BCVI
Biología Computacional y Virómica Integrativa

Centro de Investigación en Dinámica Celular

# Ensamblajes de Genomas Virales

## Modulo_2 - Ejemplo Sars-CoV-2

Paola Rojas-Estevez

M.Sc Biología Computacional

Grupo Genómica de Microorganismos Emergentes
Instituto Nacional de Salud de Colombia

# Links a utilizar

https://github.com/BCVI/2do-Workshop-Genomica-Viral

https://colab.research.google.com/


GitHub

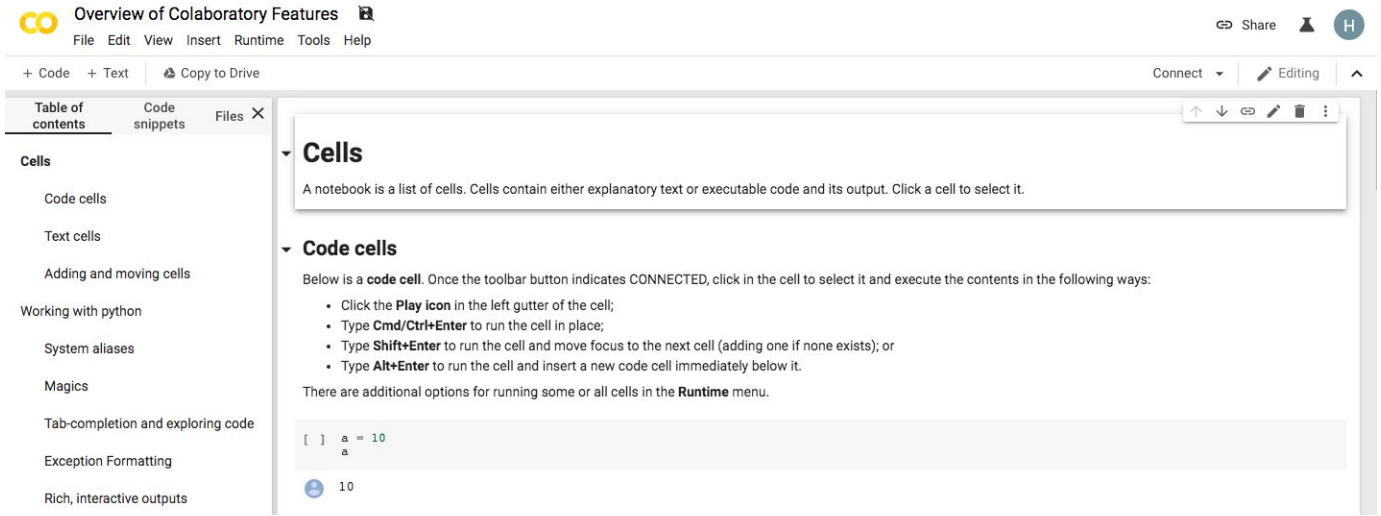
# Repositorio GitHub

## Carpeta Digital de Código

https://github.com/BCVI/2do-Workshop-Genomica-Viral

# Google Colab

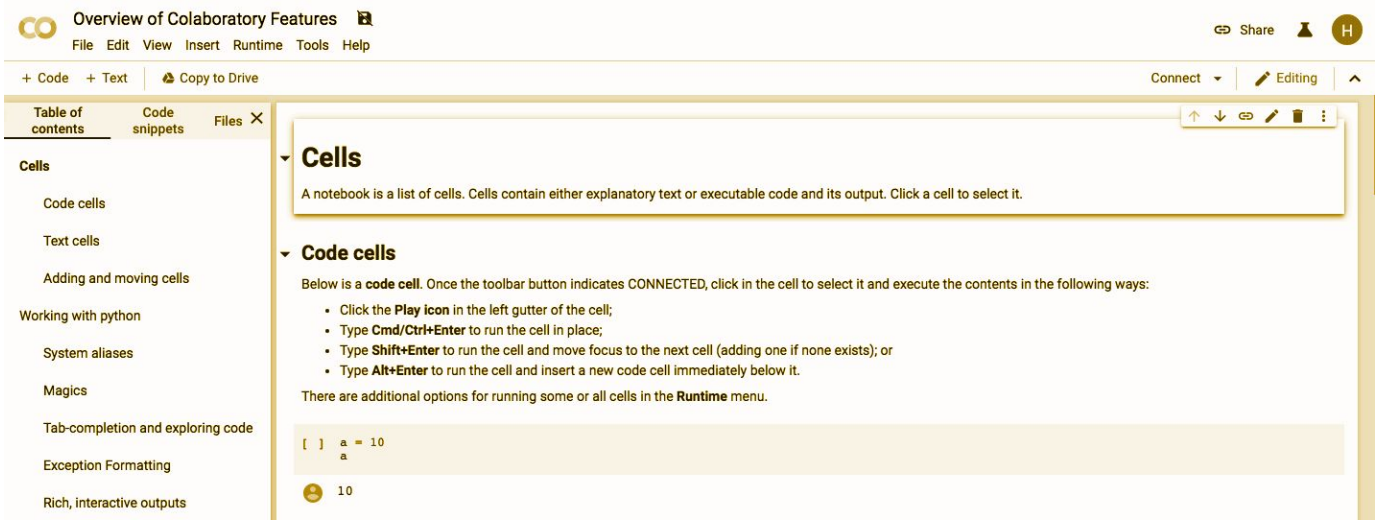
## Entorno de Jupiter notebook personalizado por google




1. Escribir Notas
2. Ejecutar líneas de código con Python

https://colab.research.google.com/

# Google Colab

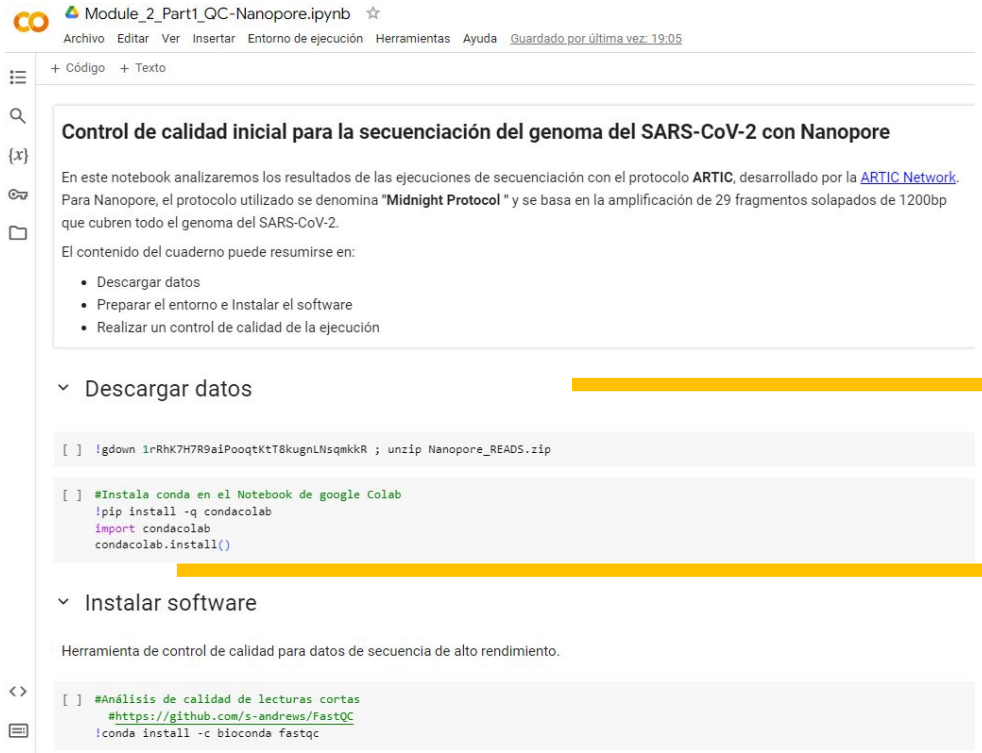



**"Colaboratory"** → permite programar y ejecutar Python/bash en tu navegador

- No requiere configuración
- Acceso a GPUs sin coste adicional
- Permite compartir contenido fácilmente

https://colab.research.google.com/

# Notebooks Colab




Actividades paso a paso

Generalidades de los programas seleccionados

Texto apoyo a línea comando

Comandos por ejecutar paso a paso

https://colab.research.google.com/

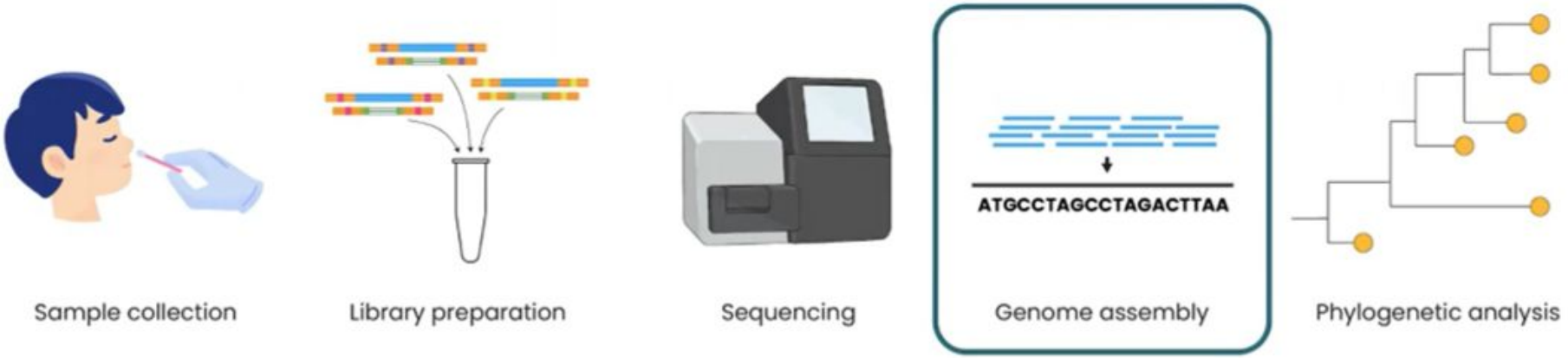
BCVI
Biología Computacional y Virómica Integrativa

Centro de Investigación en Dinámica Celular

# Ensamblajes de Genomas Virales

## Modulo 2 - Calidad / Ensamblaje

# Workflow epidemiología genómica



# Secuenciación Genómica

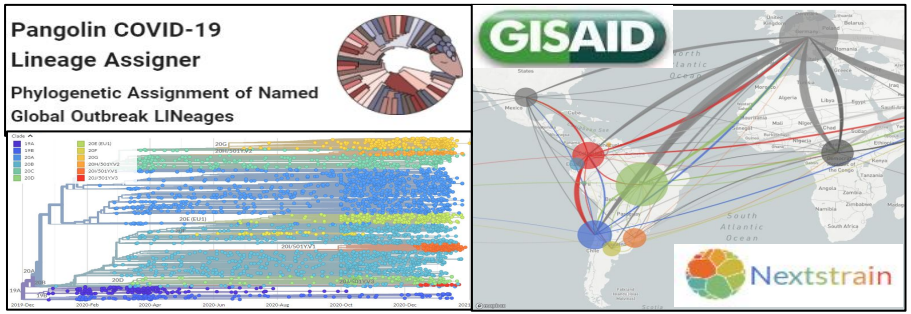
illumina

Nanopore

## Vigilancia Genómica

DENV, CHIKV, ZIKV, Rabies, Measles, POXVIRUS, H1N1, VIH

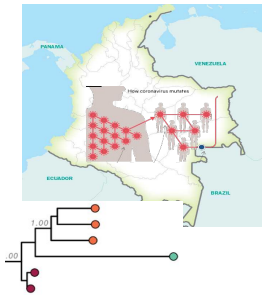
Metagenómica de virus emergentes y virodiversidad




## SARS-CoV-2 emergence

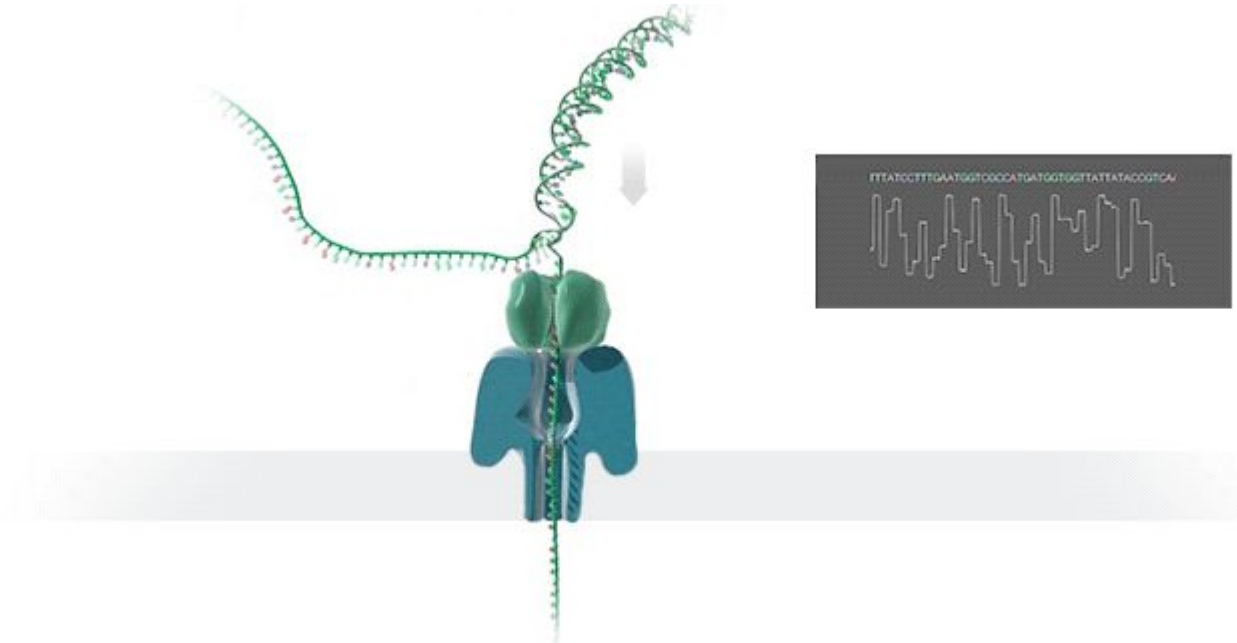
NGS Protocols

ARTICnetwork



# Secuenciación Nanopore

## Archivos fast5/pods5: datos de señal eléctrica

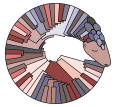


https://nanoporetech.com/how-it-works

# Paso a Paso Análisis

1. Llamado de bases guppy_basecaller
2. Control de Calidad artic guppyplex
3. Ensamblaje artic minion


4. Estimación del Linaje Pangoline / Nextclade



5. Revisión de Mutaciones NEXTSTRAIN

# Control de Calidad

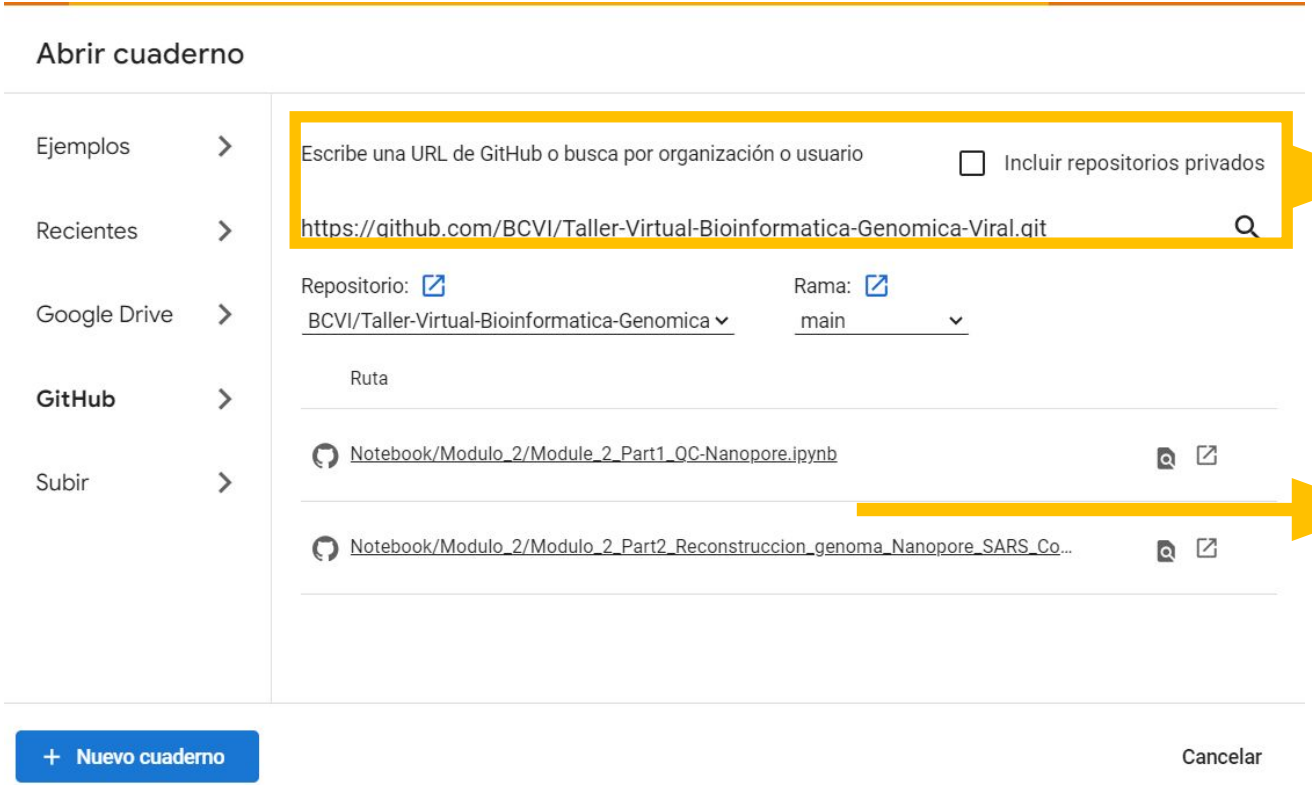
## Formato fastq

Todos los secuenciadores producen datos en un formato llamado **fastq**. Todas las secuencias con un fastq están representadas por 4 líneas:

```
@SEQ_ID                  <---- Nombre Secuencia
AGCGTGTACTGTGCATGTCGATG   <---- Bases de la secuencia
+                        <---- Linea separador
%%).1***-+*''))**55CCFF   <---- Puntaje de Calidad ASCII
```

La calidad de las secuencias se representa como un carácter del código ASCII. Consulte aquí para obtener una explicación. Los valores numéricos corresponden a los valores de calidad phred

# Cargar los notebooks




Copiar URL del curso

Notebook por módulo

https://colab.research.google.com/

BCVI
Biología Computacional y Virómica Integrativa

Centro de Investigación en Dinámica Celular

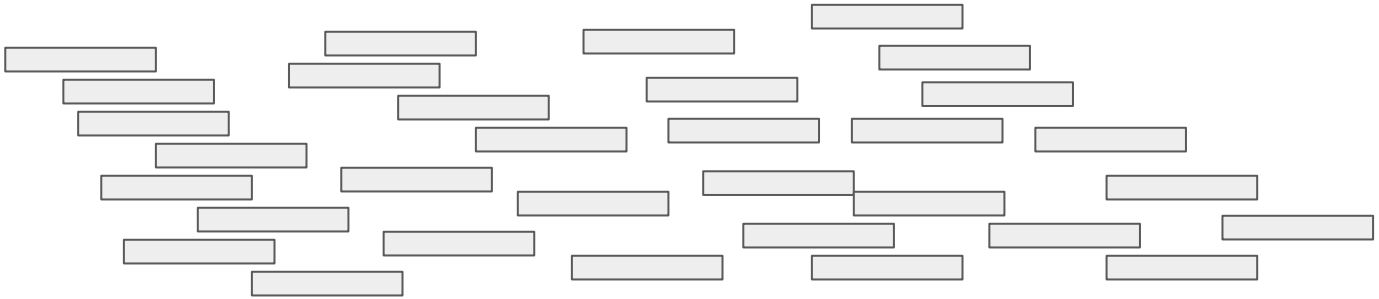
# Ensamblajes de Genomas Virales

## Módulo 2 - Práctica 2 [Ensamblaje]

# Métodos para ensamblar un Genoma

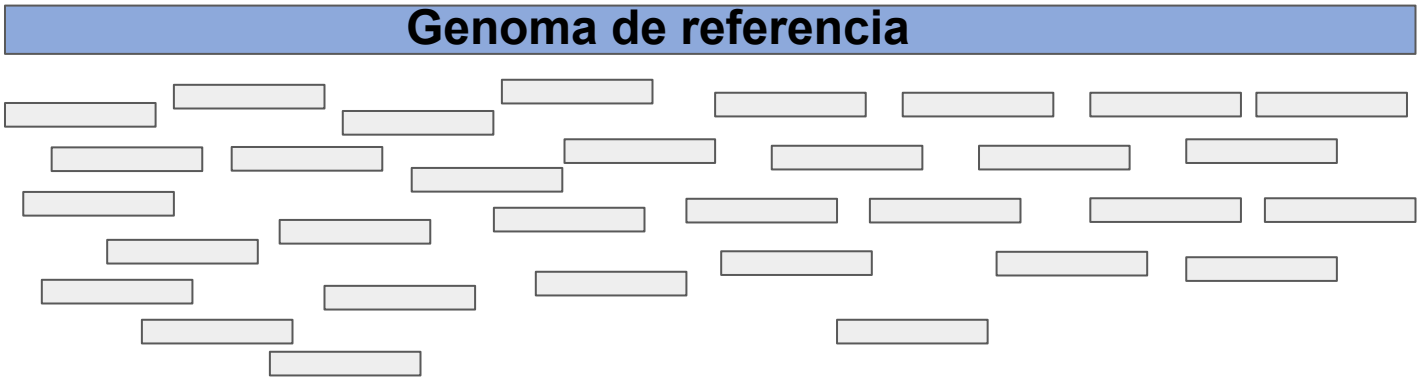
**Ensamblaje *de novo,*** Se infiere la relación de las lecturas por sobrelapamiento

En otras palabras, Comparamos lecturas entre sí.



# Métodos para ensamblar un Genoma

**¿Qué pasa si en cambio comparamos las lecturas con el genoma de referencia?**



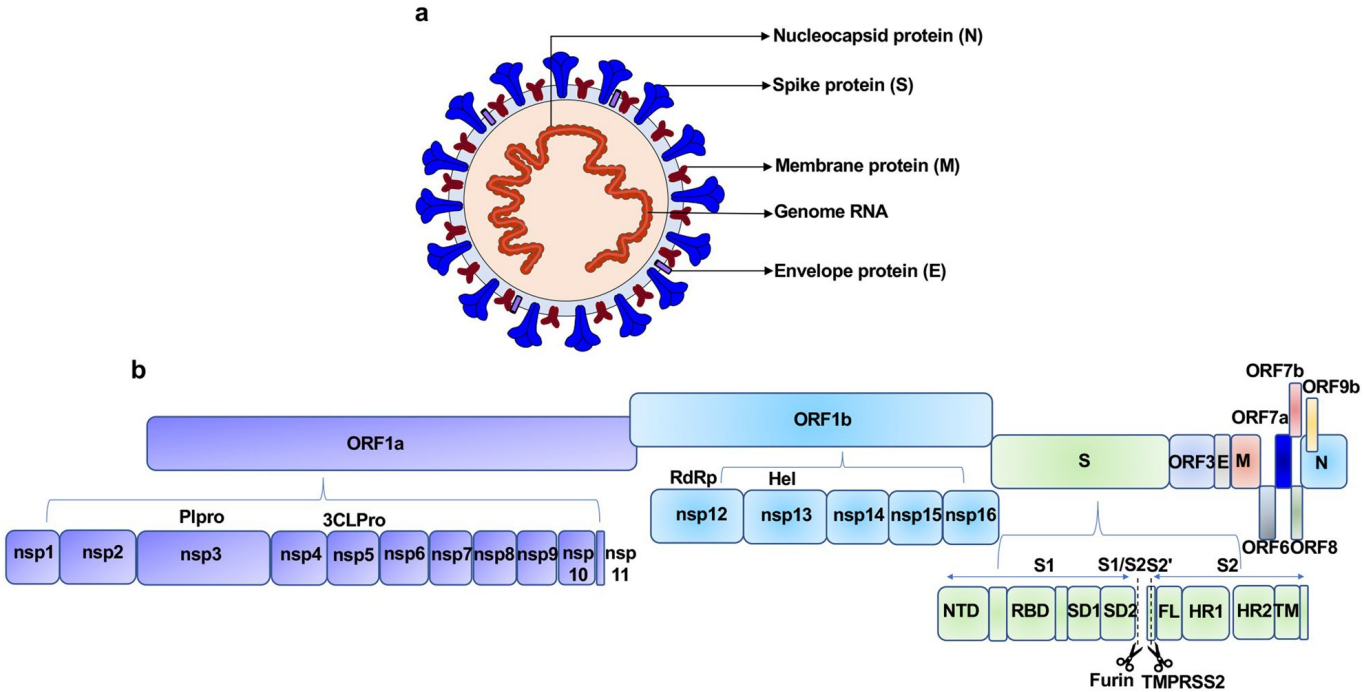
# Software

**BWA (Burrows-Wheeler Aligner)**
🔗 https://github.com/lh3/bwa
📌 Uso: Lecturas cortas de Illumina

**Bowtie2**
🔗 https://github.com/BenLangmead/bowtie2
📌 Uso: Alineamiento rápido de lecturas cortas

**Minimap2**
🔗 https://github.com/lh3/minimap2
📌 Uso: Lecturas largas de Nanopore y PacBio

# 29,903 bases - Genoma de SARS-CoV-2




Zhang et al. 2021. https://doi.org/10.1038/s41392-021-00653-w

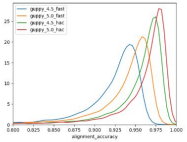
# La importancia de las mutaciones


https://outbreak.info/compare-lineages

## Guppy - Basecalling

**guppy_basecaller** -i /mnt/fast5 -s /mnt/output_dir -c dna_r9.4.1_450bps_hac.cfg -m template_r9.4.1_450bps_hac.jsn --barcode_kits "EXP-NBD196" --require_barcodes_both_ends --compress_fastq --trim_barcodes **-x "cuda:0"** --gpu_runners_per_device [num] --num_callers [num] --chunks_per_runner [num] --chunk_size [num]

- Archivos de entrada (fast5) y salida (fastq)
- Opciones del programa
- Modelo para redes neuronales (fast o hac)
- Barcode demultiplexing
- Opciones de optimización



https://mirror.oxfordnanoportal.com/software/analysis/ont-guppy_6.0.1_linux64.tar.gz

## 2. Control de calidad

- Se corre por cada barcode o muestra
- Remoción de lecturas quiméricas y filtro por tamaño

artic guppyplex --min-length 400 --max-length 700 --directory output_directory/barcode03 --prefix run_name

- Directorio donde se guardan las secuencias filtradas
- Prefijo con el que se guardan los archivos de salida

https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html

## 3. Ensamblaje (Minimap2 y Nanopolish)

- Se corre por cada barcode o muestra

artic minion --normalise 200 --threads 4 --scheme-directory ~/artic-ncov2019/primer_schemes --read-file run_name_${barcode}.fastq --fast5 directory path_to_fast5 --sequencing-summary path_to_sequencing_summary.txt nCoV-2019/V3 samplename

- Directorio donde están los datos crudos fast5
- Prefijo con el que se guardamos los archivos de salida en el paso de control de calidad
- Sequencing summary de Guppy. Único para todos los barcodes.
- Nombre de la muestra (Usualmente el barcode que se analiza)

## 4. Estimación del linaje

PANGOLIN:/ https://pangolin.cog-uk.io
https://cov-lineages.org/resources/pangolin/installation.html

Importante mantener actualizado

## 5. Revisión de mutaciones

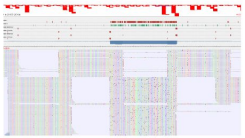
NEXTSTRAIN: https://clades.nextstrain.org/

Disponible en web y línea de comandos

## 6. Métricas

- Profundidad promedio por posición

**samtools depth** aln_${Barcode}-sorted.bam



- Cobertura

**samtools mpileup** aln_${Barcode}-sorted.bam

ó

Usar columna "Missing (N)" del análisis en Nextclade 1-N/29903

- Archivo BAM ordenado producto del alineamiento.

## 7. Reportar secuencias

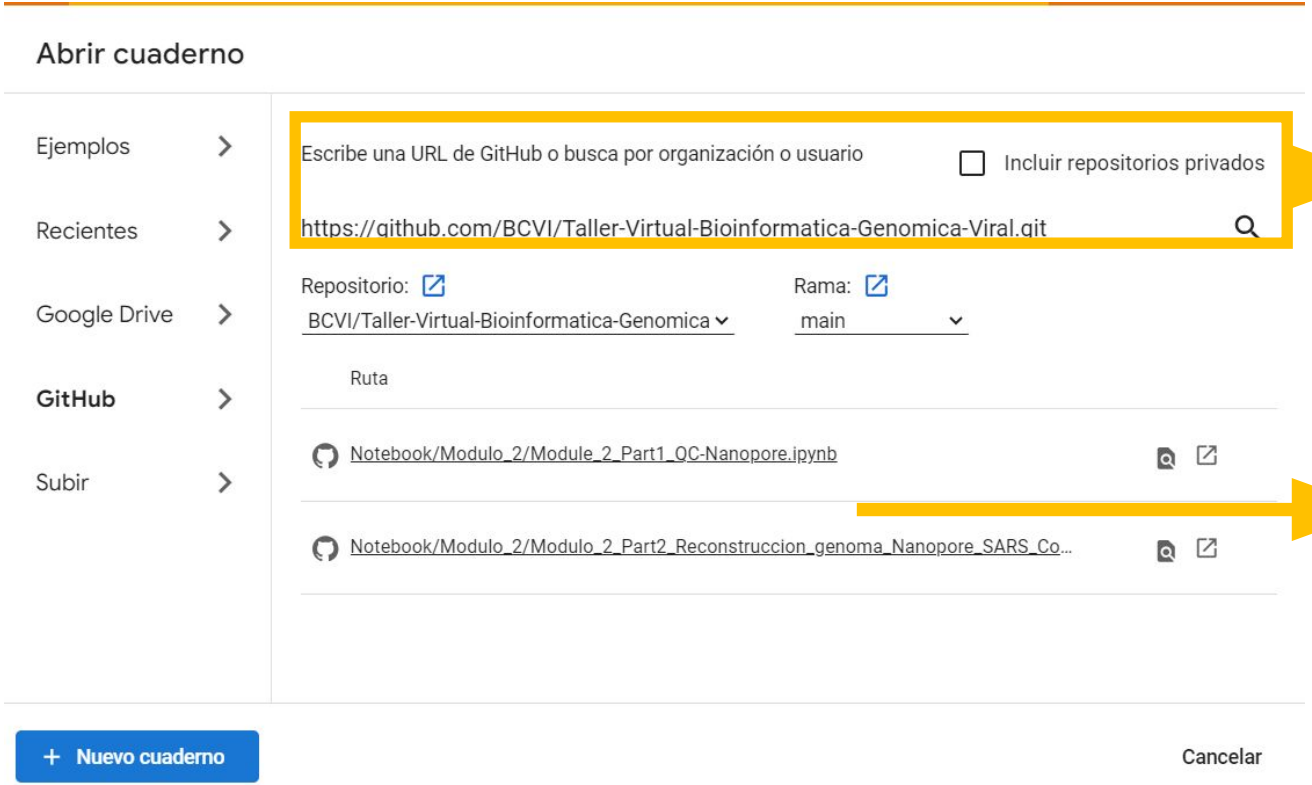
| COLUMNAS | DESCRIPCIÓN | EJEMPLO |
|---|---|---|
| **Institución que secuencia** | Nombre de la institución que realizó el experimento de secuenciación | Instituto Nacional de Salud |
| **Código lab secuenciación** | Asignado durante secuenciación | INS-COV-VG-005 |
| **Tipo de muestreo** | Descriptor estudio o tipo muestreo (Rutinario, Probabilístico, Hospitalizados, Fallecidos, Vacunados, Investigación propia, ...) | Probabilístico |
| **Lab_origen_de_la_muestra** | Nombre del laboratorio de donde proviene la muestra | SURA, IPS Saludpass, Laboratorio clínico... |
| **Código de origen de la muestra** | Código que asignó el laboratorio de origen (Si está disponible) | C-346465; 2021081563113 |
| **Documento de identidad** | Número de identificación del paciente (Si está disponible) | 1105111111 |
| **Procedencia** | Departamento de procedencia en código ISO (Ver tabla abajo) | DC; ARA; SAP... |
| **Fecha de muestra** | Fecha de toma de muestra (Si está disponible) | 12/02/2022 |
| **Frameshifts** | Cambios en el marco de lectura (reportado por nextclade o su anotación) | ORF3a:257-276 |
| **Deleciones** | Deleciones en el genoma (reportado por nextclade o su anotación) | 26158-26162 |
| **Inserciones** | Inserciones en el genoma (reportado por nextclade o su anotación) | 21991:ACT |
| **Sustituciones aminoácidos** | Todas las sustituciones de aminoácidos en el genoma (reportado por nextclade o su anotación) | N:T205I,N:D377Y,ORF1a:T1055A,ORF1a:T1538I,ORF1a:T3255I,ORF1a:Q3729R,ORF1b:P314L,ORF1b:A1131V,ORF1b:P1342S,ORF3a:Q57H,ORF3a:N257X,ORF8:T11K,ORF8:P38S,ORF8:S67F,S:T95I,S:R346K,S:E484K,S:N501Y,S:D614G,S:P681H,S:D950N |

## 8. Someter a base de datos

- Mínimo 50% de cobertura (mínimo GISAID)
- Se relacionan secuencias ingresadas con metadatos, info pacientes, información geográfica, institución que provee secuencia.
- https://www.gisaid.org/
- Códigos ISO departamentos: https://en.wikipedia.org/wiki

EPI2ME

# Cargar los notebooks




Copiar URL del curso

Notebook por módulo

https://colab.research.google.com/

# Gracias

Contacto:
crojas@ins.gov.co
paolarojasestevez@gmail.com