



La salud
es de todos

Minsalud



INSTITUTO NACIONAL DE SALUD

Ciencia, Tecnología e Innovación

Visítenos en
www.ins.gov.co





INSTITUTO
NACIONAL DE
SALUD

Anotación de genomas

Mauricio Pacheco-Montealegre
Magister en Biología Computacional

Unidad de Secuenciación y Genómica
Dirección de Investigación en Salud Pública
Instituto Nacional de Salud



La salud
es de todos

Minsalud

¿Qué es Anotación?

WordReference.com |

1. Adición de notas o comentarios a un escrito
2. Apunte o toma de un dato por escrito

>sec1

AAAAATGCCCGCTTCGGATTTCGGATTAGGCTTAGGCTTCGGACTATC
GGATTTTCGGGACCCTTGGACCCTTGGCACTTTCAACGGACTTACAC
GGTTACCGGGGACCATTGGCACTTACGG



>sec1

AAAAATG**CCCGCTTCGGATTTCGGATTAGG**CTTAGGCTTCGGACTATC
GGATTTTCGGGACCCTTGGACCCTTGGCACTTTCAACGGACTTACAC
GGTTACCGGGGACCATTTGGCACTTACGG



¿Qué es Anotación?

La anotación del ADN de un genoma consiste en:

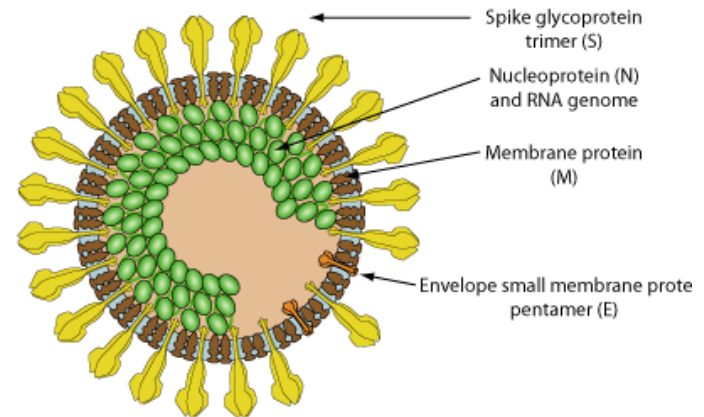
1. *Identificación de genes, regiones codificantes y motivos*
2. *Proceso de identificación y etiquetado de todas las características relevantes en una secuencia del genoma*

Anotación Estructural



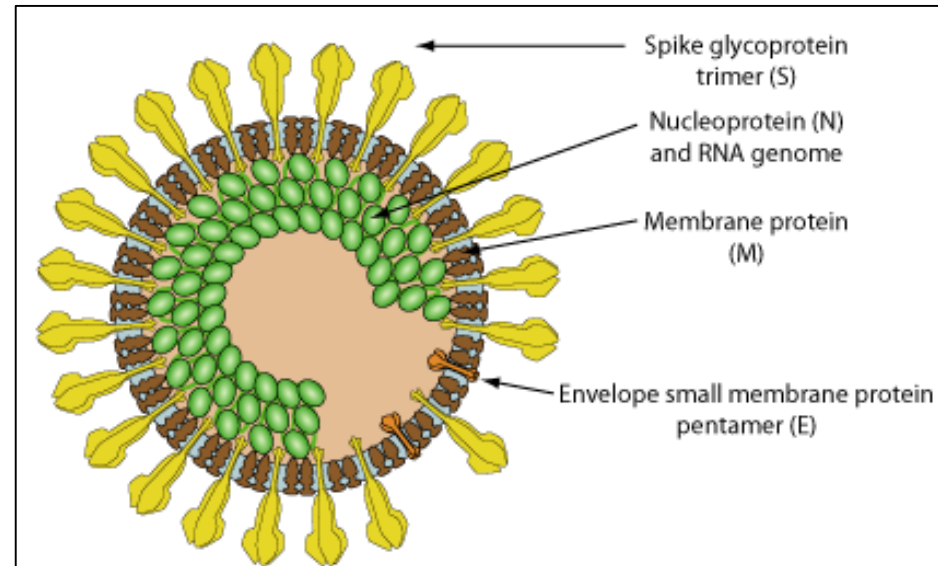
Identificación
Identificación de genes
Identificación de regiones

Anotación Funcional



Características funcionales y
físicas de los productos génicos
Perfil metabólico

COVID19: SARS-CoV-2



Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

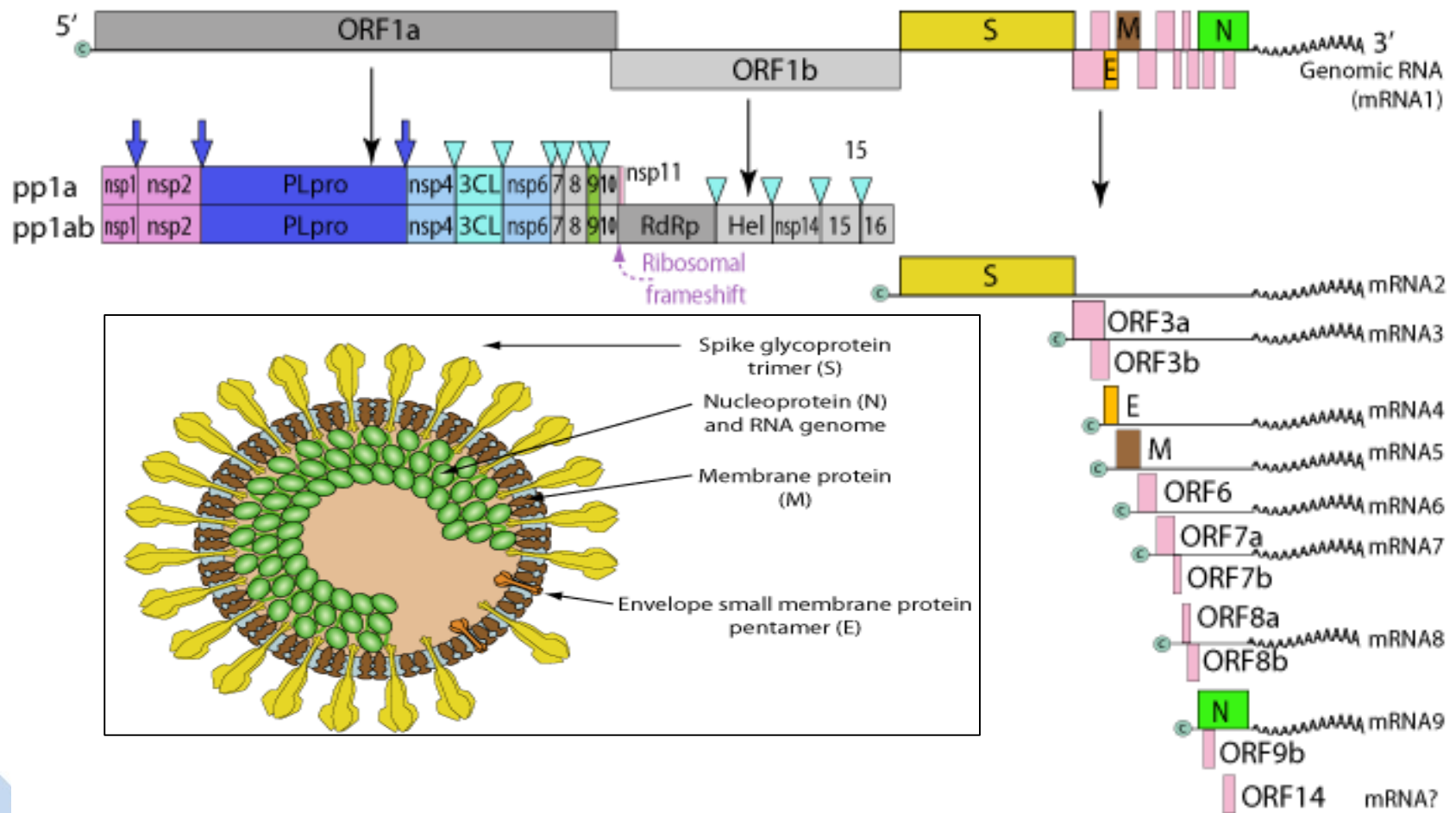
NCBI Reference Sequence: NC_045512.2

[FASTA](#) [Graphics](#)

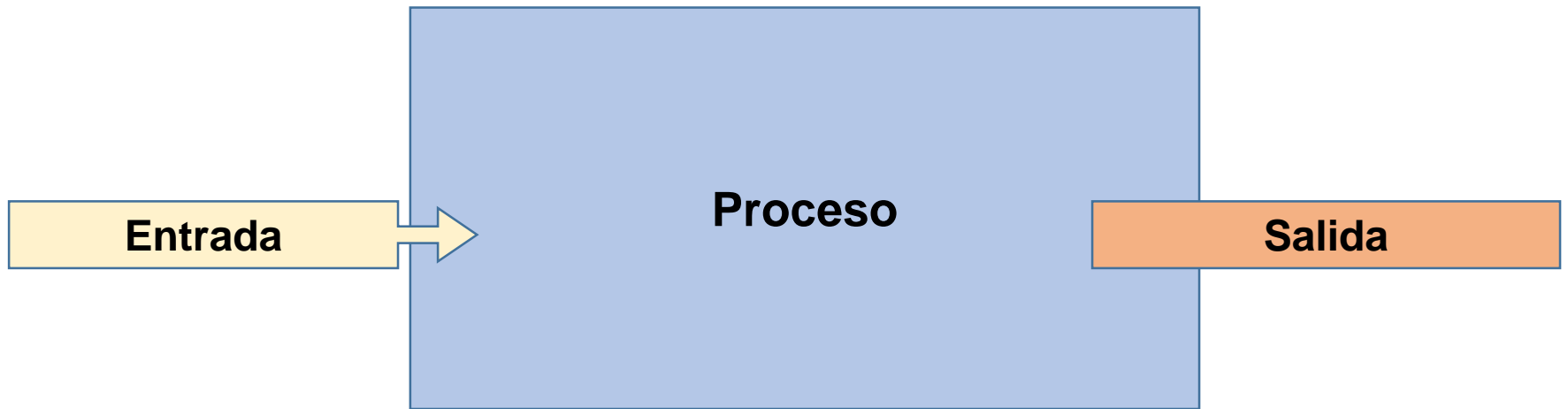
Go to: ☐

LOCUS	NC_045512	29903 bp ss-RNA	linear	VRL 18-JUL-2020
DEFINITION	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.			

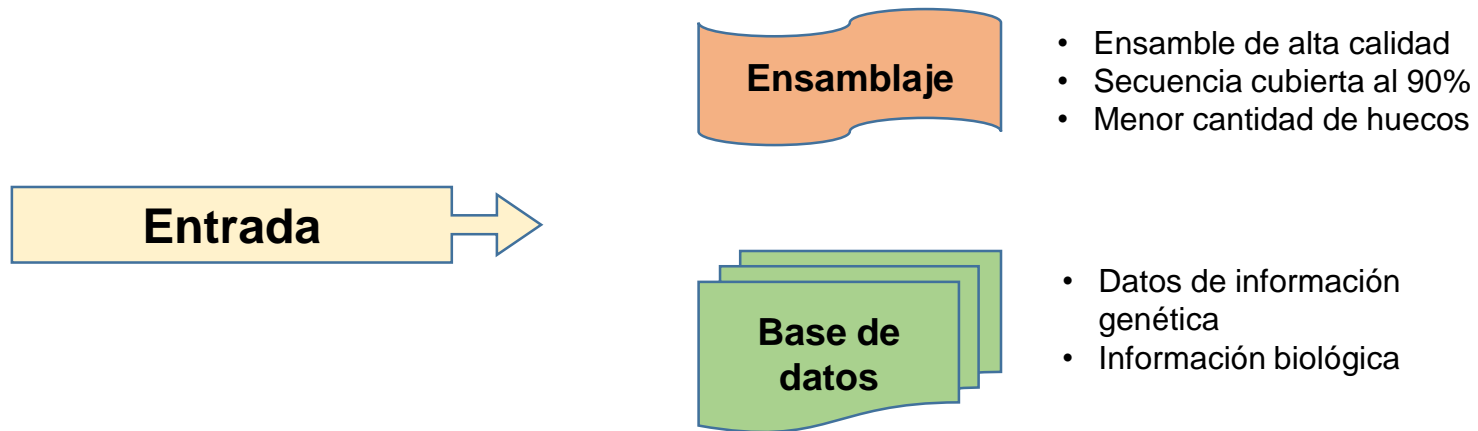
Anotación Funcional



Anotar el genoma



Anotar el genoma



Clasificación de ensamblajes de genomas virales y uso de las secuencias

Functional potential, host prediction, taxonomic classification*, diversity & distribution*	New taxonomic groups	New reference species
Finished genome Complete genome with extensive annotation		
High-quality draft genome Predicted $\geq 90\%$ complete		
Genome fragment(s) Predicted $< 90\%$ complete or no estimated genome size		

1. **Potencial funcional:** análisis del contenido de genes
2. **Predicción de host:** predicción de host *in silico*.
3. **Clasificación taxonómica:** clasificación del contig a grupos establecidos
4. **Diversidad y distribución:** incluye agrupamiento de vOTU y estimación de abundancia relativa
5. **Nuevos grupos taxonómicos:** delimitación de nuevos grupos
6. **Nueva especie de referencia:** se refiere a la propuesta de una nueva entrada en ICTV

Bases de datos



<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/>



<https://talk.ictvonline.org/taxonomy/>



ViPTree : the Viral Proteomic Tree server version

<https://www.genome.jp/viptree/>

1.9

VOGDB



Virus Orthologous Groups

<http://vogdb.org/>



<https://www.genome.jp/kegg/>

Pfam

Anotar el genoma

Proceso

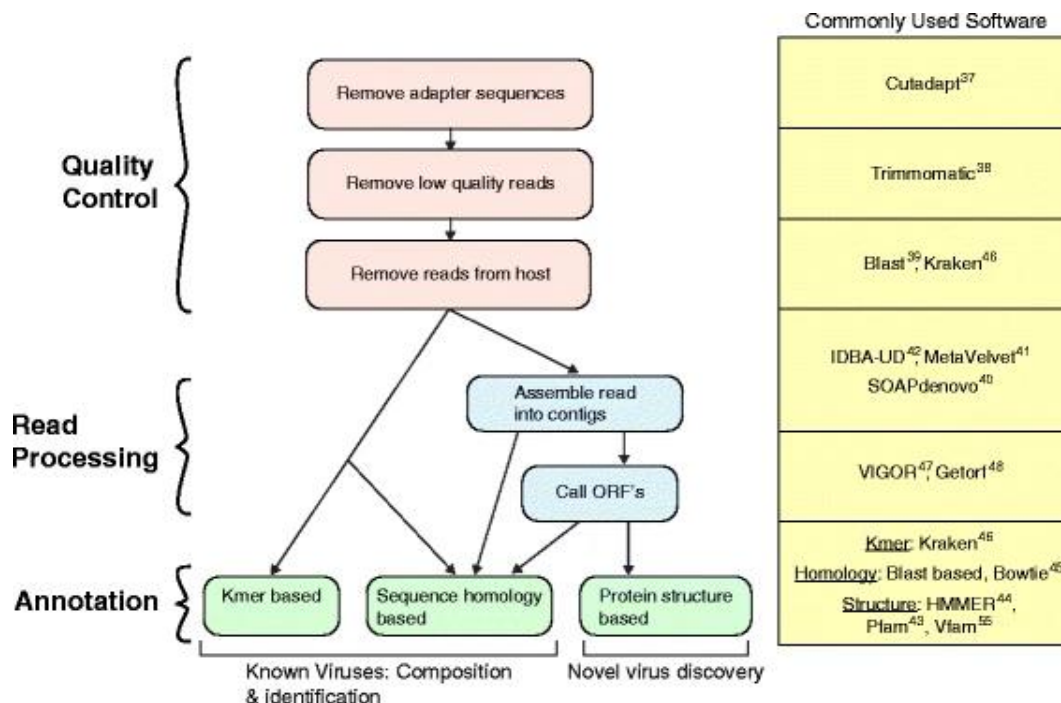
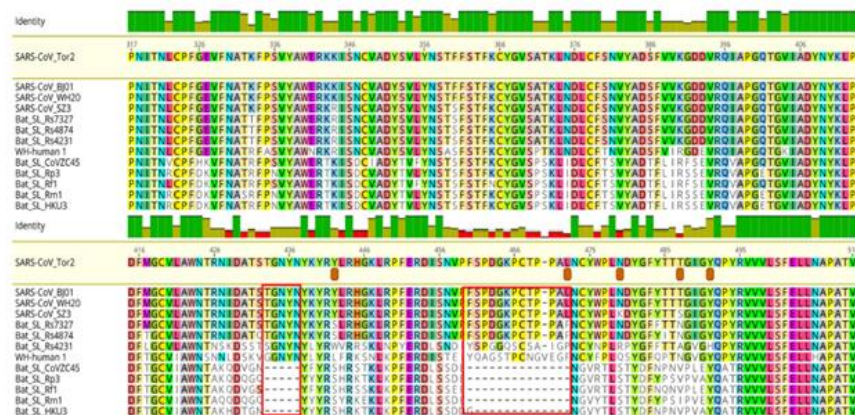
Preparación del genoma: Identificación y descarte de las partes del genoma que no contienen genes.

Fase computacional: Identificación de los elementos del genoma con base en información previa

Fase de anotación: Combinar los elementos mapeados, adjuntarles información biológica y finalmente definir un conjunto óptimo de anotaciones

Validación: Inspecciones manuales, comprobaciones experimentales y medidas de calidad

Programas



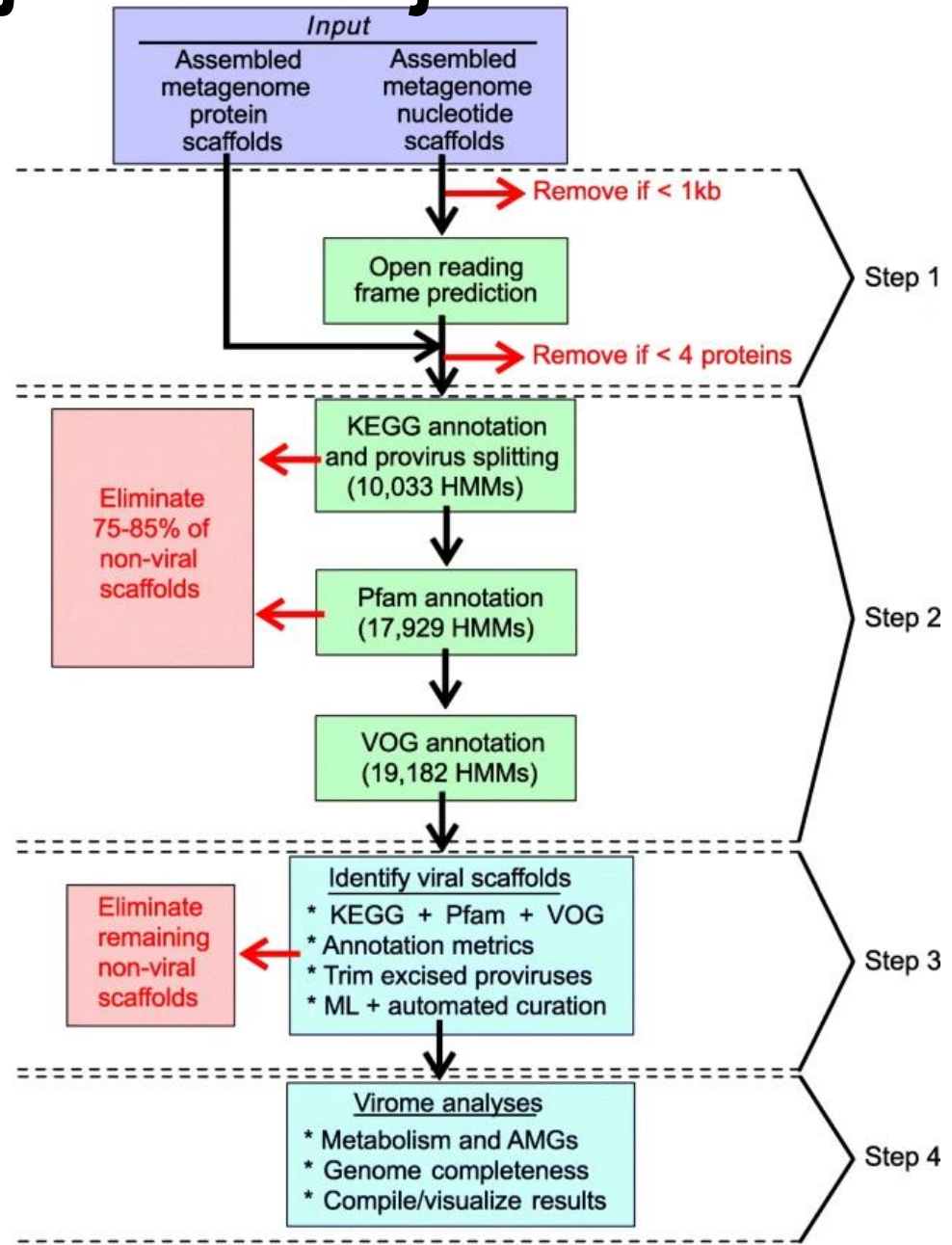
Programas: Flujo de trabajo automático



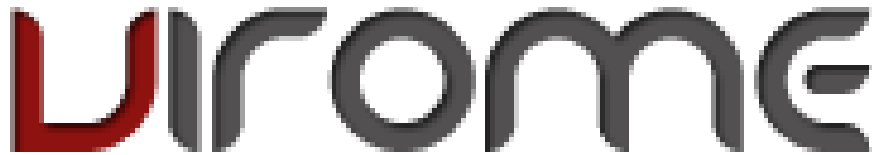
VOGDB



Virus Orthologous Groups



Herramientas online

The logo for Virome, featuring the word "Virome" in a stylized font. The "Vi" is in red and the "rome" is in grey.

<http://virome.dbi.udel.edu/app/#view=Browse;>

**VGAS (viral genome
annotation system)**

<https://cefg.uestc.cn/vgas/>

Anotar el genoma

Salida

- Conjunto diverso de datos biológicos localizados a lo largo del genoma de interés
- Conocimiento de los genes que hay en la secuencia y donde se encuentran

Fuentes de error más frecuentes:

1. No filtrar las regiones en el genoma que no contienen genes
2. Fallas al elegir los programas computacionales
3. Los datos de referencia contienen errores
4. Se utiliza un genoma de referencia con anotaciones erróneas

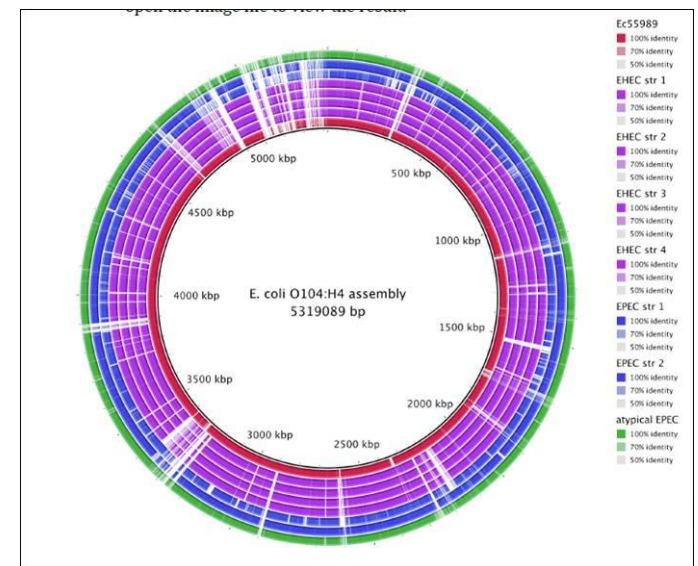
Reporte de anotación

Escribir en el texto que es y que lo compone. Se utilizan datos (proporciones, porcentajes o enteros) para dar solidez a lo reportado

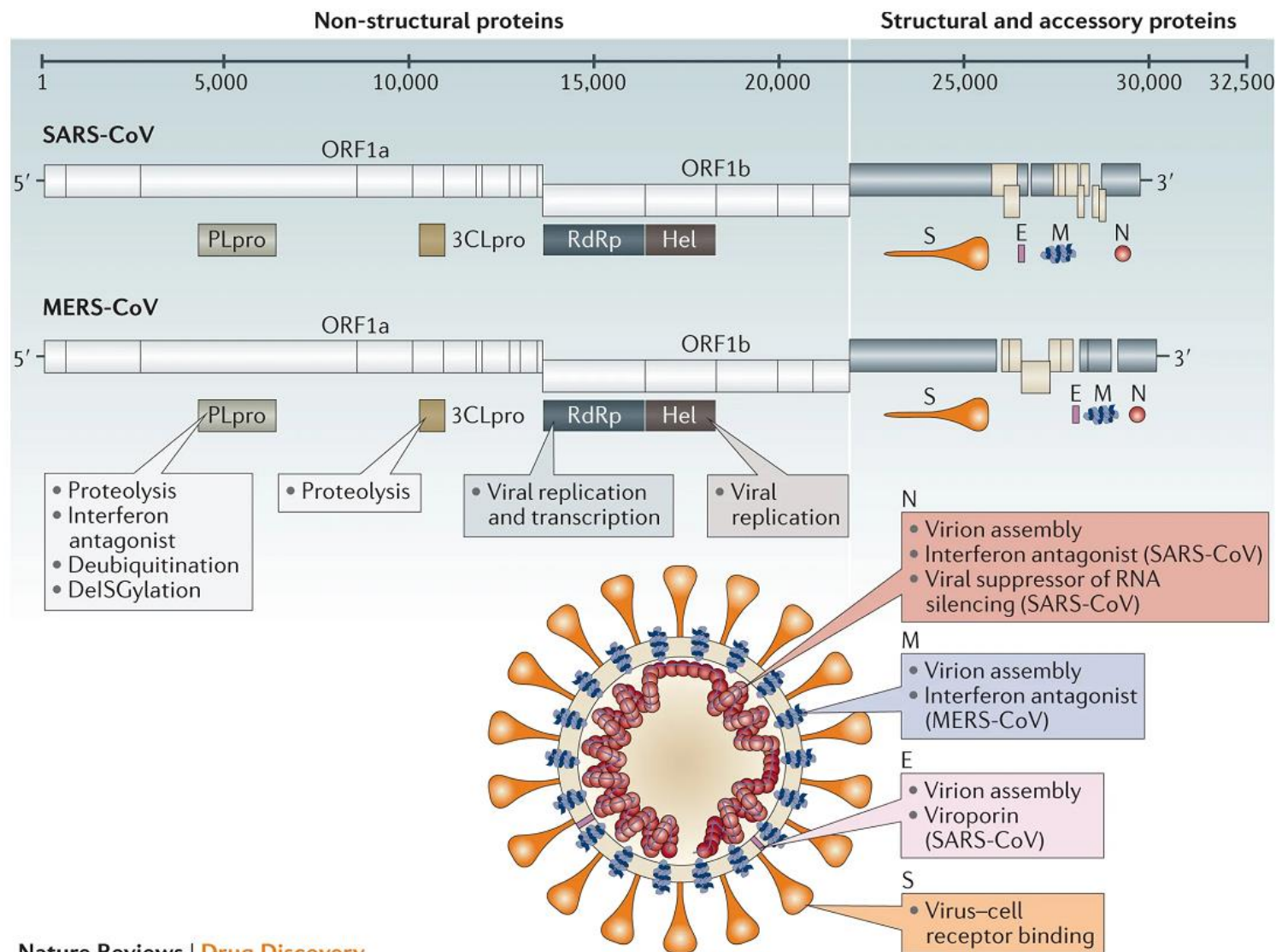
Summary of genomic analysis	
Total nucleotide change (mutations)	1516
Coding region	1247
Non-coding region	269
Total synonymus mutations	503
Total amino acid change or substitution	744
Region	Amino Acid Mutation Number
Polyprotein	412
S	120
M	15
E	11
N	82
ORF3a	48
ORF6	5
ORF7a	22
ORF7b	3
ORF8	16
ORF10	10

Reporte de anotación

Mostrar ilustraciones de procesos bioinformáticos para soportar los cambios a nivel del AND que reportan los investigadores

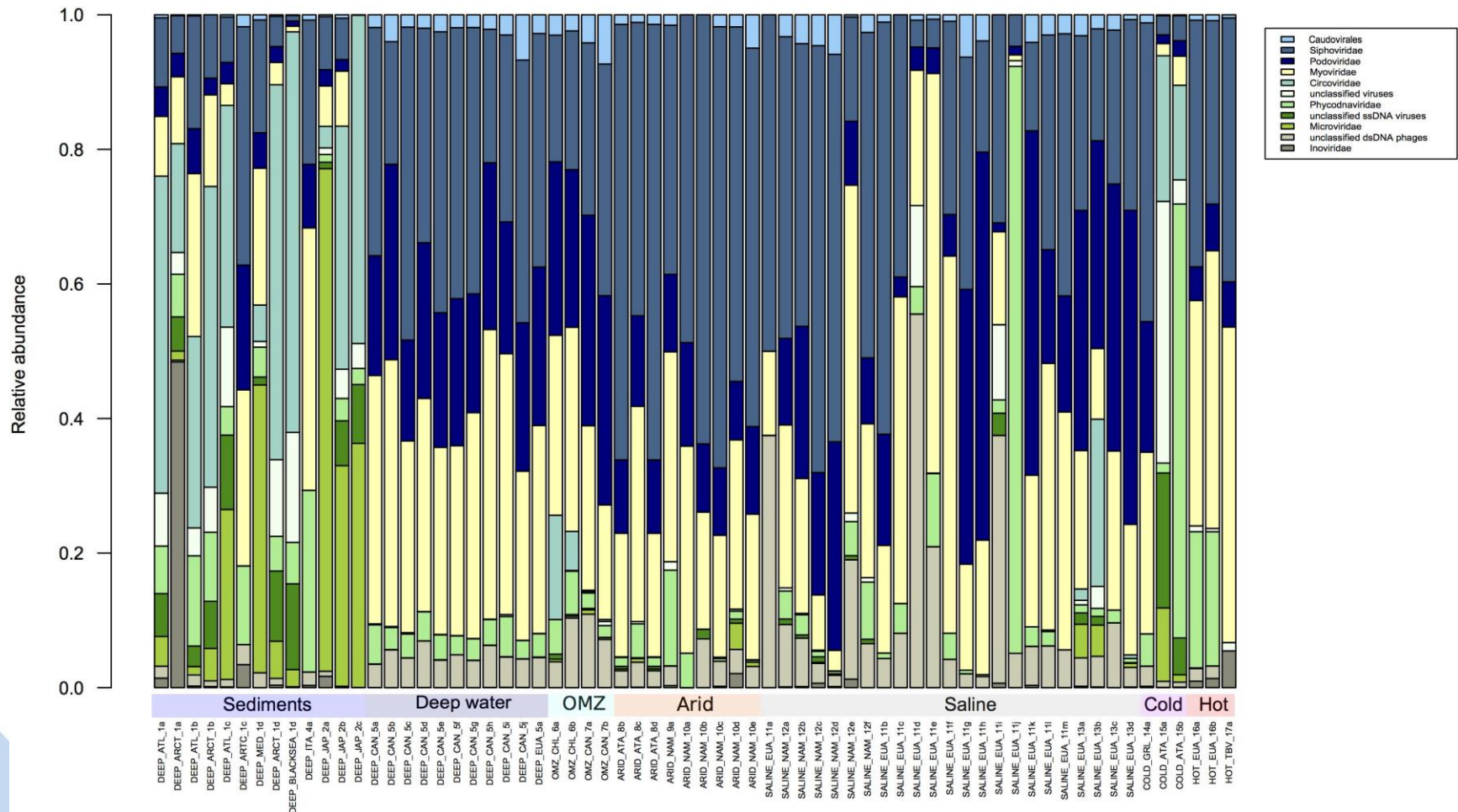


Reporte de anotación

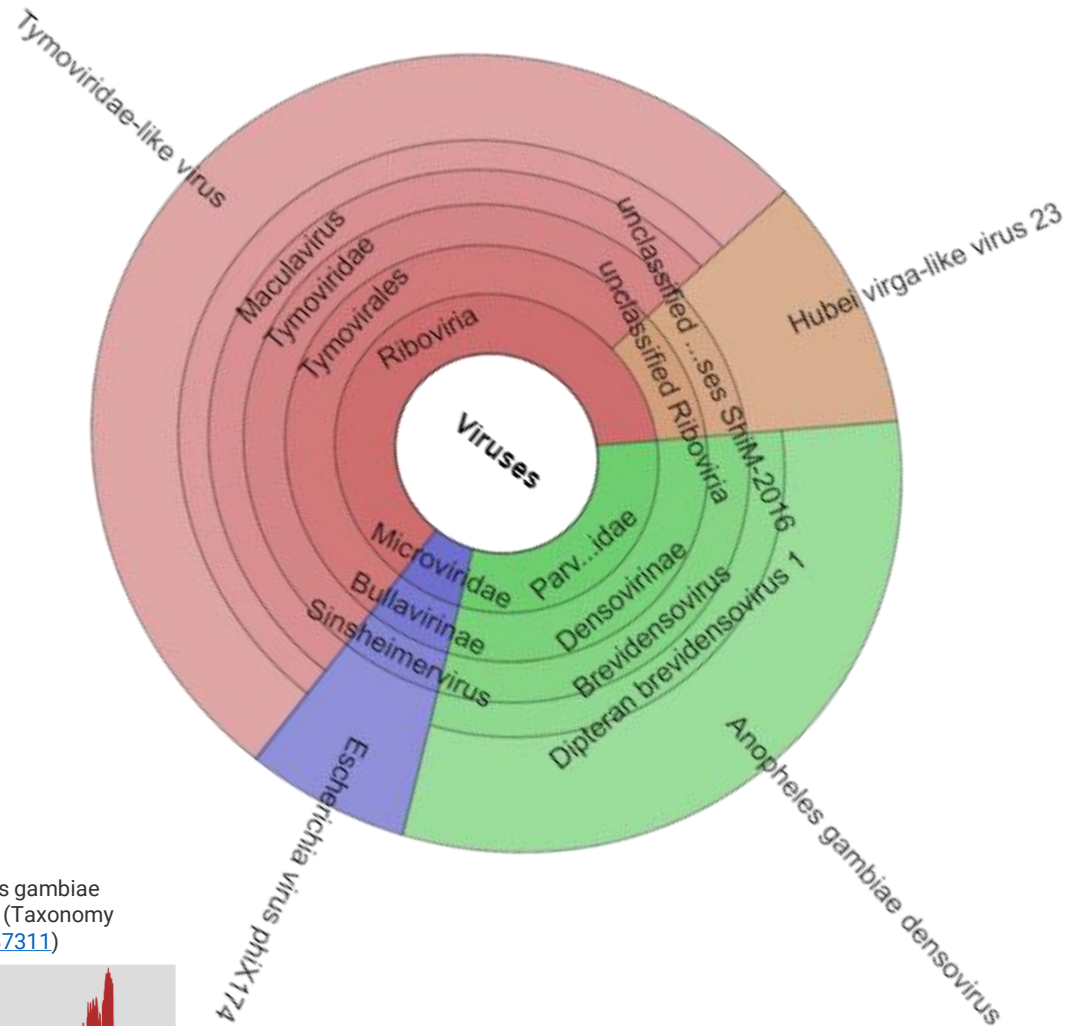


Reporte de anotación

Mostrar ilustraciones de estadísticas para soportar los resultados que reportan los investigadores

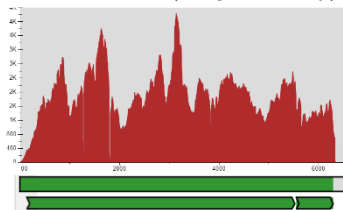


Resultados preliminares de la tesis de doctorado de K. Laiton

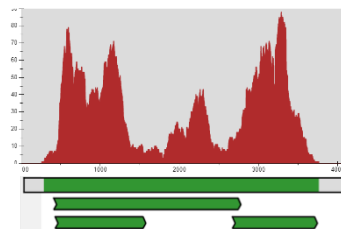


Culex originated Tymoviridae-like virus was identified at high coverage and depth in a mosquito pool of the genus *Culex*.

Culex originated Tymoviridae-like virus (Taxonomy ID: [1236047](#))
Reference Genome
[NC_018703.1](#) (Length: 6471bp)



Anopheles gambiae densovirus (Taxonomy ID: [487311](#))



Rural area
CIST0019 (*Culex* sp)

El reto de anotar y clasificar metagenomas virales

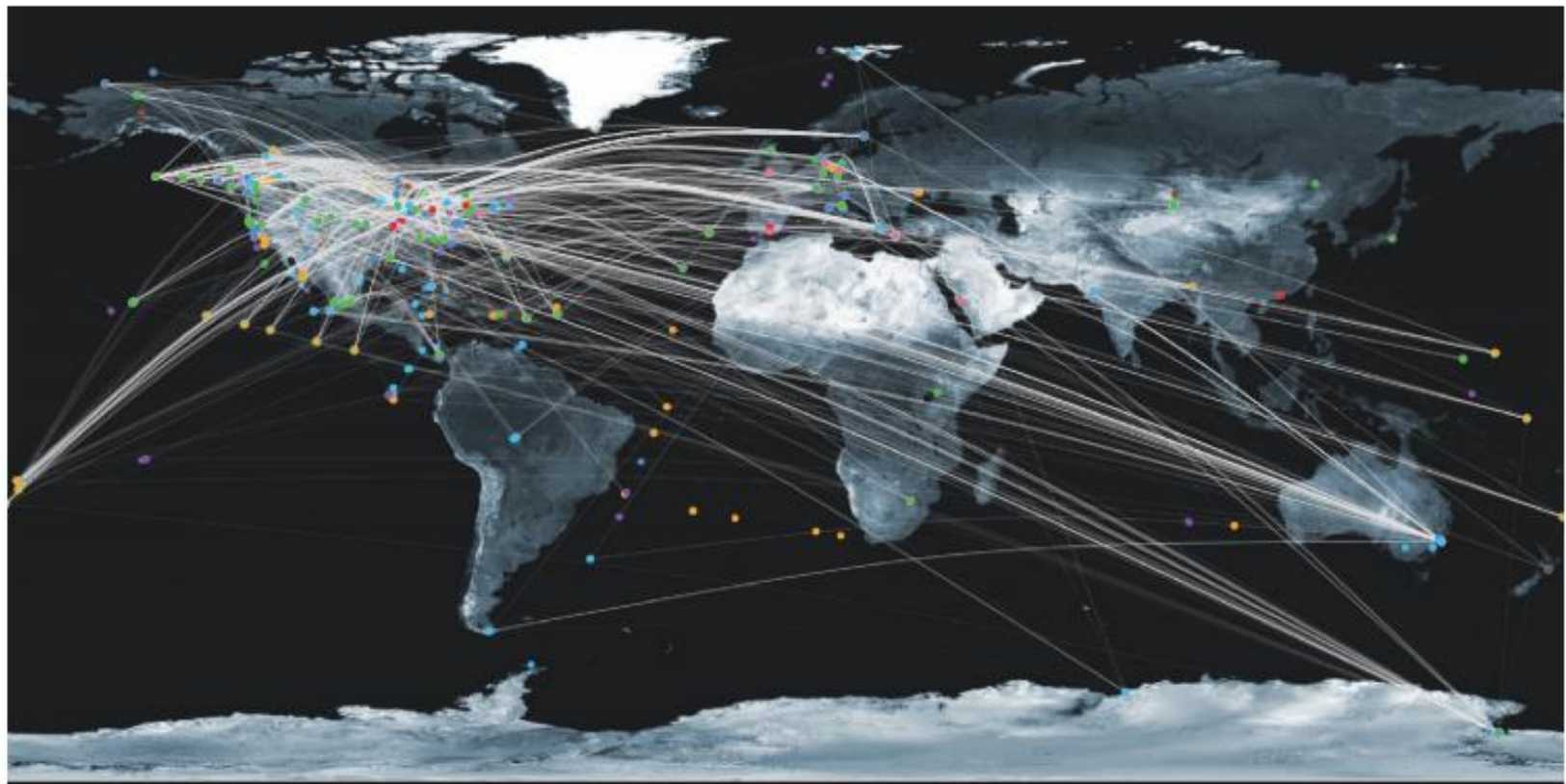
Limitada comprensión de la dinámica de sus poblaciones en los diferentes ecosistemas

Carencia de secuencias genéticas conservadas entre las especies

La mayoría de los genes secuenciados no tienen similitud suficiente como para poder ser comparados con las bases de datos

En los datos metagenómicos de muestras ambientales, se encuentran fragmentos truncados de material genético, elementos de ADN libre y marcos abiertos de lectura incompletos

1. Mejorar el esfuerzo de secuenciación
2. Re-Secuenciar
3. Utilizar todos los tipos de datos:
Lecturas procesadas, contigs, ORF y genes



- Marine
- Freshwater
- Non-marine Saline and Alkaline
- Thermal springs

- Host-associated (other)
- Host-associated (plants)

- Terrestrial (soil)
- Terrestrial (other)
- Area of 10 pixel size

Extended Data Figure 10 | Global connectivity of viral diversity from different habitat types. Geographic location of metagenomic samples containing the same viral groups and singletons represented by a white connecting line across metagenomes from different habitats. Only samples sharing 2 or more viral groups or singletons that are more distant than

10 pixels (area shown as a red square in the figure) are connected. The colours of the samples (circles) indicate the habitat type according with the legend. A freely available equirectangular projection of the world map was used as a background image (<http://visibleearth.nasa.gov/view.php?id=57752>).

-
- Taller