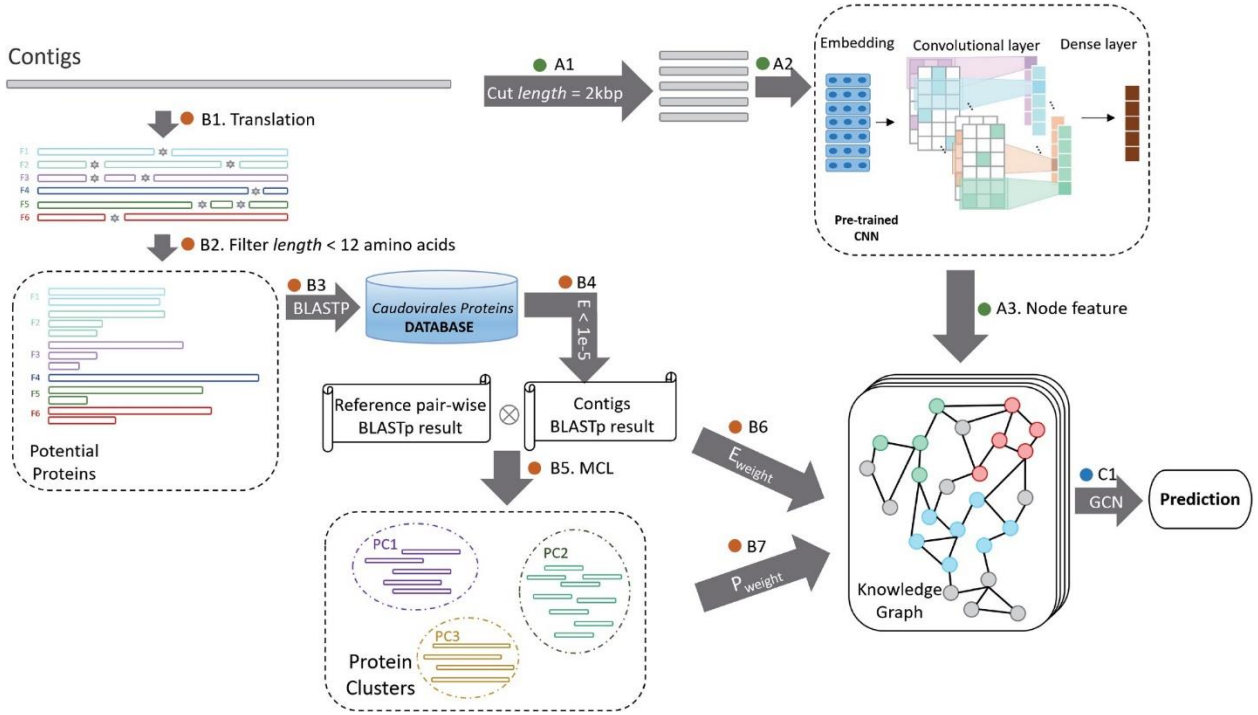


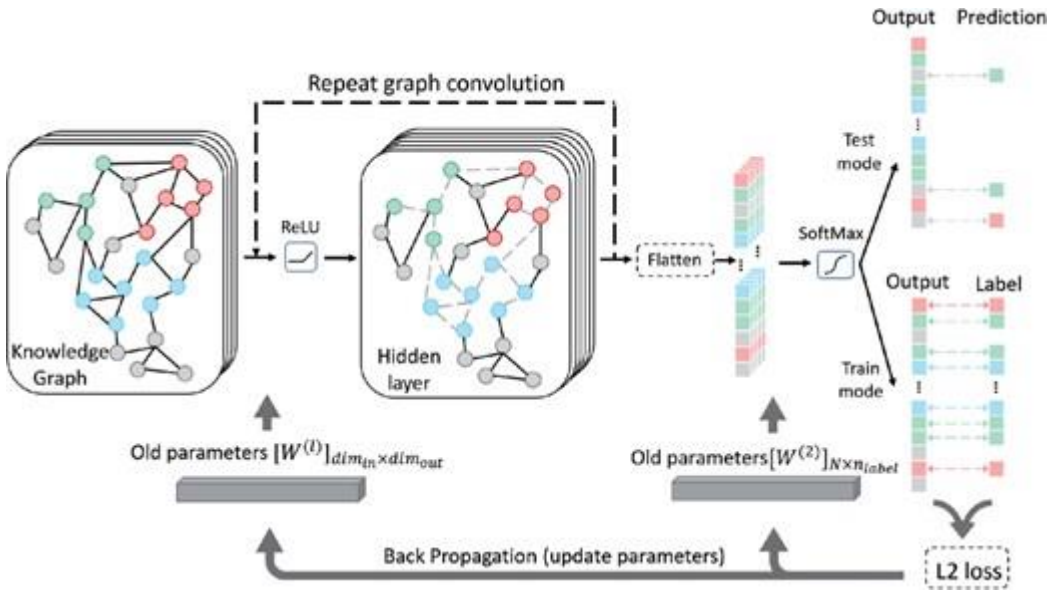
- Si alinea más del 95% y su fracción de alineamiento es mayor a 85%
  - Asignarle misma especie: Corticovirus Cr39582
- Si alinea menos de 95% pero más del 50%, tiene una longitud similar a genomas del genero Corticovirus y tiene un gen asociado a dicho genero (proteoma similar)
  - Asignarle el mismo genero: Corticovirus
- Si alinea menos de 50%, dicho fago no pertenece a ese grupo de virus y por lo tanto permanece sin nombre.



ANOTAR

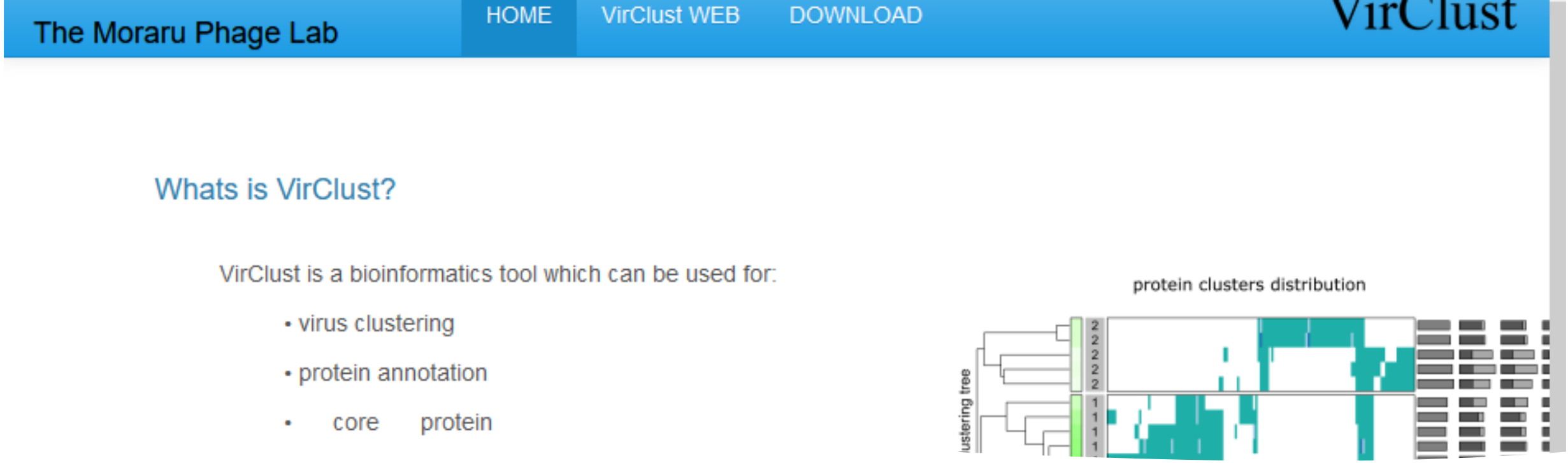
- PhageGCN es una herramienta que ayuda a anotar, su entrada son solo los genomas.





Ejercicio - Comparación de genomas.

- Virus Intergenomic Distance Calculator - VIRIDIC (https://rhea.icbm.uni-oldenburg.de/viridic/) – pensado para virus de Archaea y Bacteria, pero puede usarse para genomas monopartitos de virus eucariotas
  - Similitud, fracción alineada, longitud de los genomas
  - Genera grupos de genomas (hasta 65% similitud), clustering jerarquico
- Entrada = un único archivo fasta con todas las secuencias, agregar un nombre de usuario de as de 6 letras modificar los parámetros
- Salida = comparación de los genomas



Ejercicio Comparacion de proteomas

- Virus Intergenomic Distance Calculator - VirClust (https://rhea.icbm.uni-oldenburg.de/virclust/) – pensado únicamente para virus de Archaea y Bacteria
  - Predice los genes en los genomas (anotación)
  - Genera grupos ortólogos de proteínas genomas
  - Usa clustering jerárquico para realizar agrupamiento
  - Identifica proteínas core (presentes en todos los genomas usados)
- Entrada = un único archivo fasta con todas las secuencias, agregar un nombre de usuario de as de 6 letras modificar los parámetros
- Salida = genes en cada genoma, grupos ortólogos de proteínas, proteínas core, árbol.

HANDS-ON