



Curso Teórico-Práctico

# EPIDEMIOLOGÍA GENÓMICA



UNA HERRAMIENTA PARA  
FORTALECER LA VIGILANCIA DE  
AGENTES INFECCIOSOS



---

Día 2: *NGS, epidemiología aplicada, consenso y Calidad*  
Bogotá / Septiembre 24 del 2024

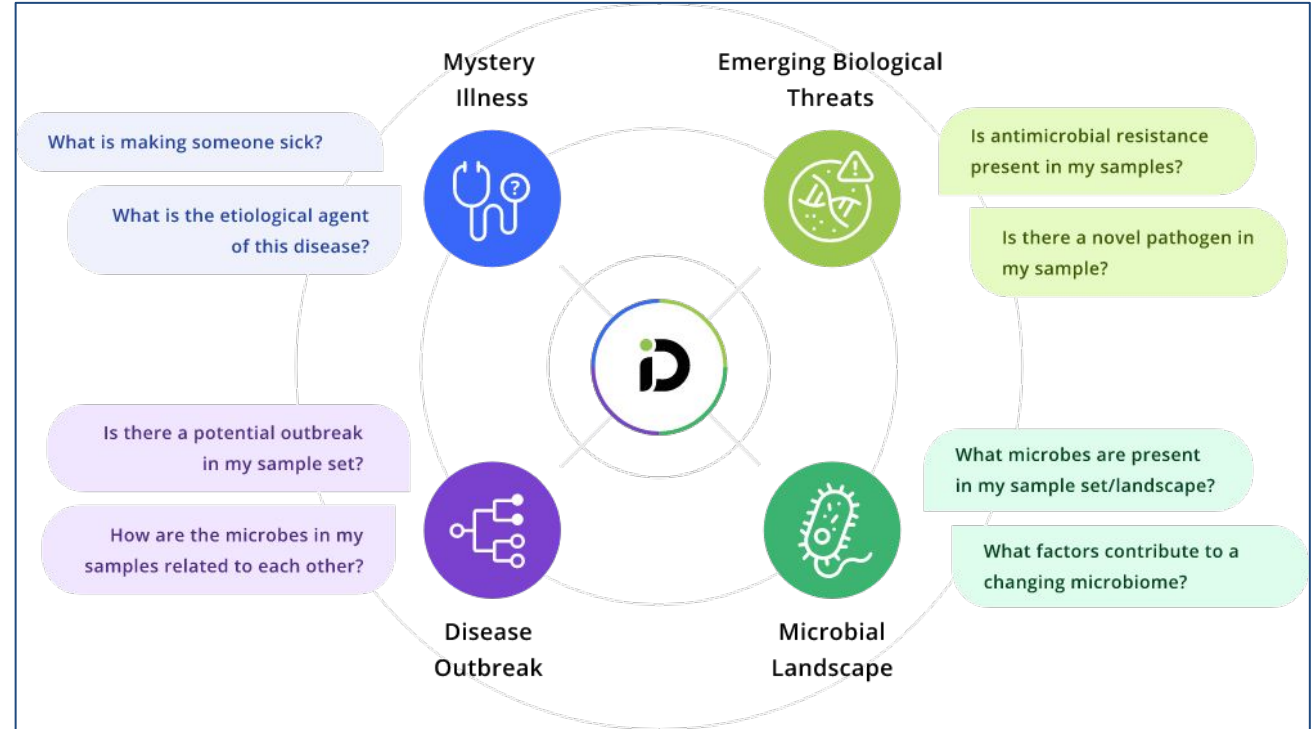
# Consenso de genomas virales en **cz** ID

Héctor Ruiz  
M.Sc.  
Genómica de Microorganismos  
Emergentes

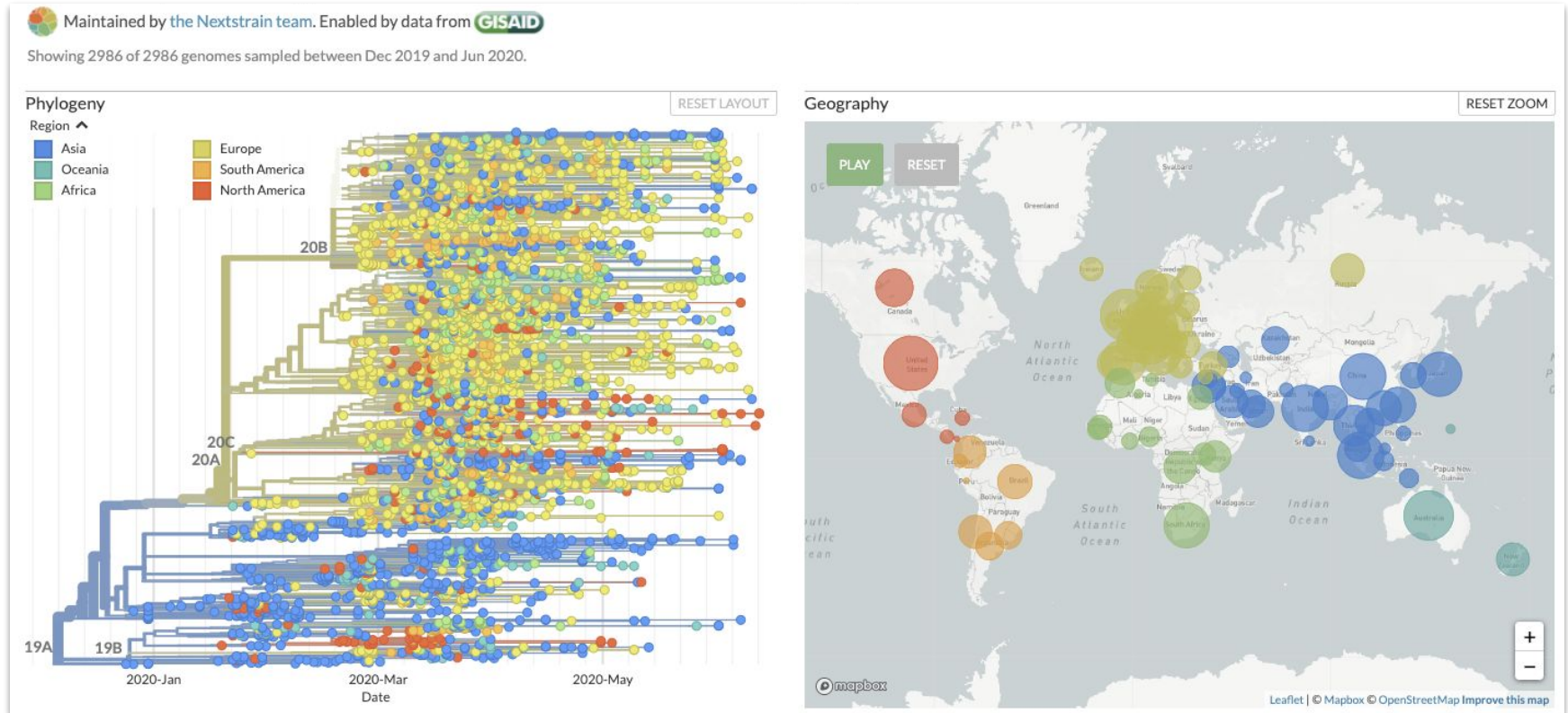
## CZ ID Tiene muchos usos

### CZ ID Pipelines:

- Análisis de metagenómica
- Análisis de genes de resistencia a antimicrobianos
- **Ensamblaje de genomas consenso**



# Las secuencias genómicas son importantes! Lo que secuenciamos son genomas de consenso



**Genomas de consenso:** Representan el nucleótido más frecuentemente observado en cada sitio del genoma en el momento de la recolección de muestras

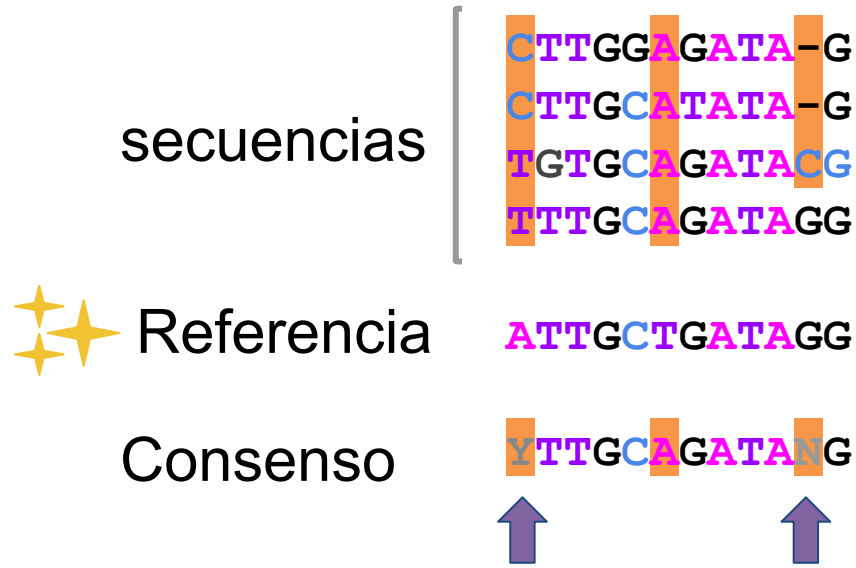
secuencias	<div>ATTGGAGATAC</div> <div>CTTGCATATAG</div> <div>ACTGCA GATAG</div> <div>ATTGCA GATCG</div>
✨ ✨ Referencia	ATTGCTGATAG
Consenso	ATTGCA GATAG

**Genomas de consenso:** Algunas posiciones pueden tener una gran diversidad y no puede aplicarse la regla de la mayoría para decidir una base. Estos sitios pueden definirse con el código de ambigüedad

IUPAC Nucleotide Code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)

IUPAC Nucleotide Code	Degenerate Base
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	Any base

**Genomas de consenso:** Algunas posiciones pueden tener una gran diversidad y no puede aplicarse la regla de la mayoría para decidir una base. Estos sitios pueden definirse con el código de ambigüedad



Ns = "bases no existentes"

# Flujo de trabajo para generar genomas de consenso

Secuencias  
sin procesar  
(fastq)



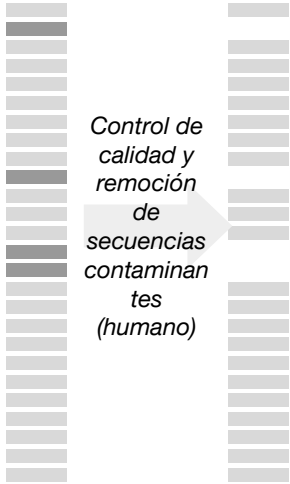
*Control de  
calidad y  
remoción  
de  
secuencias  
contaminan  
tes  
(humano)*





# Flujo de trabajo para generar genomas de consenso

Secuencias  
sin procesar  
(fastq)



*Control de  
calidad y  
remoción  
de  
secuencias  
contaminan  
tes  
(humano)*

*Alineamiento/  
Mapeo*

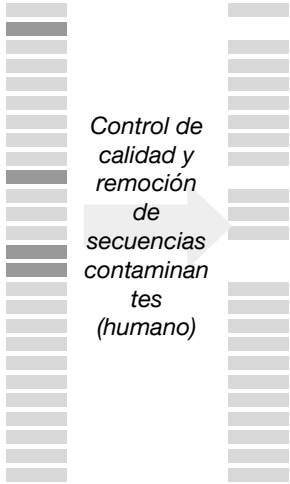
Secuencias mapeadas a la  
referencia



Genoma de referencia

# Flujo de trabajo para generar genomas de consenso

Secuencias  
sin procesar  
(fastq)



Control de  
calidad y  
remoción  
de  
secuencias  
contaminan  
tes  
(humano)

Alineamiento/  
Mapeo

Secuencias mapeadas a la  
referencia



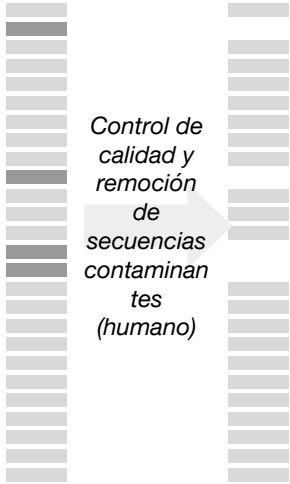
Genoma de referencia

Recortar  
los  
primers



# Flujo de trabajo para generar genomas de consenso

Secuencias  
sin procesar  
(fastq)



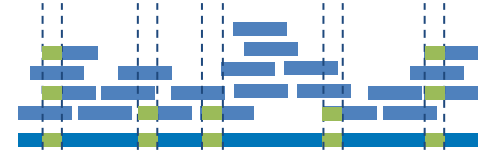
Control de  
calidad y  
remoción  
de  
secuencias  
contaminan  
tes  
(humano)

Alineamiento/  
Mapeo

Secuencias mapeadas a la  
referencia

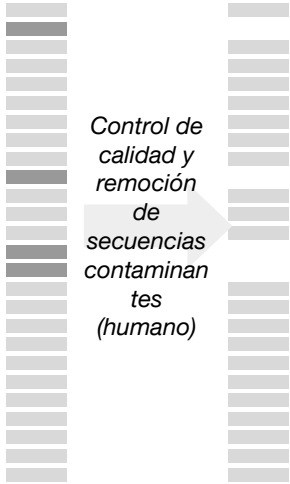


Recortar  
los  
primers



# Flujo de trabajo para generar genomas de consenso

Secuencias  
sin procesar  
(fastq)



Control de  
calidad y  
remoción  
de  
secuencias  
contaminan  
tes  
(humano)

Alineamiento/  
Mapeo

Secuencias mapeadas a la  
referencia



Recortar  
los  
primers



Genoma consenso



Determinar  
el consenso



# Flujo de trabajo para generar genomas de consenso

Secuencias  
sin procesar  
(fastq)

Control de  
calidad y  
remoción  
de  
secuencias  
contaminan  
tes  
(humano)

Alineamiento/  
Mapeo

Secuencias mapeadas a la  
referencia

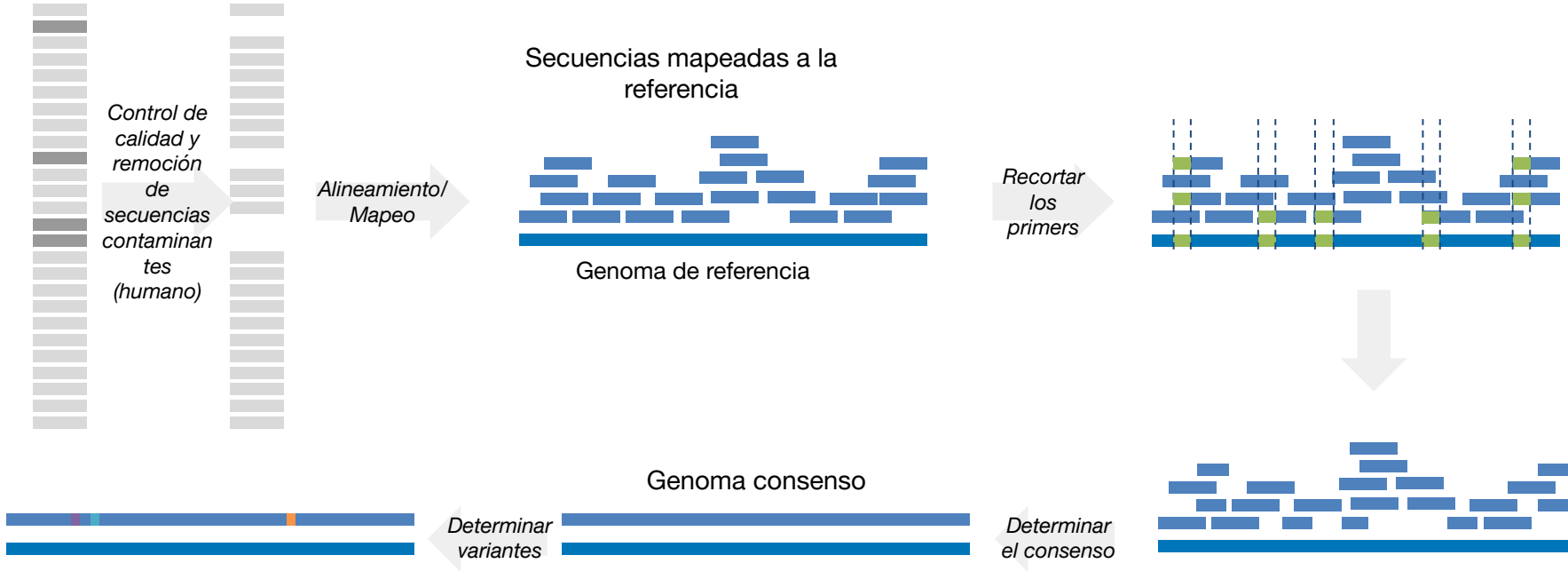
Genoma de referencia

Recortar  
los  
primers

Genoma consenso

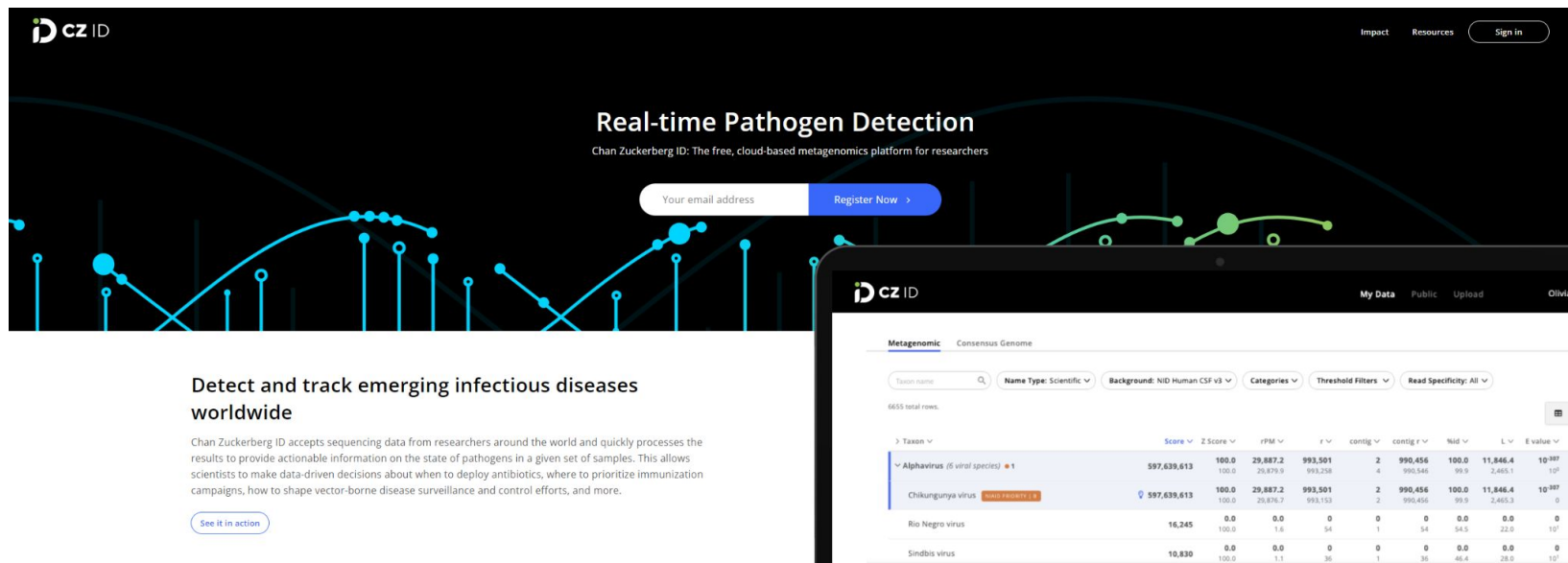
Determinar  
variantes

Determinar  
el consenso



# CZ ID para reconstruir genomas consenso

## Módulo de genomas consenso virales en CZ ID



**CZ ID** Impact Resources [Sign in](#)

### Real-time Pathogen Detection

Chan Zuckerberg ID: The free, cloud-based metagenomics platform for researchers

Your email address [Register Now](#)

#### Detect and track emerging infectious diseases worldwide

Chan Zuckerberg ID accepts sequencing data from researchers around the world and quickly processes the results to provide actionable information on the state of pathogens in a given set of samples. This allows scientists to make data-driven decisions about when to deploy antibiotics, where to prioritize immunization campaigns, how to shape vector-borne disease surveillance and control efforts, and more.

[See it in action](#)

**CZ ID** My Data Public Upload Olivia

**Metagenomic** Consensus Genome

Taxon name  Name Type: Scientific Background: NID Human CSF v3 Categories Threshold Filters Read Specificity: All

6855 total rows.

Taxon	Score	Z Score	rPM	r	contig	contig r	Nid	L	E value
Alphavirus (6 viral species) 1	597,639,613	100.0	29,887.2	993,501	2	990,456	100.0	11,846.4	10 <sup>-987</sup>
Chikungunya virus	597,639,613	100.0	29,887.2	993,501	2	990,456	100.0	11,846.4	10 <sup>-987</sup>
Rio Negro virus	16,245	0.0	0.0	0	0	0	0.0	0.0	0
Sindbis virus	10,830	0.0	0.0	0	0	0	0.0	0.0	0

# Automatización de construcción de genomas virales (metagenómica opcional)

## Select Samples

Rather use our command-line interface? [View CLI Instructions.](#)

1  
Samples

2  
Metadata

3  
Review





## Select Project

Project

Select project

[+ CREATE PROJECT](#)

## Analysis Type

- ☐  **Metagenomics**  
Run your samples through our metagenomics pipeline. Our pipeline supports Illumina and Nanopore technologies.
- ☐  **Antimicrobial Resistance**  
Run your samples through our antimicrobial resistance pipeline. Our pipeline supports metagenomics or whole genome data. It only supports Illumina. You can also run the AMR pipeline from within an existing project by selecting previously uploaded mNGS samples. You can check out the AMR pipeline on Github [here](#).
- ☐  **Viral Consensus Genome**  
Run your samples through our Illumina supported pipeline to get viral consensus genomes using your own reference sequence. Pipeline report does not link to Nextclade.
- ☐  **SARS-CoV-2 Consensus Genome**  
Run your samples through our Illumina or Nanopore supported pipelines to get consensus genomes for SARS-CoV-2. Send consensus genomes to Nextclade.

## Select Files

Your Computer S3 Bases

Select Input Files [MORE INFO](#)

Drag and drop your files here, or [click to use a file browser](#).

Continue

Cancel

## INPUTS:

- Genoma de referencia
- Archivo Fastq de lecturas
- Archivo BED [opcional]

## Tips:

- Usar un genoma de referencia estándar usado por la comunidad, o aquel usado en el diseño de los primers
- Para mantener consistencia en el análisis posteriores como construcción de filogenias debe usar la misma referencia para el ensamblaje

FASTA ▼

## Chikungunya virus, complete genome

NCBI Reference Sequence: NC\_004162.2

[GenBank](#) [Graphics](#)

>NC\_004162.2 Chikungunya virus, complete genome  
ATGGCTGCGTGAGACACACGTAGCCTACCACTTTCTTACTGCTCTACTCTGCAAAGCAAGAGATTAAGAA  
CCCATCATGGATCCTGTGTACGTGGACATAGACGCTGACAGCGCCTTTTTGAAGGCCCTGCAACGTGCGT  
ACCCCATGTTTGAGGTGGAACCTAGGCAGGTACACCGAATGACCATGCTAATGCTAGAGCGTTCTCGCA  
TCTAGCTATAAACTAATAGAGCAGGAAATTGATCCCGACTCAACCATCCTGGATATTGGTAGTGCGCCA  
GCAAGGAGGATGATGTGCGACAGGAAGTACCACTGCGTTTGCCCGATGCGCAGTGCGAAGATCCCGAGA  
GACTCGCCAATTATGCGAGAAAGCTAGCATCTGCCGACAGGAAAAGTCTGGACAGAAACATCTCTGGA  
GATCGGGGACTTACAAGCAGTAATGGCCGTGCCAGACACGGAGACGCCAACATTCTGCTTACACACAGAT  
GTATCATGTAGACAGAGAGCAGACGTCGCGATATACCAAGACGTCTATGCTGTACACGACCCACGTCGC  
TATACCACAGGCGATTAAAGGGGTCGATTGGCGTACTGGGTAGGGTTTGACACAACCCCGTTTCATGTA  
CAATGCCATGGCGGGTGCCTACCCCTCATACTCGACAAATTGGGCAGATGAGCAGGTACTGAAGGCTAAG  
AACATAGGATTATGTTCAACAGACCTGACGGAAGGTAGACGAGGCAAATTGCTATTATGAGAGGAAAAA

Send to: ▼

- ☒ Complete Record  
☐ Coding Sequences  
☐ Gene Features

### Choose Destination

- ☒ File ☐ Clipboard  
☐ Collections ☐ Analysis Tool

Download 1 item.

Format

FASTA ▼

Show GI ☐

Create File



- Usado para quitar primers y maximizar la precisión de la detección de SNPs
- Solo usado para secuenciación de amplicones, o con enriquecimiento (e.g. MSSPE)
- A la posición inicial del primer respecto a la referencia se le resta 1 (índice empieza en cero)
- Crear un archivo separado por tabulación y cambiar la extensión a .BED

### Requerido

Reference accession #	Start position	End position	Primer name	Orientation
MN908947.3	555	573	Primer_CoV36	—
MN908947.3	943	965	Primer_CoV37	—
MN908947.3	1376	1397	Primer_CoV35	—
MN908947.3	1708	1731	Primer_CoV38	—
MN908947.3	2161	2180	Primer_CoV34	—
MN908947.3	2491	2512	Primer_CoV39	—
MN908947.3	2872	2890	Primer_CoV33	—
MN908947.3	3306	3330	Primer_CoV40	—
MN908947.3	3758	3777	Primer_CoV32	—

# Agregar metadatos

## Upload Metadata

This metadata will provide context around your samples and results in IDseq.

1

Samples

2

Metadata

3

Review

**Required fields:** We require the following metadata to determine how to process your data and display the results: Host Organism, Sample Type, Water Control, Nucleotide Type, Collection Date, Collection Location. Please be as accurate as possible! [View Full Metadata Dictionary](#).

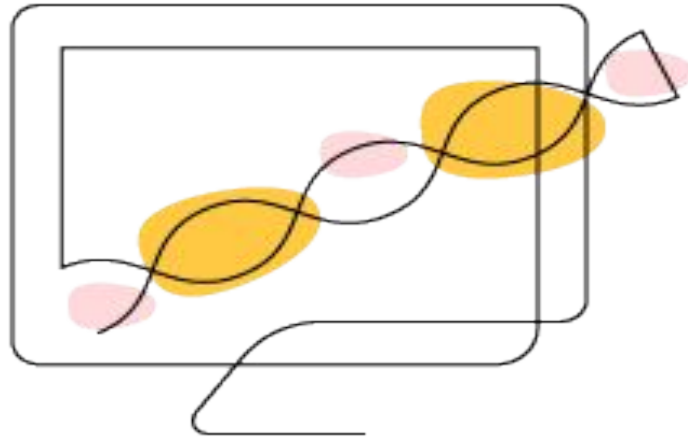
**Available organisms for host subtraction:** Human, Mosquito, Tick, Mouse, Cat, Pig, C.elegans, Carp, Chicken, Bee, Salpingoeca rosetta, Bat, Rat, Field Vole, Bank Vole, Rabbit, Water Buffalo, Horse, Taurine Cattle, Turkey, Barred Hamlet, Orange Clownfish, Tiger Tail Seahorse, Torafugu, Avian, White Shrimp.

Manual Input

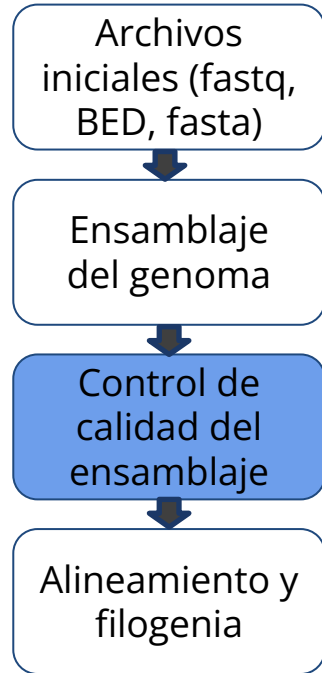
CSV Upload

Sample Name	Host Organism	Sample Type	Water Control	Nucleotide Type	Collection Date	Collection Location	
upload_file	<input type="text"/>	<input type="text"/>	<input type="checkbox"/> No	<input type="text"/>	<input type="text" value="YYYY-MM-DD"/>	<input type="text" value="Enter a city, region or country"/>	<input type="button" value="⊕"/>

El pipeline se ejecuta automaticamente en la nube



# Flujo de trabajo para generar genomas de consenso



¿Puedo utilizar el genoma de consenso ensamblado para análisis posteriores?

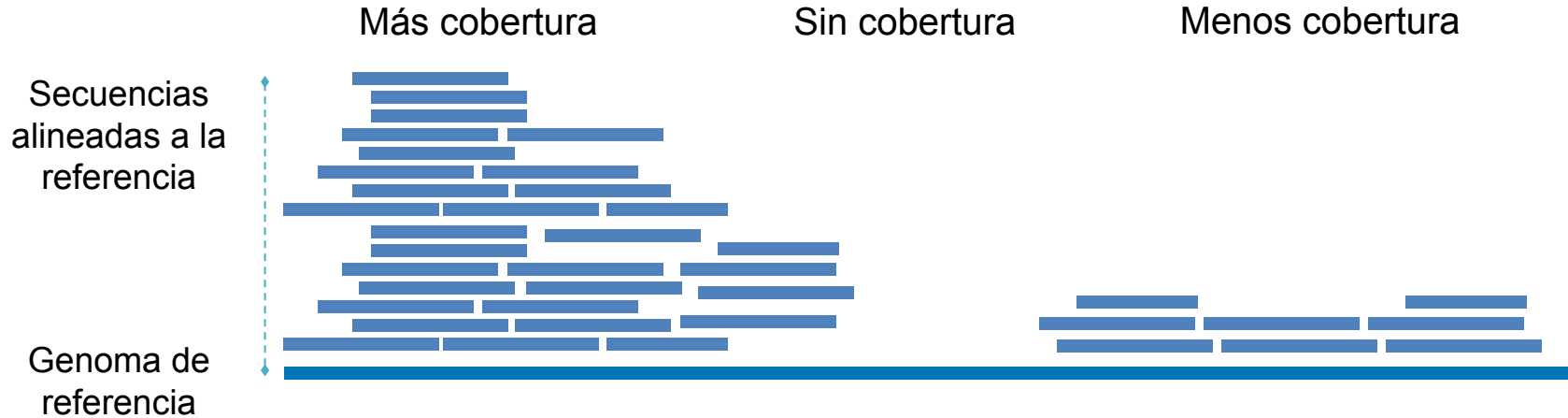
# Control de calidad del genoma consenso: Parametros

- **Secuencias mapeadas** - Número de secuencias alineadas a la referencia.
- **Contenido de GC** - Porcentaje de bases en el consenso que son guanina (G) o citosina (C) (debe ser cercano al %GC de la secuencia de referencia).
- **SNPs (polimorfismos de nucleótido único)** - Representan variaciones unicas de nucleótidos único entre la referencia y el genoma consenso.
- **% Identidad (id)** - Porcentaje de nucleótidos en el genoma consenso que son idénticos a los de la accesión de referencia.

# Control de calidad del genoma consenso: Parametros

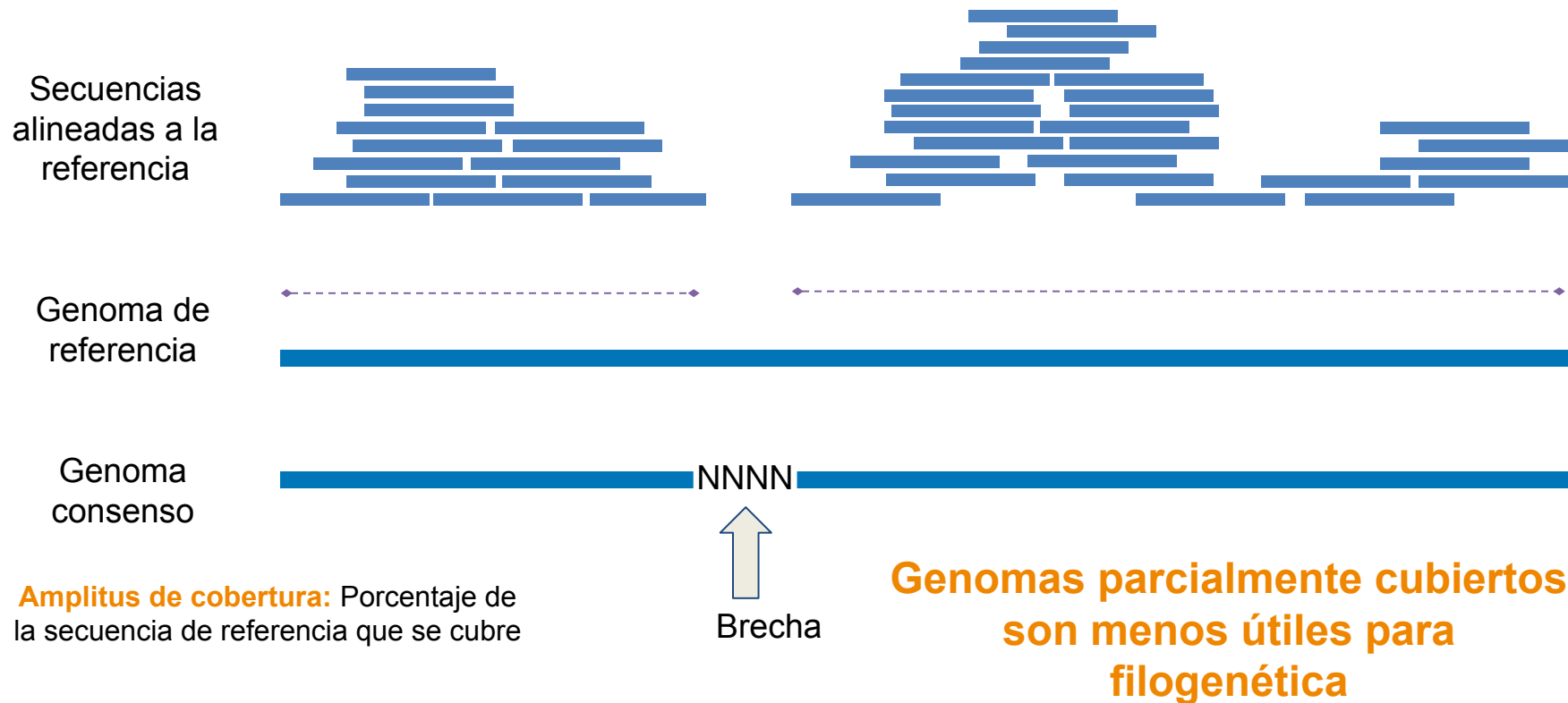
- **Bases Informativas** - Número de nucleótidos que son A,T,C, o G. - En CZID el consenso solo se determina si tiene 10 o más secuencias como soporte
- **% Genoma Llamado** - Porcentaje del genoma que cumple con los umbrales para llamar bases de consenso. Cuanto más se acerque este número al 100%, mejor.
- **Bases faltantes** - Número de Ns en el genoma de consenso.
- **Bases Ambiguas** - Número de bases degeneradas (no-C, T, G, o A) en el genoma consenso. En CZID el genoma de consenso sólo contiene nucleótidos que son detectados al menos con una frecuencia del 75%

# ¿Hay suficiente cobertura?



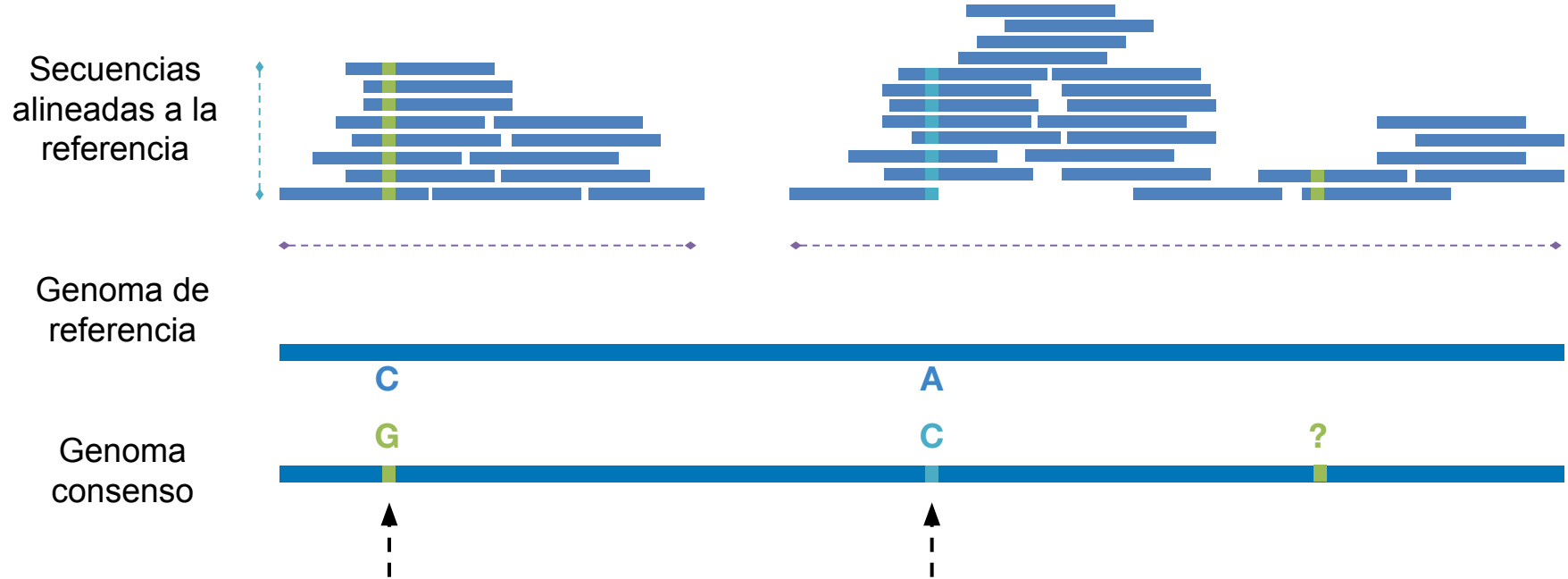
**Profundidad de la cobertura:** # de veces que un nucleótido es secuenciado

# ¿Se recuperó la mayoría del genoma (amplitud de cobertura)?





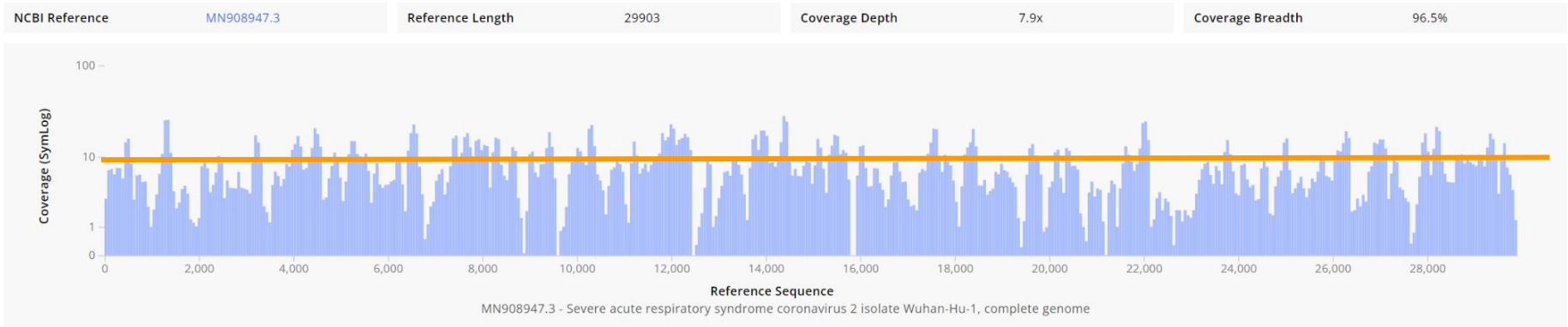
# Evaluación de la profundidad para identificar SNPs



**SNP:** Polimorfismos de nucleótido único, variaciones en un solo par de bases de una secuencia de ADN. El número de SNP aceptables depende del patógeno

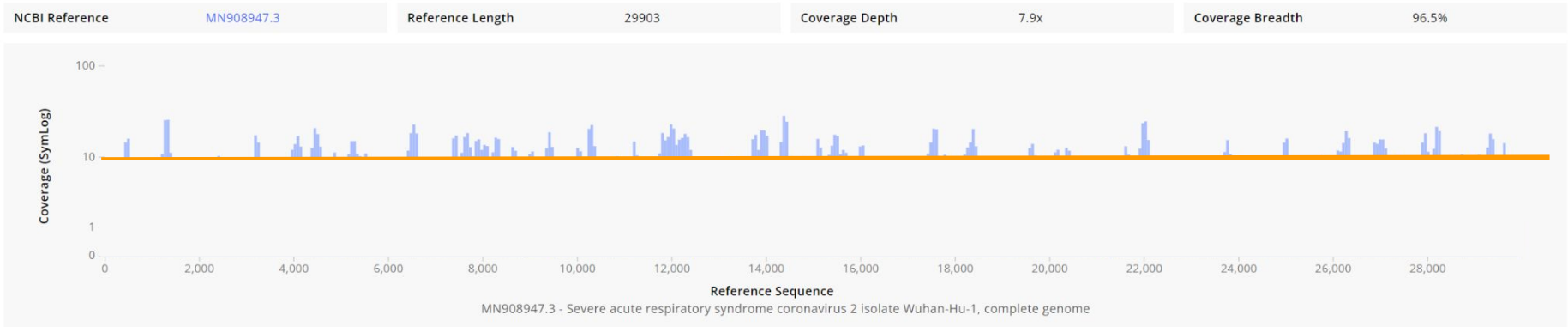
## Importancia de buena profundidad

- Las regiones con una cobertura inferior a 10 secuencias se denominarán N en un genoma de consenso.
- El genoma de consenso que se muestra a continuación sería de baja calidad (la mayoría de las llamadas de base serían Ns)



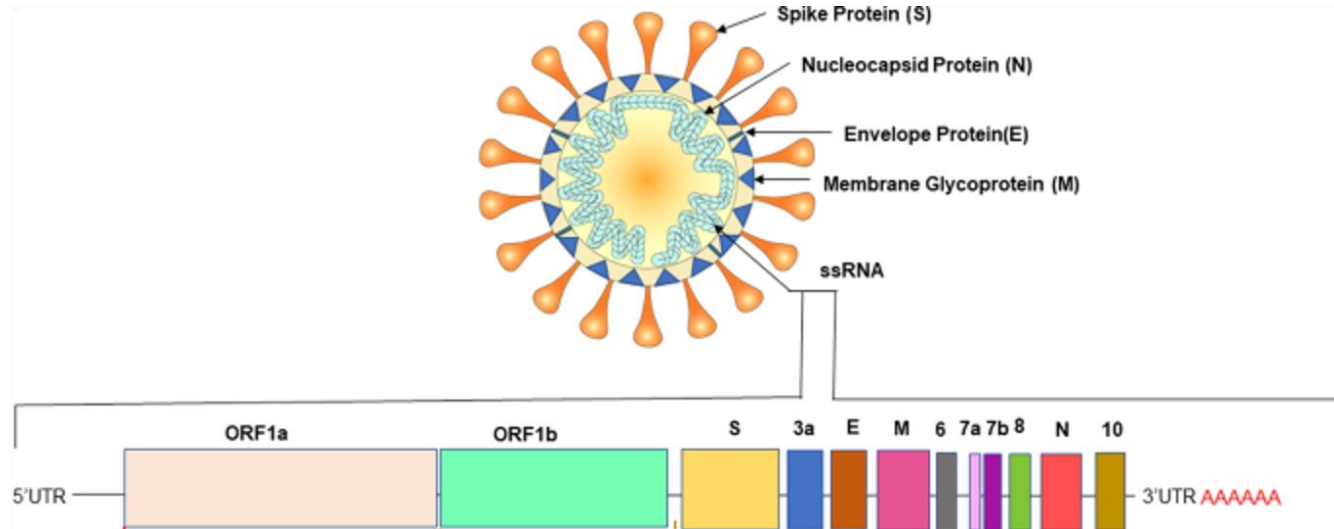
## Importancia de buena profundidad

- Las regiones con una cobertura inferior a 10 secuencias se denominarán N en un genoma de consenso.
- El genoma de consenso que se muestra a continuación sería de baja calidad (la mayoría de las llamadas de base serían Ns)



# ¿Con que puedo continuar?

- El genoma incluye las regiones codificantes previstas?
  - Encontrar marcos de lectura abiertos (ORF) y compararlos con genomas conocidos



# ¿Con que puedo continuar?

- ¿Qué serotipo o variante representa?
- Utilice las herramientas de tipificación disponibles para determinar el serotipo o linaje del virus
- [Nextclade](#) - SARS-CoV-2, Influenza (Gripe), Mpox (Virus del Mono), Virus Sincitial Respiratorio (VSR), Dengue
- [GenomeDetective](#)- Arbovirus, virus de la hepatitis, enterovirus, virus del sarampión
- Análisis filogenético

# Gracias

Contacto:

[hruiz@ins.gov.co](mailto:hruiz@ins.gov.co)

[hector.genomica@gmail.com](mailto:hector.genomica@gmail.com)

Algunas de las diapositivas usadas en esta presentación hacen parte del material disponible en Help Center de CZID

<https://chanzuckerberg.zendesk.com/hc/en-us>