Curso Teórico-Práctico

EPIDEMIOLOGÍA GENÓMICA

UNA HERRAMIENTA PARA FORTALECER LA VIGILANCIA DE AGENTES INFECCIOSOS

Día 3: ***Análisis Metagenómicos y Plataformas para Divulgación***

Bogotá / Septiembre 25 del 2024

Análisis de metagenómica (*mNGS*)

Paola Rojas - Estevez
M.Sc Biología Computacional
Grupo Genómica de Microorganismos Emergentes
Instituto Nacional de Salud de Colombia

Flujo de trabajo de mNGS

Diseño del estudio

- Muestras
- Controles
- Almacenamiento
- Condiciones de laboratorio

Extracción de ácidos nucleicos

(1-2 hrs)

- Tipos
- Controles
- Control de calidad

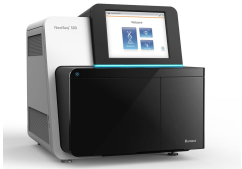
Preparación de librerías

(1-2 días)

- Remoción hospedero
- Fragmentación
- Generación de dsDNA
- End-repair
- Ligación de adaptadores
- Barcoding
- Limpieza
- Pooling
- Control de calidad

Secuenciación

(0.5-3 días)



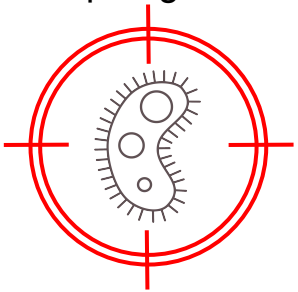
- Plataforma
- Longitud de reads
- Concentración de carga

Análisis de datos

(0.5-4 hrs)

- Control de calidad de datos
- Remoción hospedero
- Alineamiento
- Blast
- Puntuación de hits

Identificación de patógenos



Flujo de trabajo de mNGS

Diseño del estudio

Preparación de librerías

Secuenciación

Extracción de ácidos nucleicos

$

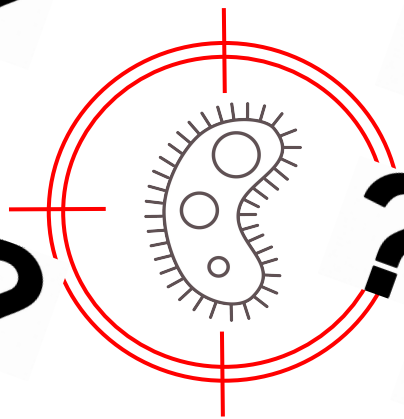
Análisis de datos

(0.5-4 hrs)

IDSEQ

- Control de calidad de datos
- Remoción hospedero
- Alineamiento
- Blast
- Puntuación de hits

Identificación de patógenos



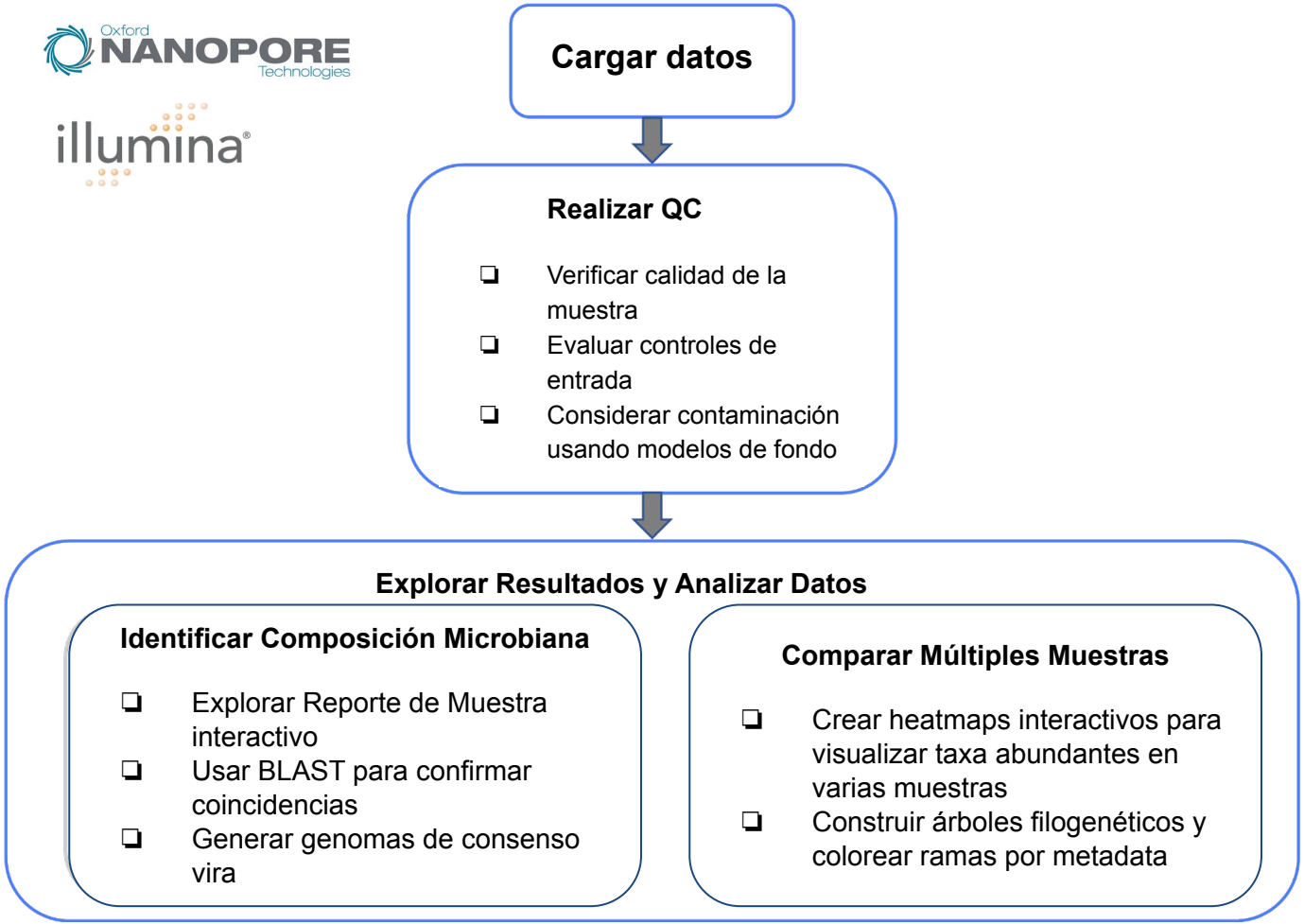
- **Software gratuito**
- **Cloud-based pipeline**
- **Número libre de muestras**
- **Número libre de trabajos**
- **Pipeline de fondo**
- **Instructivo detallado**
- **Asistencia técnica y servicios:**
  help@czid.org

Módulos

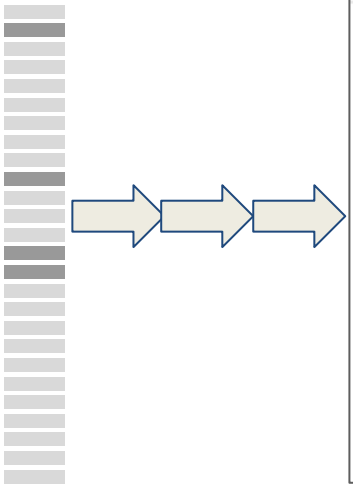
Módulo	Datos de entrada	Soporte de tecnología de secuenciación
Metagenómica (mNGS)	Datos metagenómicos imparciales *(sin datos de amplicón/16 s/18 s)*	Illumina y Nanopore
Resistencia a los antimicrobianos	Secuenciación del genoma completo y mNGS	Ilumina
SARS-CoV-2	Secuenciación de amplicones	Illumina y Nanopore
Genoma de consenso viral	Secuenciación de amplicones y WGS	Ilumina

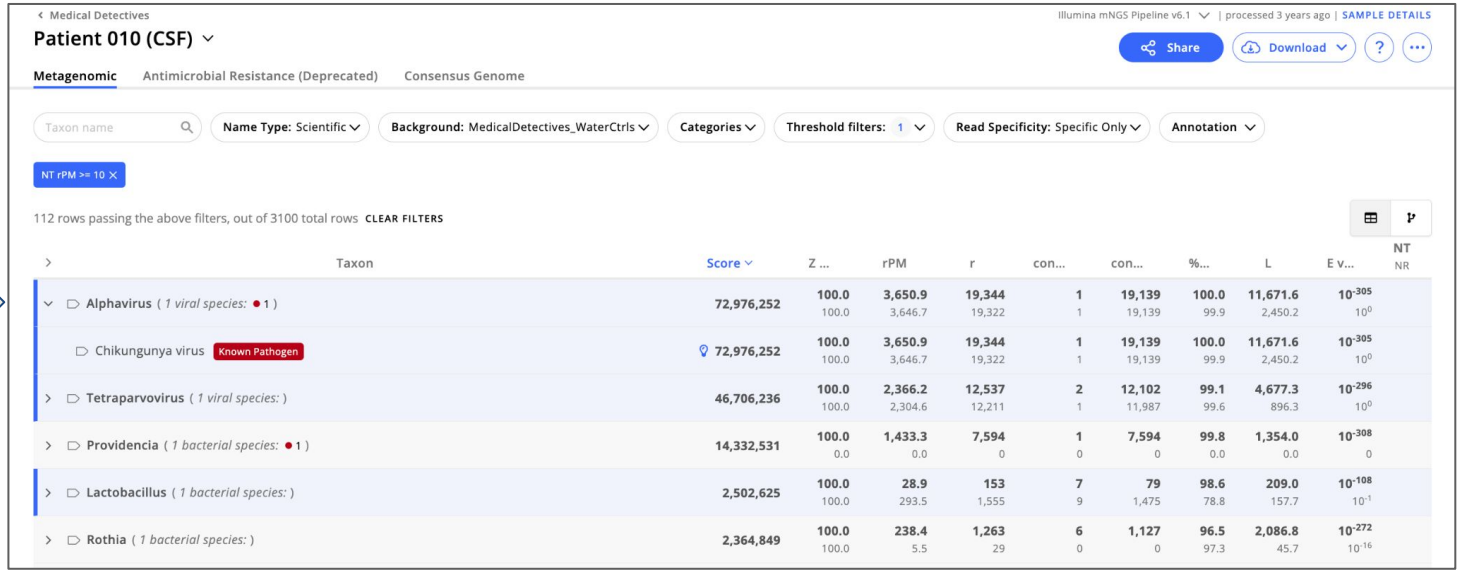
¿Cómo abordar el análisis de datos de secuencias metagenómicas de próxima generación (mNGS) a través de CZ ID?

Descripción general del análisis de datos de mNGS



Raw reads



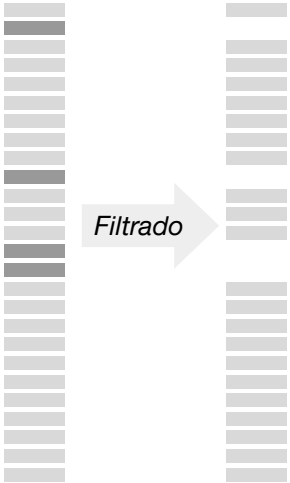


Objetivo Pipeline

- **Asignar lecturas** filtradas de calidad a **taxones** mediante mapeo o alineación con bases de datos de referencia
- **Ensamblar** contigs y **mapear** lecturas a contigs **para aumentar el soporte para taxones** identificados (alineación basada en ensamblaje)

Lecturas sin procesar

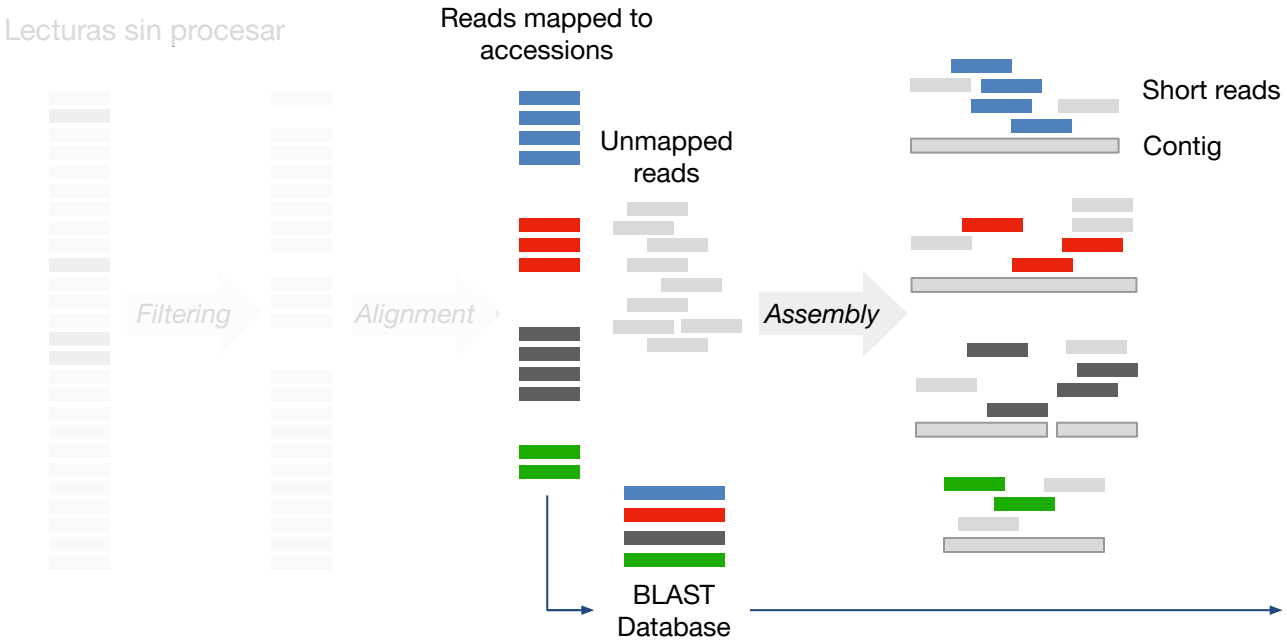
Lecturas sin procesar



Paso 1: Control de calidad y filtrado del host

- Filtrar lecturas de **baja calidad/ complejidad y adaptadores** (fastp)
- **Eliminación** de lecturas del **host** (Bowtie seguido de HISAT2)
- **Eliminación** de lecturas **humanas** (Bowtie seguido de HISAT2)
- Contraer **lecturas duplicadas** (CZID-dedup)
- **Submuestreo** a 2 millones de lecturas o 1 millón de pares de lecturas

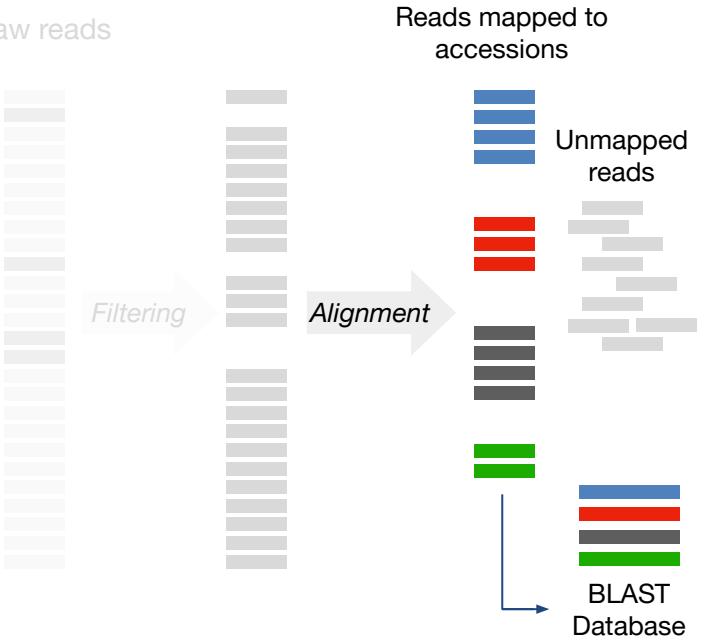
Bioinformatic tool/software



Paso 3: ensamblar contigs y asignar lecturas a contigs

- Ensamble las lecturas en contigs mediante **ensamblaje de novo** (SPAdes).
- Alinee las lecturas con los contigs ensamblados (Bowtie2).

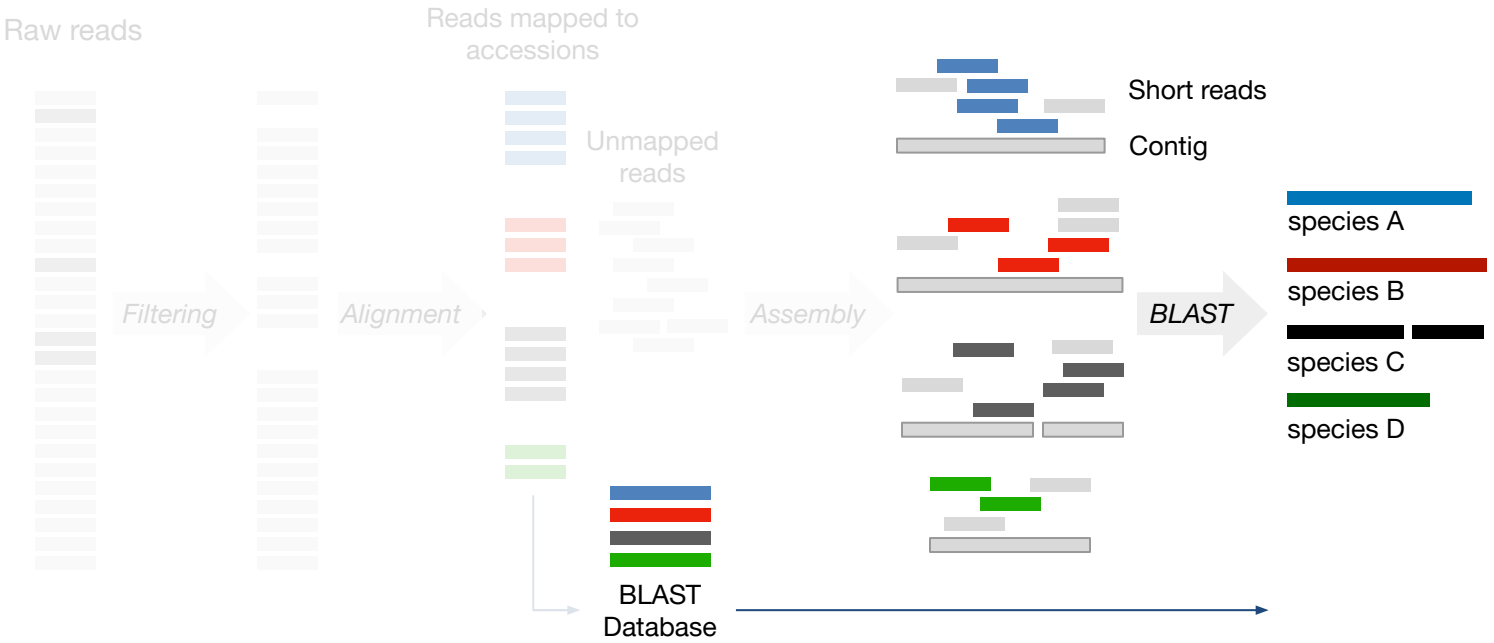
Bioinformatic tool/software



Paso 2: Alineamiento de lecturas con las bases de datos del NCBI

- Mapeo de las lecturas en bd de nucleótidos (**NT**) y proteínas (**NR**) para obtener una asignación preliminar para cada lectura (minimap2 y DIAMOND).
- Genere una base de datos BLAST personalizada con todos los taxones identificados con asignaciones preliminares.

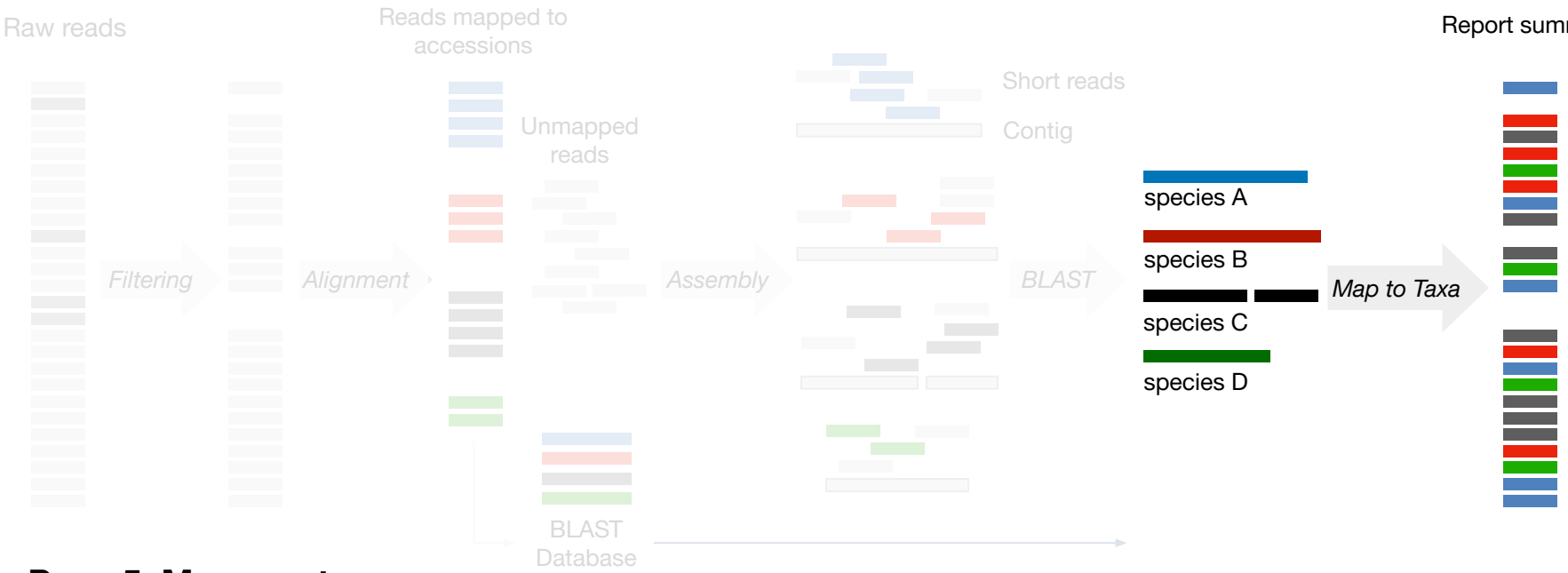
Bioinformatic tool/software



Paso 4: Alinear los contigs con la base de datos personalizada

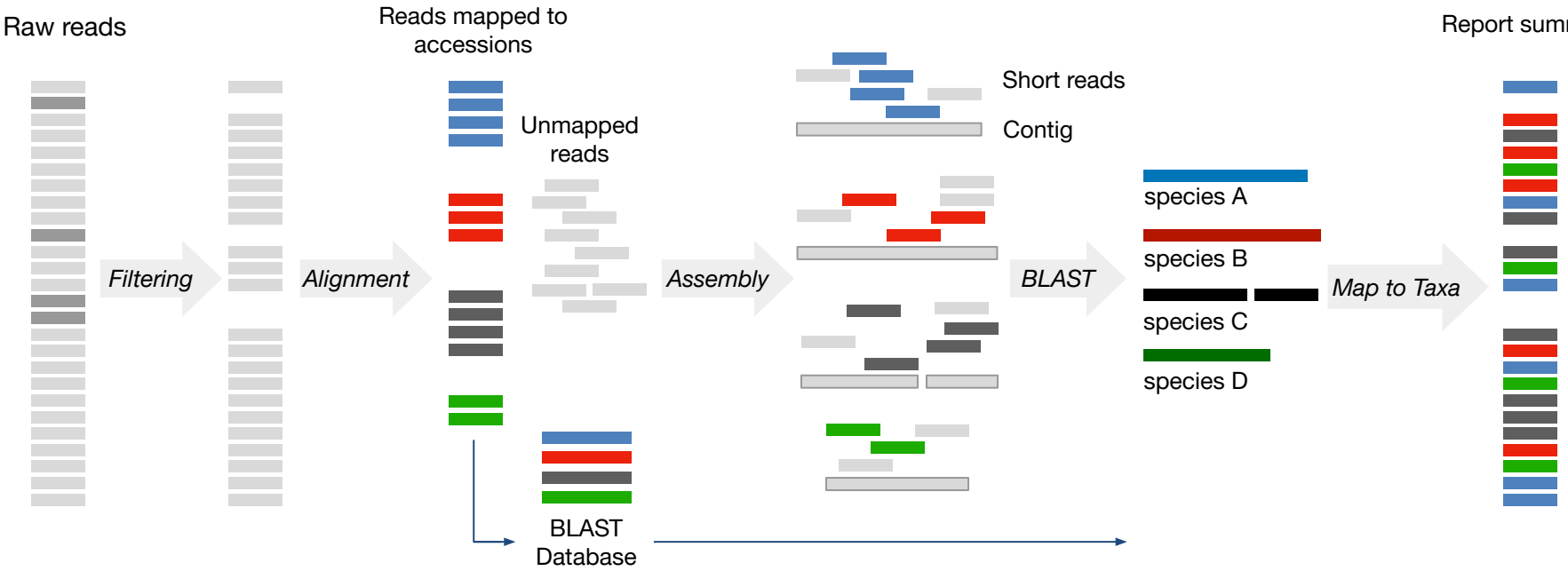
- Se alinea los contigs en la bd generada con las asignaciones preliminares (BLAST)

Bioinformatic tool/software

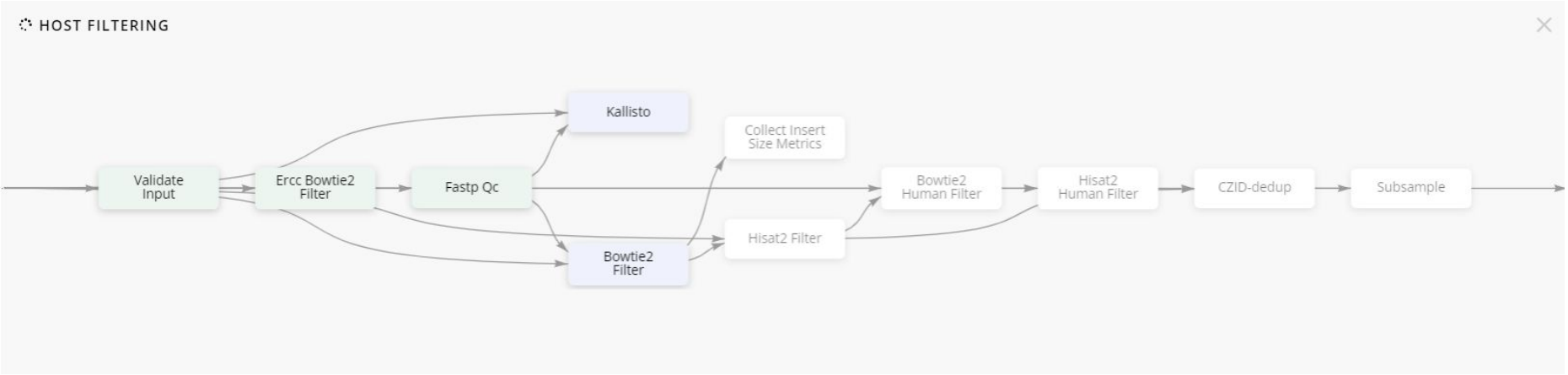


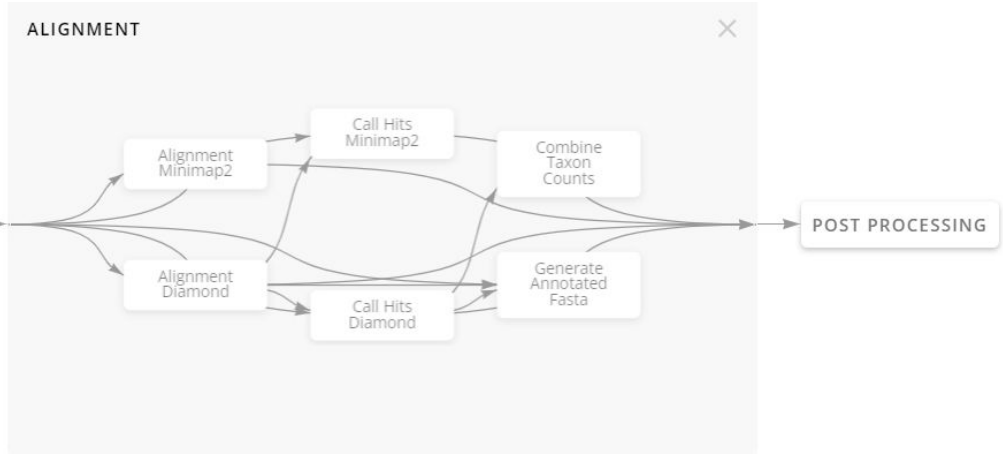
Paso 5: Mapeo a taxones

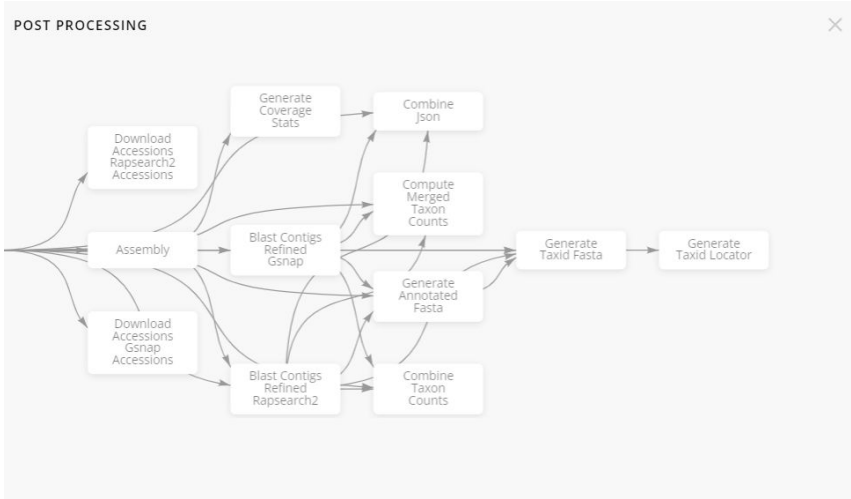
- **Asigna a cada lectura una accesión final** en función de:
  1. Accesión de contig, o
  2. Accesión lectura no se ensambló en un contig
- **Calcular estadísticas** (p. ej., lecturas por millón para cada taxón)


Raw reads
Filtering
Alignment
Reads mapped to accessions
Unmapped reads
BLAST Database
Assembly
Short reads
Contig
BLAST
species A
species B
species C
species D
Map to Taxa
Report summary

Pipeline


HOST FILTERING
Validate Input
Ercc Bowtie2 Filter
Fastp Qc
Kallisto
Collect Insert Size Metrics
Bowtie2 Filter
Hisat2 Filter
Bowtie2 Human Filter
Hisat2 Human Filter
CZID-dedup
Subsample


ALIGNMENT
Alignment Minimap2
Alignment Diamond
Call Hits Minimap2
Call Hits Diamond
Combine Taxon Counts
Generate Annotated Fasta
POST PROCESSING


POST PROCESSING
Download Accessions Rapsearch2 Accessions
Assembly
Download Accessions Gsnap Accessions
Generate Coverage Stats
Blast Contigs Refined Gsnap
Blast Contigs Refined Rapsearch2
Combine Json
Compute Merged Taxon Counts
Generate Annotated Fasta
Combine Taxon Counts
Generate Taxid Fasta
Generate Taxid Locator

Carga de Datos

Para cargar nuevas muestras en CZ ID, inicie sesión en la aplicación con su correo electrónico y contraseña.

Haga clic aquí para obtener instrucciones paso a paso sobre cómo cargar datos en CZ ID para el análisis mNGS.

https://chanzuckerberg.zendesk.com/hc/en-us/articles/360035296573-Upload-mNGS-Illumina-Data-through-the-Web-App

Informe de la Muestra: Métricas y análisis

La muestra del informe presenta métricas para ayudar a identificar taxones de interés.

NT = DB nucleótidos
NR = DB proteínas

Si yo veo:	**Bacteria**	**Virus**	**Eukaryote**
Solo NT	Región Conservada (16s)	?	Región Conservada (18s)
Solo NR	No es un hit Real	Podría ser un virus divergente	No es un hit Real
Ambos	Hit real	Hit real	Hit real

¿Búsqueda más estricta de secuencias similares?

¿Qué pasa con las secuencias no codificantes?

La muestra del informe presenta métricas para ayudar a identificar taxones de interés.

NT = nucleotide database
NR = protein database

Si yo veo:	**Bacteria**	**Virus**	**Eukaryote**
Solo NT	Podría estar afectando la región 16s	Potencialmente afectando la región no codificante	Región Conservada (18s)
Solo NR	No es un hit Real	Podría ser un virus divergente	No es un hit Real
Ambos	Hit real	Hit real	Hit real

Identificar patógenos conocidos, debe centrarse en los valores de las coincidencias en la base de datos NT. Por el contrario, si sospecha de un patógeno nuevo, debe centrarse en los valores de la base de datos NR.

Añadir métricas complementarias a cada taxón

rPM	Por cada millón de lecturas secuenciadas nos indica el **número de lecturas que se alinean con el taxón**	Alto rPM = Más confianza en el resultado
r	Número de reads alineado por taxón	Alto r = Más confianza en el resultado
Contig	Número de contigs ensamblados	contig largos = Más confianza
Contig r	Número de Lecturas que se alinean a contigs ensamblados	Alto r = Más confianza
L	Longitud promedio de alineación de contigs y lecturas	Largo= Más confianza, unless it is divergent
E value	Significativo: el valor medio esperado de alineamiento	Bajo = Más confianza en el resultado

Los filtros son útiles para eliminar resultados falsos, pero no son perfectos

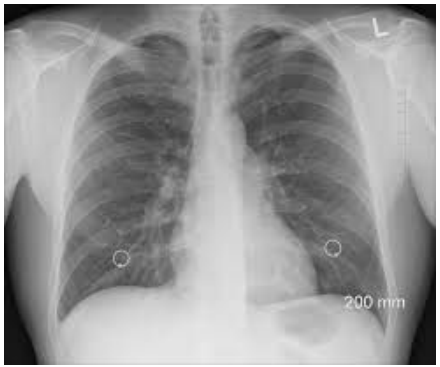
Usar los siguientes filtros es un buen punto de partida:

- **NT L >= 50 [eliminar taxones con longitudes de alineamiento cortas]**
  - Solo quiero ver taxones con al menos 50 pb que se alineen con la base de datos de nucleótidos en NCBI.

- **NT rPM >= 10 [eliminar taxones con baja abundancia]**
  - Solo quiero ver taxones con al menos 10 lecturas por millón que se alineen con la base de datos de nucleótidos. Las lecturas por millón son un valor normalizado que tiene en cuenta el total de lecturas.

- **NR rPM >= 1 [eliminar taxones que solo tienen coincidencias con regiones no codificantes, por ejemplo, 16S/18S]**
  - Solo quiero ver taxones con al menos 1 lectura que se alinee con la base de datos de proteínas en NCBI.

Análisis de mNGS con una y múltiples muestras



🧪**Actividad:** Análisis de una muestra

Culex erythrothorax

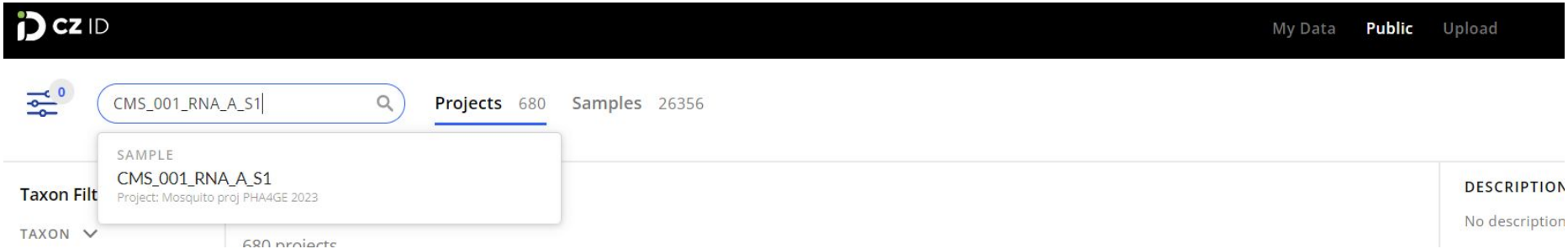
- ❏ Un solo individuo
- ❏ Secuenciación: Illumina
- ❏ Este mosquito se alimentó de sangre

¡Veamos qué taxones podemos encontrar!

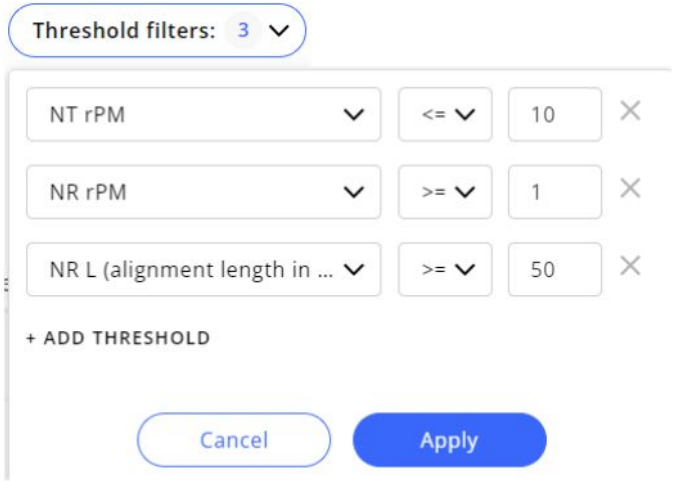
⏳ Tiempo: ~ 30 minutos para la actividad y 10 minutos para la discusión

https://github.com/instituto-nacional-de-salud/Curso-internacional-de-epidemiologia-genomica-2024/blob/main/Sesiones%20Pr%C3%A1cticas/D%C3%ADa%203/mNGS/Actividad_1_%20Pat%C3%B3genos%20transmitidos%20por%20vectores.docx

Paso a Paso



1. Acceder: [CMS_001_RNA_A_S1]



1. ¿Cuántas filas de taxones están presentes ahora?
2. ¿Cuál es el taxón más abundante según las lecturas por millón (rPM)?

¿Qué pasa con los virus divergentes?

Las mutaciones se acumulan más rápido a nivel de nucleótidos.

NR rPM ≥ 20 [Solo quiero ver taxones con 20 o más lecturas alineadas con la base de datos de proteínas en NCBI]

NR % ID ≤ 80 [Solo quiero ver taxones con menos del 80% de identidad con taxones en la base de datos de proteínas en NCBI]

1. **¿Cuántos géneros hay? ¿Cuáles son?**

A que no sabías que …

1. Este conjunto de datos proviene del artículo "‘Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay’".

```
LOCUS       QRW41732                399 aa            linear   ENV 19-MAY-2021
DEFINITION  MAG: nucleoprotein [Miglotas virus].
ACCESSION   QRW41732
VERSION     QRW41732.1
DBLINK      BioProject: PRJNA605178
            BioSample: SAMN14051441
            Sequence Read Archive: SRR11035377
DBSOURCE    accession MW434677.1
KEYWORDS    ENV; Metagenome Assembled Genome; MAG.
SOURCE      Miglotas virus (mosquito metagenome)
  ORGANISM  Miglotas virus
            Viruses; Riboviria; Orthornavirae; Negarnaviricota;
            Polyploviricotina; Ellioviricetes; Bunyavirales; Phasmaviridae;
            Orthophasmavirus; Orthophasmavirus miglotasense.
```

nuevo virus que no ha sido incluido en la base de datos de NCBI …

Pariente más cercano:
Porcentaje de identidad (38%) con el culex orthophasmavirus.

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
nucleoprotein [Miglotas virus]	826	826	57%	0.0	100.00%	QRW41732.1
nucleoprotein [Miglotas virus]	825	825	57%	0.0	99.75%	QRW41744.1
nucleoprotein [Miglotas virus]	824	824	57%	0.0	99.75%	YP_010840373.1
hypothetical protein QK753_s3gp1 [Miglotas virus]	269	269	18%	6e-83	100.00%	YP_010840372.1
hypothetical protein [Miglotas virus]	267	267	18%	3e-82	99.23%	QRW41755.1
hypothetical protein [Miglotas virus]	267	267	18%	4e-82	99.23%	QRW41737.1
nucleocapsid [Aedes japonicus bunyavirus 2]	261	261	50%	2e-76	41.29%	WCJ14327.1
nucleocapsid protein [Culex phasma-like virus]	255	255	53%	1e-73	38.74%	QHA33851.1

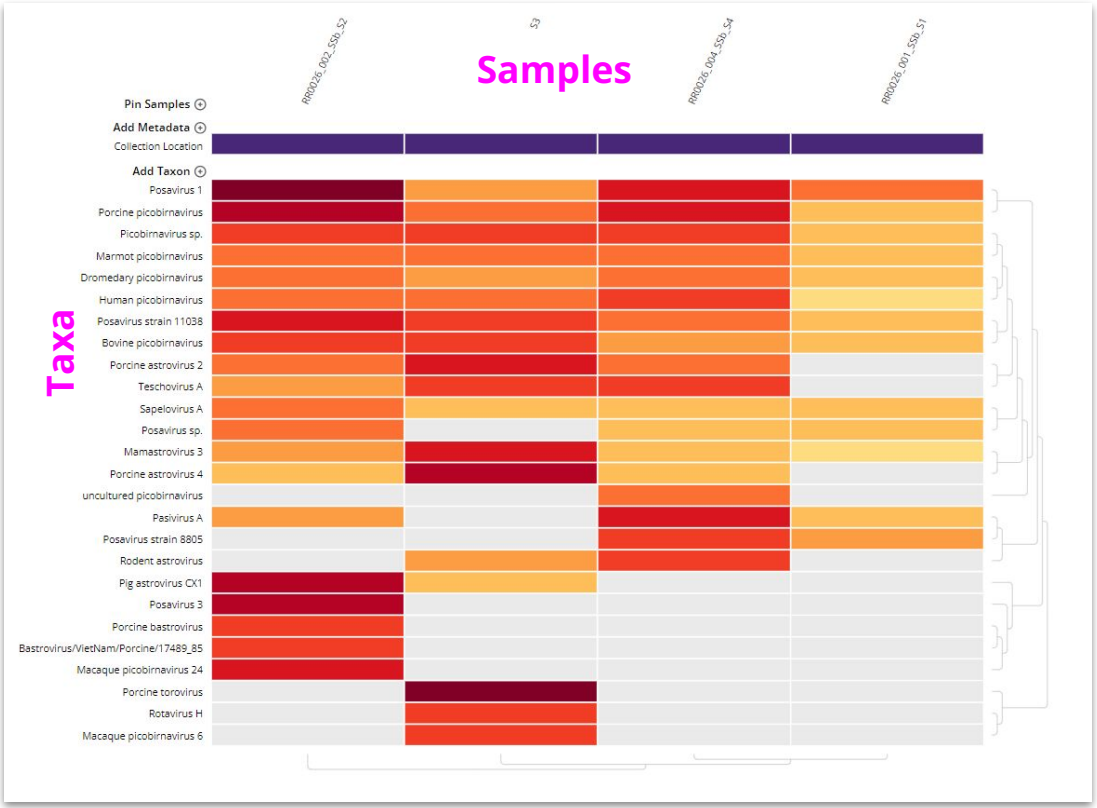
- ❏ Métricas para el análisis
- ❏ Identificación de hits
- ❏ Utilizar filtros

Análisis de mNGS múltiples muestras – Heatmap

Análisis de cohortes

- ¿Hay patrones en mis muestras??

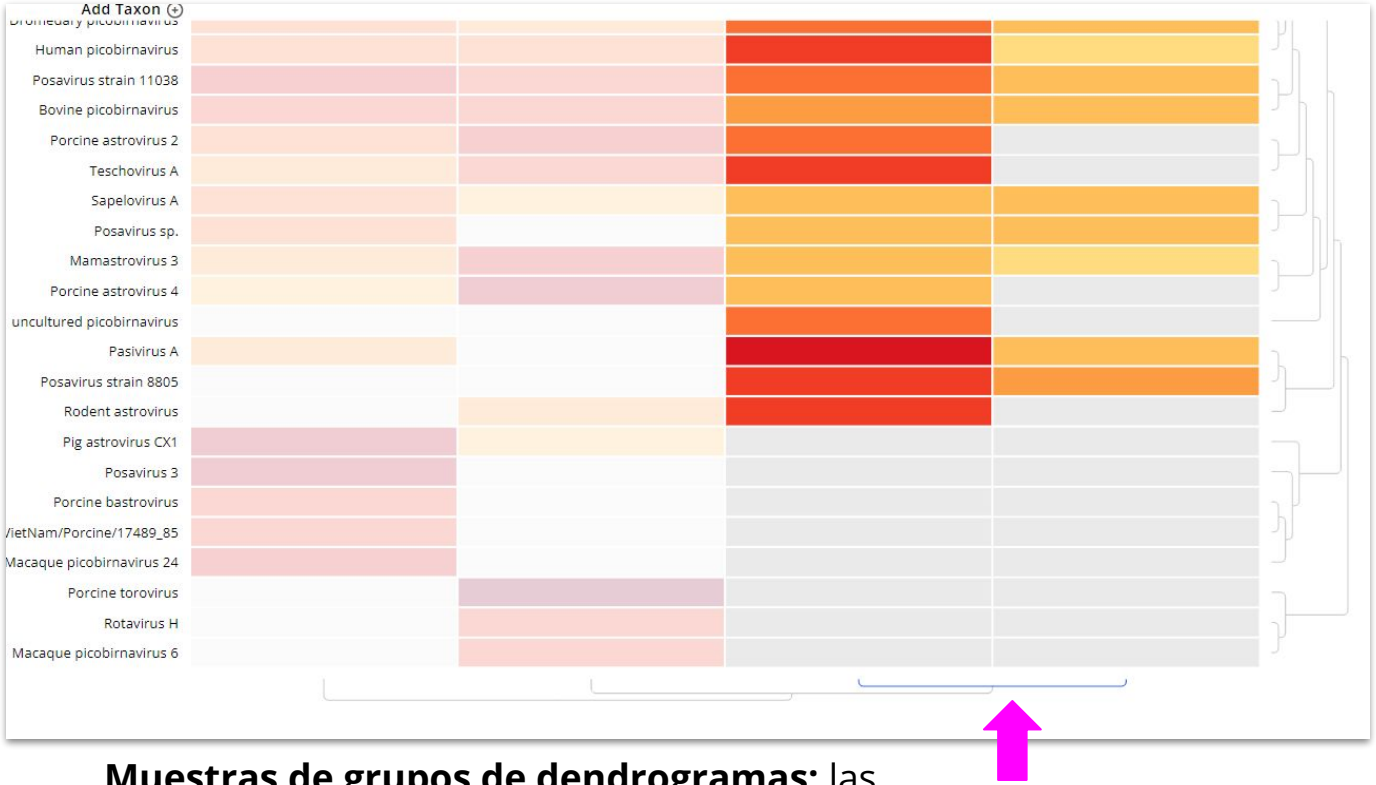
Interpretando un heatmap



Less abundant

More abundant

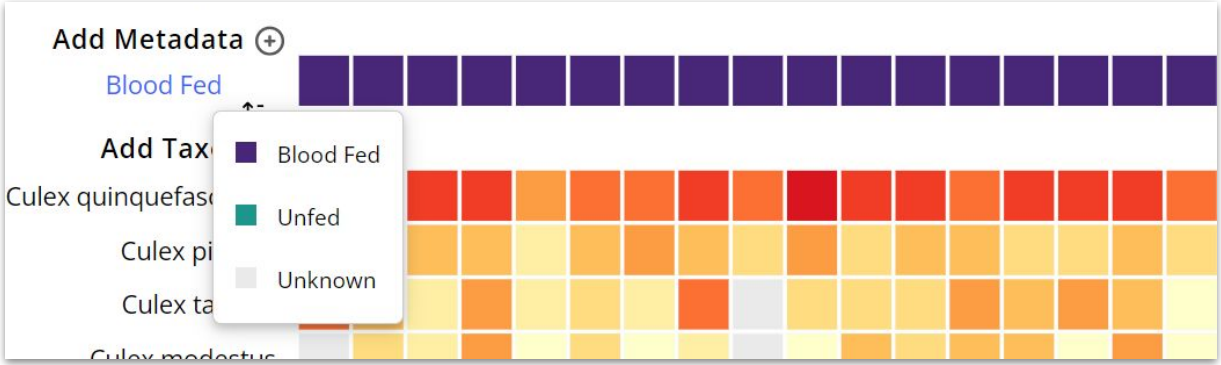
Las muestras se organizan automáticamente mediante agrupamiento jerárquico.



Agrupamientos de taxones en el dendrograma: los taxones que tienen más probabilidades de aparecer juntos en abundancias similares se agrupan

Muestras de grupos de dendrogramas: las muestras con taxones similares en abundancias similares se agrupan

Sort samples by metadata



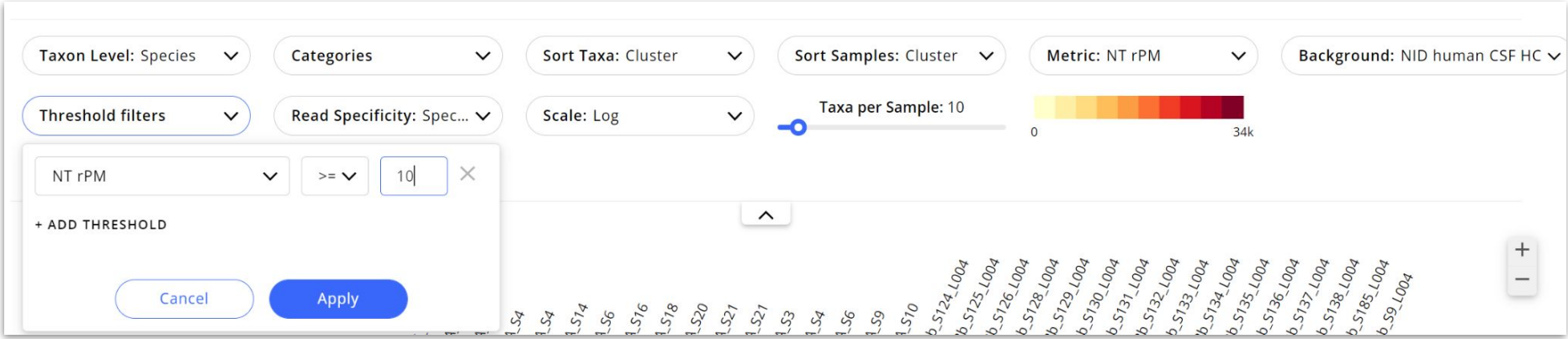
Es posible que desee organizar sus muestras por diferentes campos de metadatos en lugar de agruparlas jerárquicamente.

Puede agregar metadatos al mapa de calor y ordenarlos. En este ejemplo, compararemos los taxones presentes en mosquitos alimentados con sangre y mosquitos no alimentados.

Aplicar filtros

Al igual que en el informe de una muestra, deberá aplicar filtros para eliminar los resultados falsos.

También puede comparar los controles de agua con las muestras para ver los contaminantes antes de agregar un filtro de puntuación **Z-score filter.**



¿Qué pregunta puedo responder con este diseño experimental??

A. ¿Qué patógenos transmitidos por vectores están circulando en el ecosistema?

B. ¿Hay virus nuevos presentes y cuáles son?

C. Todas las anteriores

D. Ninguna de las anteriores

Actividad: Análisis de Múltiples Muestras

Culex erythrothorax

- ❏ 33 individuos
- ❏ Secuenciación: Illumina
- ❏ 17 alimentados con sangre
- ❏ 2 controles de agua

¡Hagamos un análisis de múltiples muestras utilizando un Heatmap!

⏳ Tiempo: ~ 30 minutos para la actividad y 10 minutos para la discusión

ENLACE GITHUB DONDE ESTÁ EL ARCHIVO.

Publicaciones Relevantes

- Single mosquito metatranscriptomics identifies vectors, emerging pathogens and reservoirs in one assay
- IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring
- Discovering disease-causing pathogens in resource-scarce Southeast Asia using a global metagenomic pathogen monitoring system
- Case Report: Cambodian National Malaria Surveillance Program Detection of Plasmodium knowlesi
- The Perfect Storm of 2019: An immunological and phylodynamic analysis of Cambodia's unprecedented dengue outbreak
- Metagenomic Pathogen Sequencing in Resource-Scarce Settings: Lessons Learned and the Road Ahead
- Pathogen genomics in public health
- IDseq—An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring
- Pathogen genomics in public health
- Unbiased metagenomic sequencing for pediatric meningitis in Bangladesh reveals neuroinvasive Chikungunya virus outbreak and other unrealized pathogens

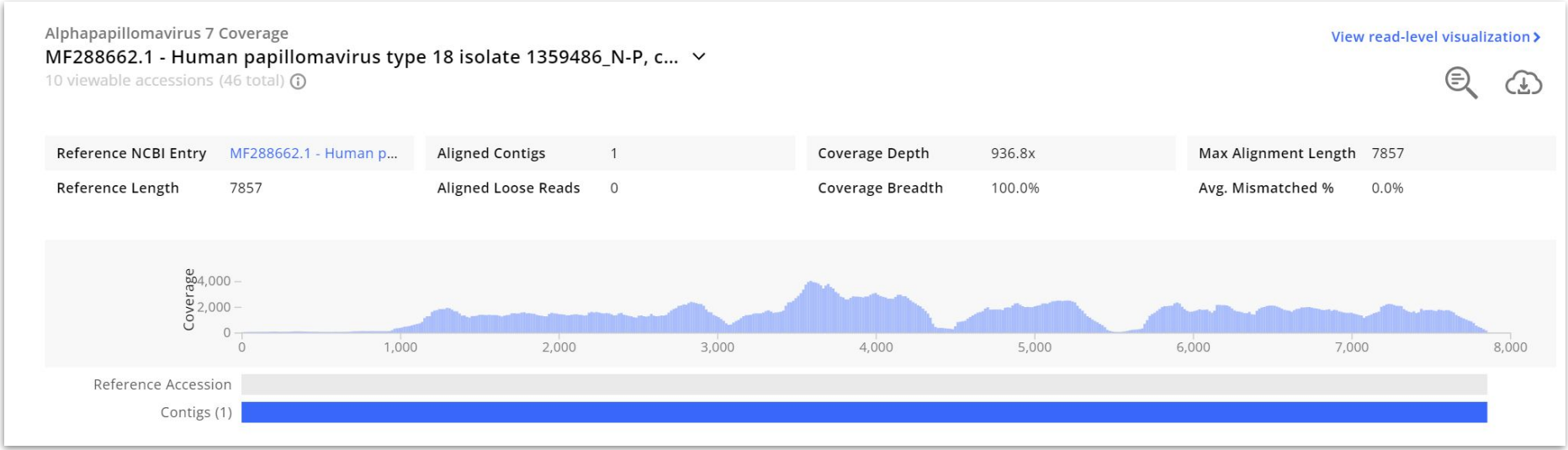
Gracias

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

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

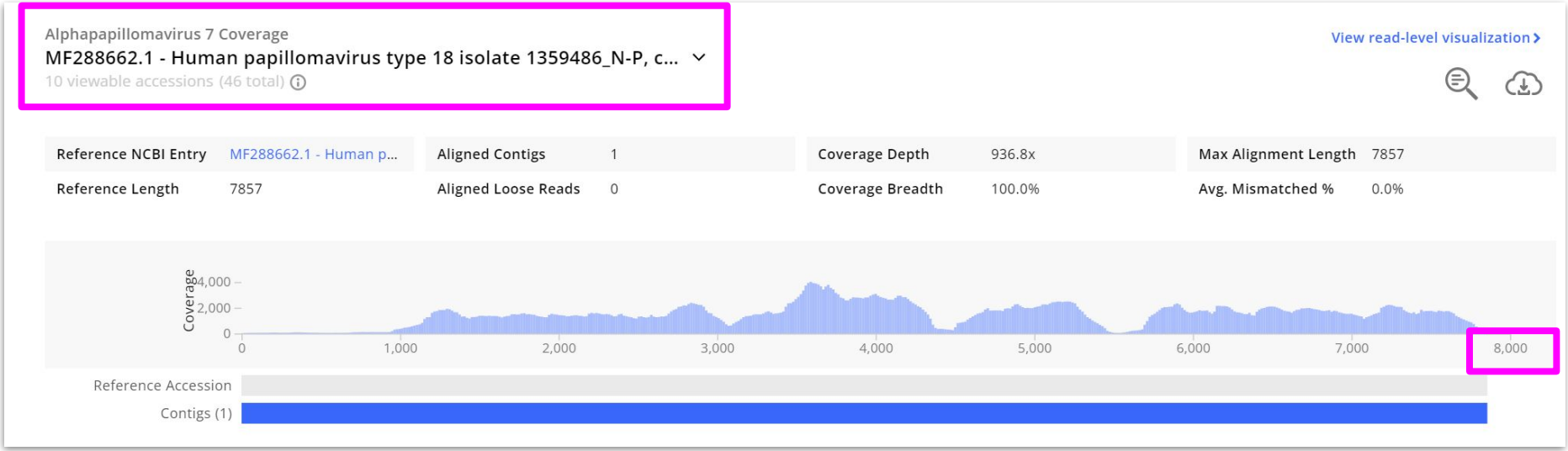
https://chanzuckerberg.zendesk.com/hc/en-us

The coverage visualization is valuable in calling hits



The coverage visualization is valuable in calling hits

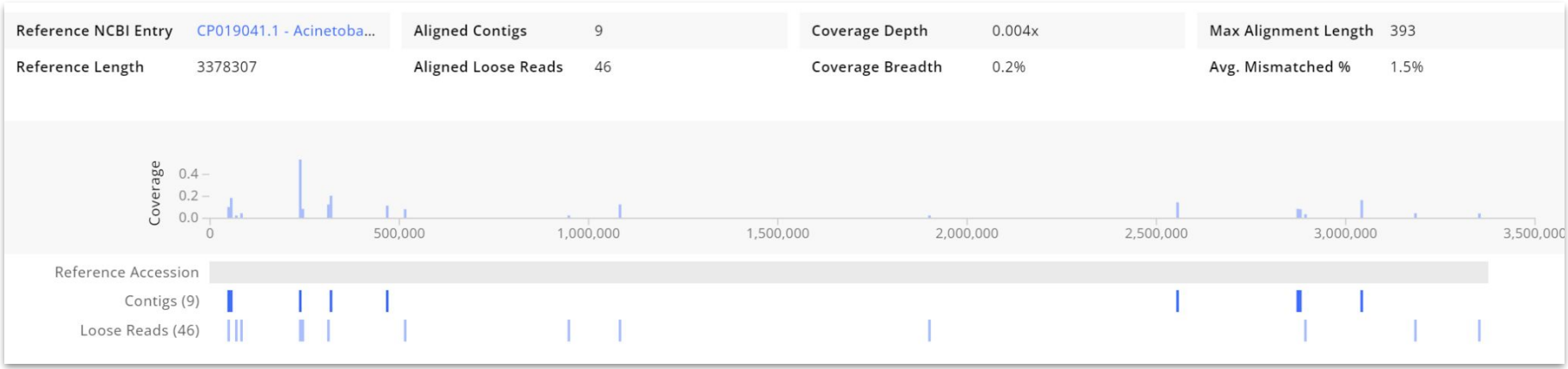
View the accession



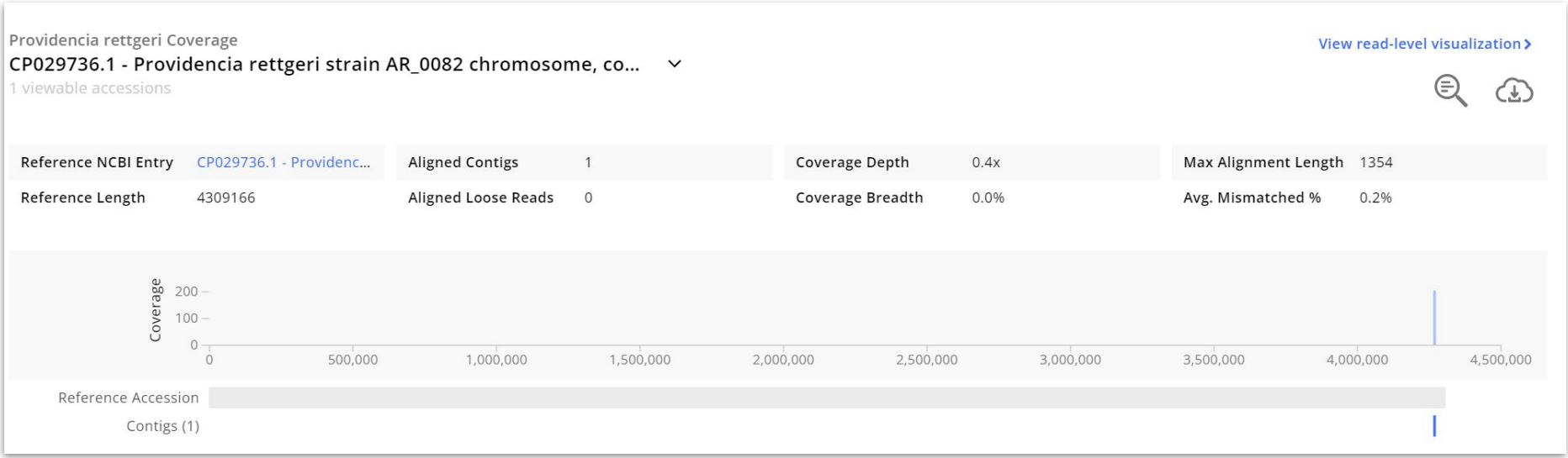
Pay attention to accession length

Take into consideration the depth and breadth of the coverage

- This is great coverage of bacterium
- Typically see less coverage, you will look for multiple peaks across the genome



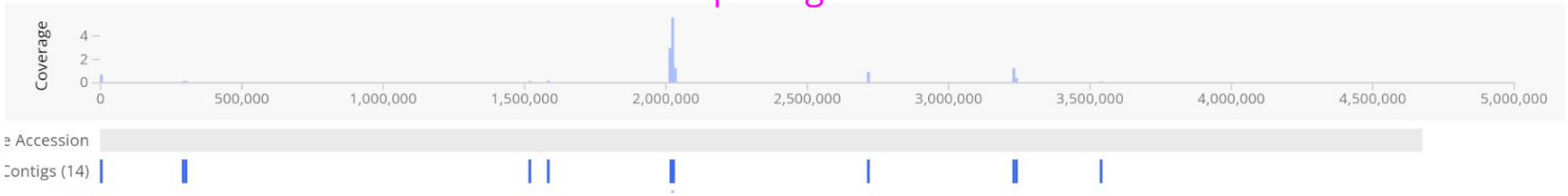
This is an example of poor coverage. I would not be comfortable calling this hit real



Which would you not consider a true hit?

A

E coli - complete genome



B

Chikungunya virus- complete genome

C

Cryptococcus neoformans