**mNGS: patógenos transmitidos por un vector**

🧪**Actividad:** Análisis de una muestra

Se analizarán datos de secuenciación de illumina de un solo mosquito hembra (*Culex erythrothorax)*. Este mosquito se alimentó de sangre. ¡Veamos qué taxones podemos encontrar!

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

1. Inicien sesión en [**CZ ID**](http://czid.org)

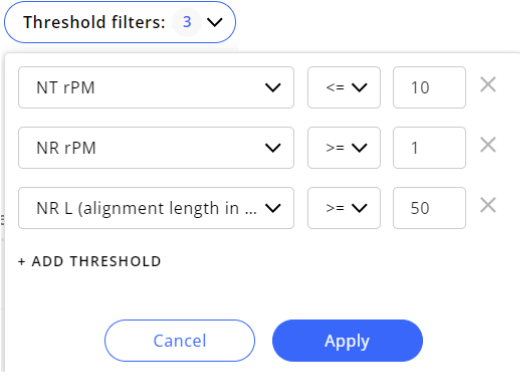
**Username:** correo

**Password:** contraseña

1. En el repositorio público de CZID buscar la muestra : [**CMS\_001\_RNA\_A\_S1**](https://czid.org/djh05)
2. Para asegurarnos de que todos los hits que vemos tengan asignaciones completas de linaje seleccione en la categoria *Read Specificity “specific Only ”*. **¿Cuántas filas de taxones están presentes?**

Dado que estamos comparando contra las bases de datos de nucleótidos y proteínas en NCBI. A menudo hay hits no confiables. Para ayudar a mitigar esto, aplicaremos filtros.

1. Apliquen los siguientes filtros para eliminar hits no confiables:



**NT rPM ≥ 10** [Solo quiero ver taxones presentes con al menos 10 lecturas por millón alineadas a la base de datos de nucleótidos].

**NR rPM ≥ 1** [Solo quiero ver taxones presentes con al menos 1 lectura por millón alineada a la base de datos de proteínas].

**NT L ≥ 50** [Solo quiero ver taxones presentes con al menos 50 bp de homología alineada a la base de datos de nucleótidos].

1. **¿Cuántas filas de taxones están presentes ahora?**
2. Para eliminar la contaminación aplique un modelo de *Background “*PHA4GE 2023”. Fíjese que al añadir el modelo ya el puntaje Z deberían ser calculados. Ahora, añada en los filtros *Z score filter.*

**NT Z-score ≥ 1** [Solo quiero ver los taxones presentes con al menos una desviación estándar mayor que el promedio de rPM alineando con la base de datos de nucleótidos en mi modelo de fondo seleccionado].

1. Ordene las columnas de rPM haciendo clic en el símbolo ^ junto al encabezado de la columna para que los taxones con más rPM se muestren en la parte superior. **¿Cuál es el taxón más abundante según las lecturas por millón (rPM)?**
2. Busque algún taxón interesante, explore su cobertura y con base en los datos mostrados añada dos observaciones.

## **¿Qué pasa con los virus divergentes?**

Los virus divergentes a menudo no tienen coincidencias en la base de datos de nucleótidos y solo muestran homología a nivel de proteínas, ya que las mutaciones se acumulan más rápido a nivel de nucleótidos.

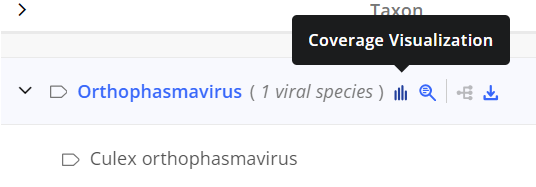
1. Elimine los filtros haciendo clic en "*Clear Filters*" (borrar filtros).
2. Cambie la categoría a virus y deseleccione *"phage"* (los fagos son virus que infectan bacterias y no infectan eucariotas).

Las métricas a las que debes prestar atención para identificar un virus divergente o novedoso son diferentes. Algunas métricas importantes son el número de lecturas que se alinean con la base de datos de proteínas (puede que solo veas unas pocas o incluso 0 lecturas alineándose con la base de datos de nucleótidos) y el porcentaje de identidad. El valor de corte para el porcentaje de identidad.

Añada los siguientes filtros:  
**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 ve? ¿Cuáles son?**

Ahora, confirme que estos son hits virales. Podemos hacer esto usando BLASTX. BLASTX traduce la secuencia de nucleótidos a una secuencia de proteínas y luego la compara con la base de datos de proteínas. Dado que no hay coincidencias a nivel de nucleótidos, no queremos usar BLASTN (comparación de secuencia de nucleótidos a nucleótidos).

Abre el género **orthophasmavirus**. Ejecuta BLASTX a nivel de especie.

1. Seleccione BLASTX > continuar > seleccionar el contig > continuar > ver informe. Qué puede observar?