Logotipo

Descripción generada automáticamente

Texto

Descripción generada automáticamente con confianza baja

**CURSO TEÓRICO-PRÁCTICO:**

**EPIDEMIOLOGÍA GENÓMICA Y DESCUBRIMIENTO DE PATÓGENOS MEDIANTE SECUENCIACIÓN DE PRÓXIMA GENERACIÓN**

**/ GENOMIC EPIDEMIOLOGY AND PATHOGEN DISCOVERY BY NEXT GENERATION SEQUENCING**

**Noviembre 8 - 12, 2021**

**Introducción a ARTIC, ensamblaje por mapeo, y análisis de variantes**

Esta sección del parte del experimento de secuenciación en MinION de Nanopore. Para los datos obtenidos durante la secuenciación ya realizamos el llamado de bases con Guppy y visualización de la calidad de las lecturas con PycoQC.

Ahora pasaremos a ensamblar y obtener las secuencias de las muestras y analizarlas.

Los comandos están entre cajas en letra Courier

1. **Instalación del ambiente de ARTIC**

El protocolo de bioinformática de ARTIC está disponible en el siguiente vínculo: <https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html>

ARTIC tiene pipelines establecidos para realizar los procesos de control de calidad y ensamblaje de secuencias de SARS-CoV-2 obtenidas con su protocolo. Estas herramientas las liberaron de forma conjunta en un ambiente de conda.

Para usar estas herramientas debemos instalar el ambiente de conda.

Primero clonamos en nuestra máquina el repositorio de ARTIC:

git clone https://github.com/artic-network/artic-ncov2019.git

Ingresamos a la carpeta descargada e instalamos el ambiente usando conda o mamba.

cd artic-ncov2019

conda env remove -n artic-ncov2019

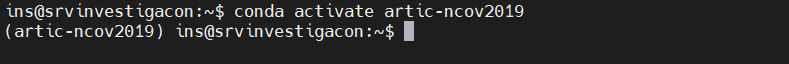
conda env create -f environment.yml

Al crear el ambiente se instalan todas las herramientas que son usadas en el pipeline. Si al usar conda se demora mucho la instalación, se puede instalar el manejador de ambientes alternativo mamba.

**Activar el ambiente de ARTIC**

Para usar las herramientas del ambiente de conda, primero debemos cargarlas al sistema activando el ambiente.

conda activate artic-ncov2019



Veremos que el ambiente de conda está activo en la terminal antes de nuestro usuario. Ya podemos usar todos los programas que contenga este ambiente, como bwa, minimap, samtools y más.

**Archivos de trabajo**

Para esta sección del taller necesitamos los archivos de setTutorial que usamos en la sección de llamado de bases con Guppy.

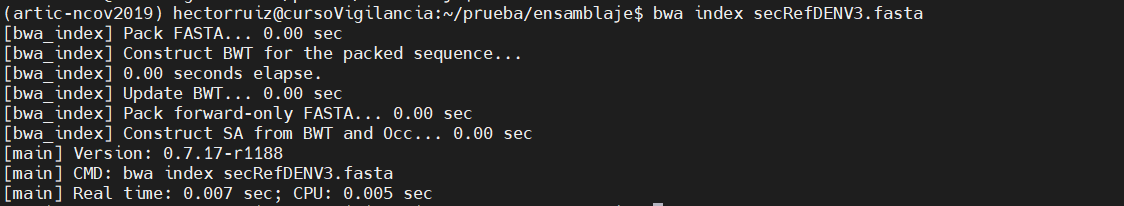
Además de esto, debemos clonar el repositorio con nuestros archivos:

git clone https://github.com/dfbautista/Curso-Epidemiologia-Genomica-Practica-Bioinformatica.git

**Parte 1. Ensamblaje por mapeo con BWA**

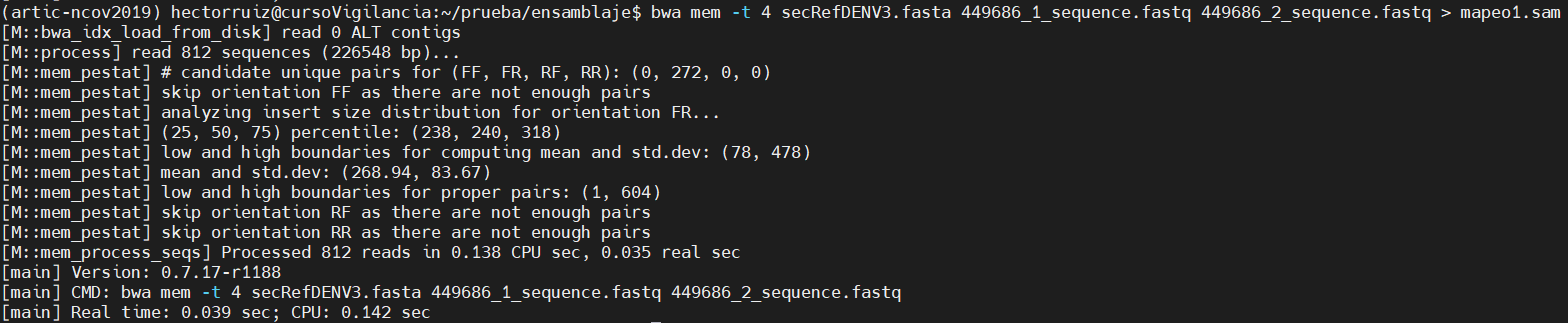
**1. Crear índice de la secuencia de referencia**

bwa index secRefDENV3.fasta



**2.Mapear lecturas en formato fasta a la referencia y obtener SAM**

bwa mem -t 4 secRefDENV3.fasta 449686\_1\_sequence.fastq 449686\_2\_sequence.fastq > mapeo1.sam



**3. Convertir SAM a BAM**

samtools view -S -b mapeo1.sam > mapeo1.bam



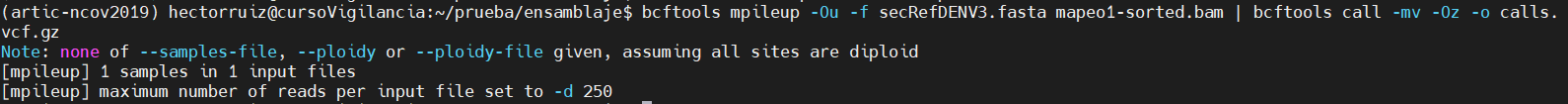
**4.Ordenar BAM**

samtools sort mapeo1.bam -o mapeo1-sorted.bam



**5.Generar archivo de llamado de variantes**

bcftools mpileup -Ou -f secRefDENV3.fasta mapeo1-sorted.bam | bcftools call -mv -Oz -o calls.vcf.gz



**6.Crear indice variantes**

bcftools index calls.vcf.gz



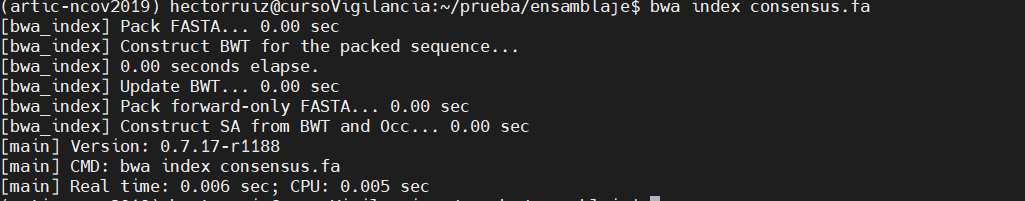
**7.Generar secuencia consenso con variantes**

cat secRefDENV3.fasta | bcftools consensus calls.vcf.gz > consensus.fa



**8.Generar secuencia índice consenso con variantes**

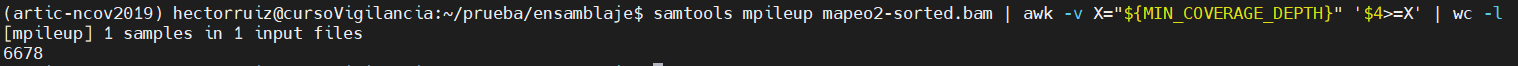
bwa index consensus.fa



**Métricas para profundidad y cobertura**

**12. Determinar cobertura con número de bases ensambladas respecto al genoma de referencia**

samtools mpileup mapeo2-sorted.bam | awk -v X="${MIN\_COVERAGE\_DEPTH}" '$4>=X' | wc –l



**13. Determinar profundidad con el número de reads**

samtools depth mapeo2-sorted.bam | awk '{sum+=$3} END { print "Average = ",sum/NR}'



**Observar variants**

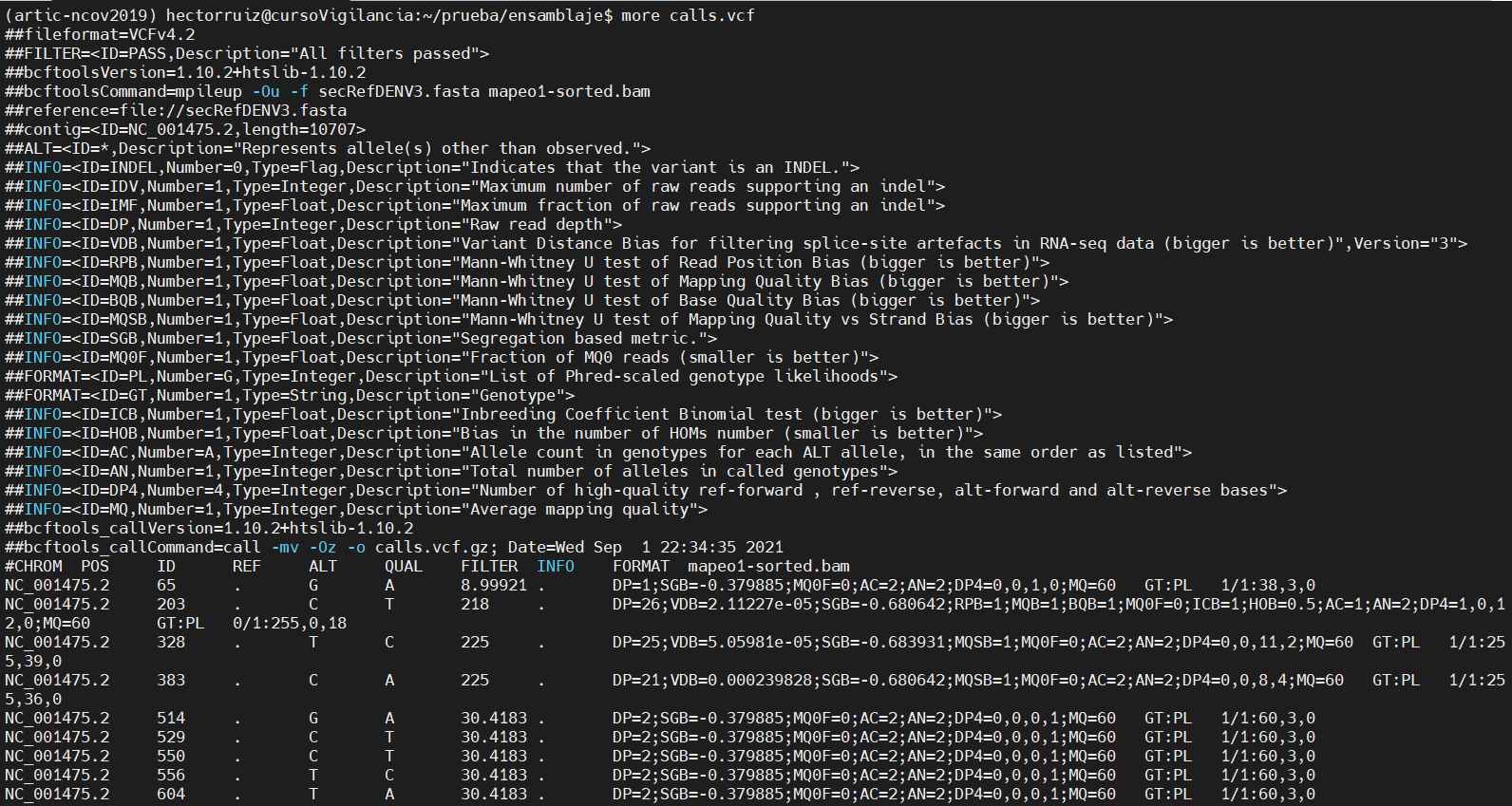
**14. Descomprimir vcf**

gunzip calls.vcf.gz



**15. Leer el archive vcf**

more calls.vcf o zcat calls.vcf.gz



Podemos usar las herramientas visuales de Tablet o IGV para observar los alineamientos de nuestras lecturas a partir de los archivos sorted.bam junto con el índice sorte.bai

IGV: <http://software.broadinstitute.org/software/igv/download>

Tablet: <https://ics.hutton.ac.uk/tablet/download-tablet/>

**Parte 2. Pipeline de ARTIC para ensamblar secuencias de SARS-CoV-2**

1. **Crear un directorio de trabajo**

Primero creemos un directorio para nuestro análisis

mkdir -p ~/taller/artic

cd ~/taller/artic

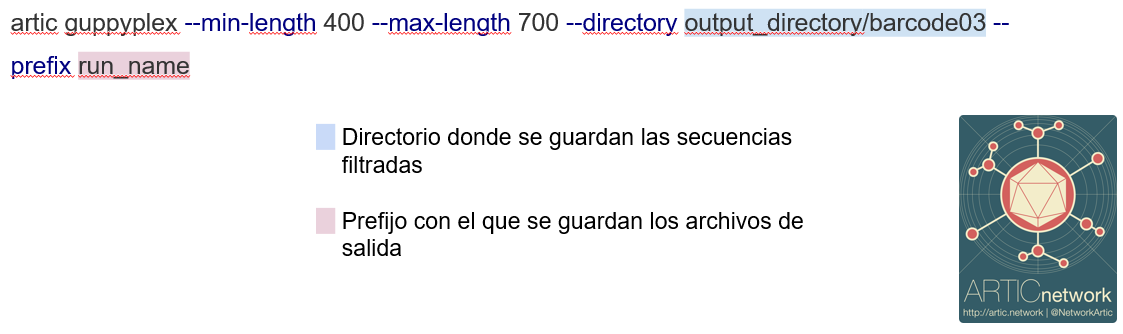
1. **Realizar el filtro por tamaño**

Las lecturas deben ser filtradas por tamaño para remover fragmentos pequeños y posibles lecturas quiméricas.

Este proceso se realiza para cada muestra o barcode:

artic guppyplex --min-length 400 --max-length 700 --directory output\_directory/barcode03 --prefix run\_name

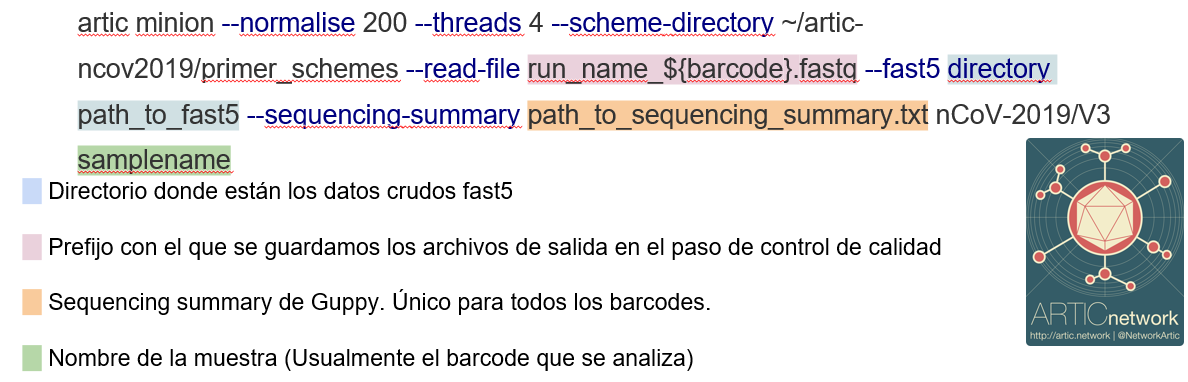
corremos el comando artic guppyplex, definiendo el tamaño mínimo y máximo que esperamos de las lecturas, y definimos un prefijo para el experimento.



1. **Realizar el ensamblaje**

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

El comando artic minion contiene todos los pasos que se realizan durante el ensamblaje. Esto incluye el alineamiento de las lecturas a la secuencia de referencia usando Minimap2, el llamado de variantes teniendo en cuenta los archivos fast5 con nanopolish, el manejo de los archivos con bcftools y samtools.



Vamos a repetir este proceso para cada barcode.

1. **Revisar los archivos de salida**

Al finalizar el proceso debemos tener una carpeta para cada barcode, y dentro de cada una podemos ubicar los archivos generados durante el ensamblaje. Estos incluyen los alineamientos de las lecturas a la secuencia de referencia en formato bam.

Para cada barcode también tenemos un archivo fasta con las secuencias.

1. **Unir las secuencias en un solo archivo fasta**

Para analizar las secuencias es más sencillo tenerlas todas juntas en un solo archivo fasta.

Para esto podemos usar el comando cat para concatenar todos los archivos y guardarlos en un solo fasta.

cat barcode\*/barcode\*.consensus.fasta > todas\_consensus.fasta

Podemos revisar que estén todos los barcodes en el archivo con el comado grep

grep “barcode” todas\_consensus.fasta

1. **Estimación de linajes**

Para saber a qué linaje perteneces las secuencias de nuestras muestras podemos usar la herramienta pangolin. Esta herramienta asignará el linaje más cercano de acuerdo a un modelo de árbol de decisión desarrollado por el Pango Network.

Para instalar pangolin podemos seguir estas instrucciones: <https://cov-lineages.org/resources/pangolin/installation.html>

conda install -c bioconda -c conda-forge -c defaults pangolin

Nuevamente si la instalación es muy demorada se puede usar mamba para instalar

Para comprobar que la instalación fue exitosa y ver las instrucciones de uso corremos el comando de ayuda:

pangolin –h

Es recomendable que cada día que usemos la herramienta pangolin, la actualicemos antes de usar ya que los modelos de decisión implementados para asignar linaje son actualizados frecuentemente para incluir nuevos linajes y mejorar la asignación de linajes existentes.

pangolin --update

El comando pangolin lo podemos correr con opciones para definir el nombre de archivo de salida, y el tipo de asignación. Vamos a usar la forma más común:

pangolin todas\_consensus.fasta

El resultado es una tabla en formato separado por comas llamado lineage\_report.csv.

Lo podemos abrir en Excel, GoogleSheets, LibreOffice.

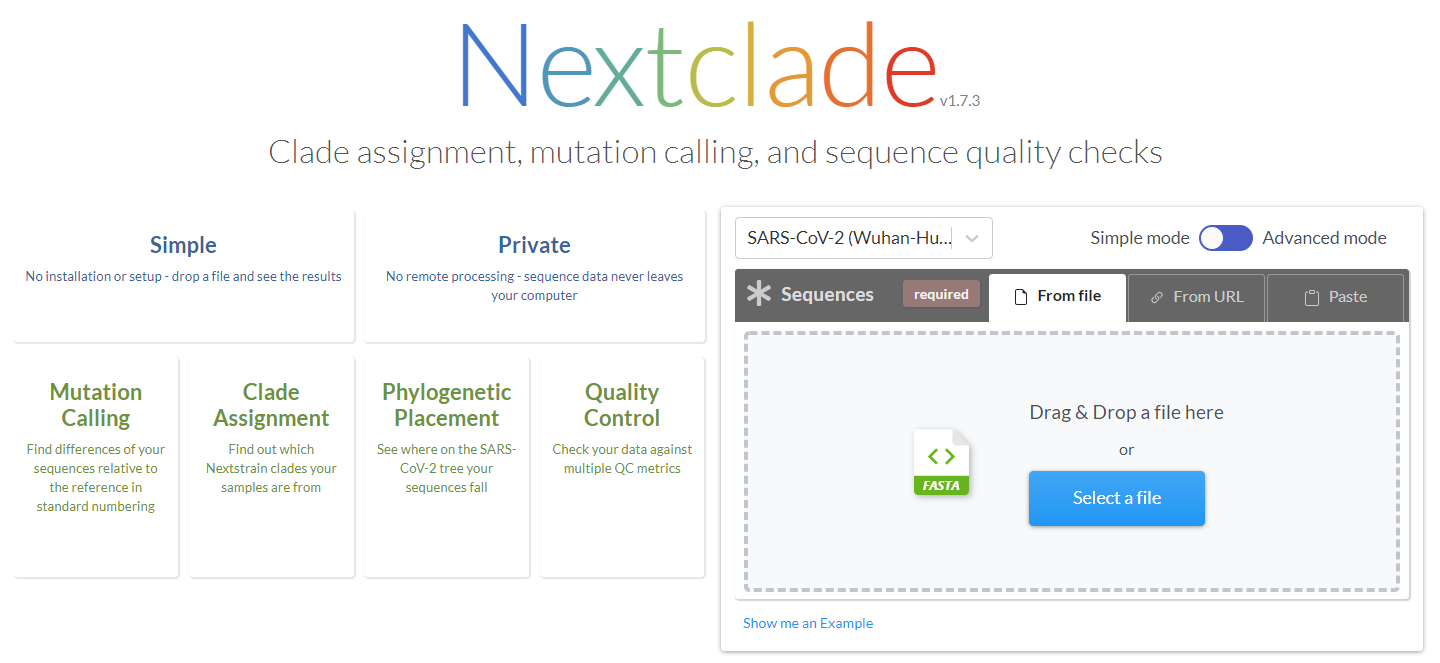
1. **Anotación de mutaciones con Nextclade**

La anotación de inserciones, deleciones, y sustituciones, así como otra información que podamos obtener de la secuencia, se puede anotar usando la herramienta Nextclade.

Se puede instalar esta herramienta desde el sitio web como un ejecutable de un solo archivo: <https://github.com/nextstrain/nextclade/releases/latest/download/nextclade-Linux-x86_64>

Vamos a usar en este tutorial la herramienta en versión online en la siguiente página:

<https://clades.nextstrain.org/>



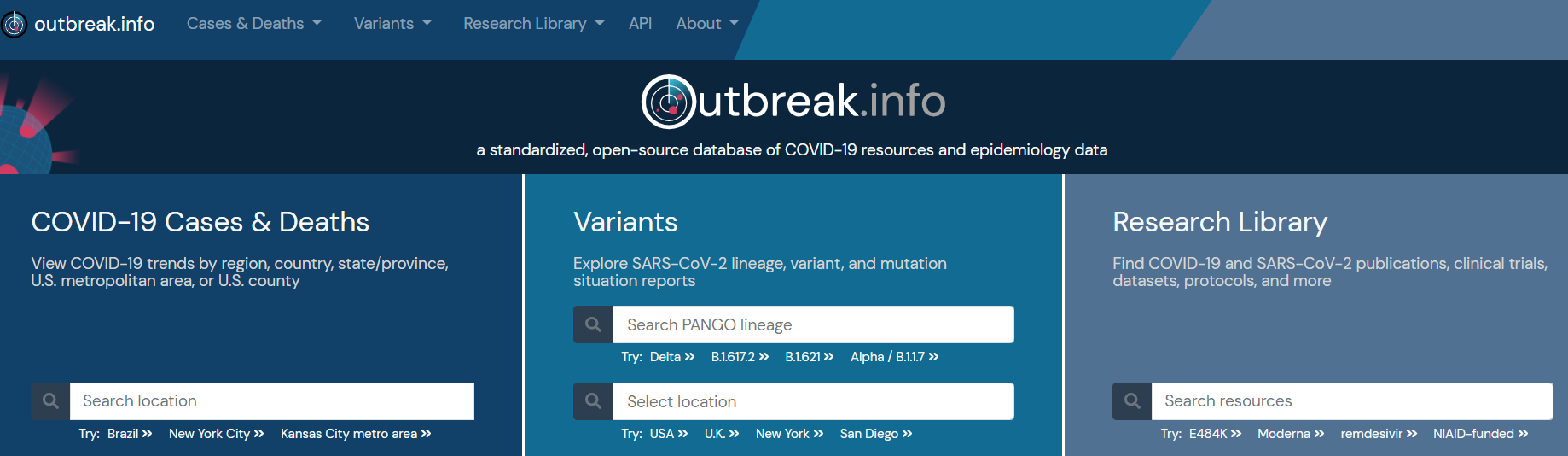
Los resultados de Nextclade los podemos visualizar en el sitio web, pero también los podemos descargar en un archivo csv que tiene información más completa.

1. **Análisis de variantes**

Con la información de los resultados de pangolin y Nextclade podemos evaluar las secuencias para saber a qué linaje es probable que pertenezcan y qué mutaciones tienen.

Es posible que las secuencias contengan mutaciones novedosas, por lo que es importante tener en cuenta cuáles son las mutaciones que contienen distintos linajes.

Para explorar esta información vamos a usar las herramientas de <https://outbreak.info/>



Particularmente nos interesa conocer qué mutaciones son frecuentes en los linajes que encontramos. <https://outbreak.info/compare-lineages>