**Análisis de calidad de secuencias**

**Actividad:**

1. Inicie sesión en la plataforma CZ-ID y diríjase al set de datos publico [Fever Project](https://czid.org/public?currentDisplay=table&currentTab=samples&mapSidebarTab=samples&projectId=15628&showFilters=true&showStats=true&updatedAt=2024-09-23T19%3A18%3A05.681Z&workflow=short-read-mngs), subido por Zoumana Isaac Traore en septiembre de 2024 y que tiene como objetivo tiene como objetivo descubrir el agente etiológico de la enfermedad febril aguda en Guinea.

CZ ID ofrece una primera visualización de dos parámetros “Passed filters” y “passed QC” identifique ubicando el cursor sobre cada título a que hacen referencia estos dos parámetros.

¿Cuál es la secuencia con mejor calidad, después de eliminar bases de baja calidad y lecturas cortas?

1. CZ ID ofrece una página de visualización de control de calidad a nivel de proyecto (PLQC) en la que puede ver varias métricas y evaluar preguntas como:

¿Hay suficientes lecturas totales?

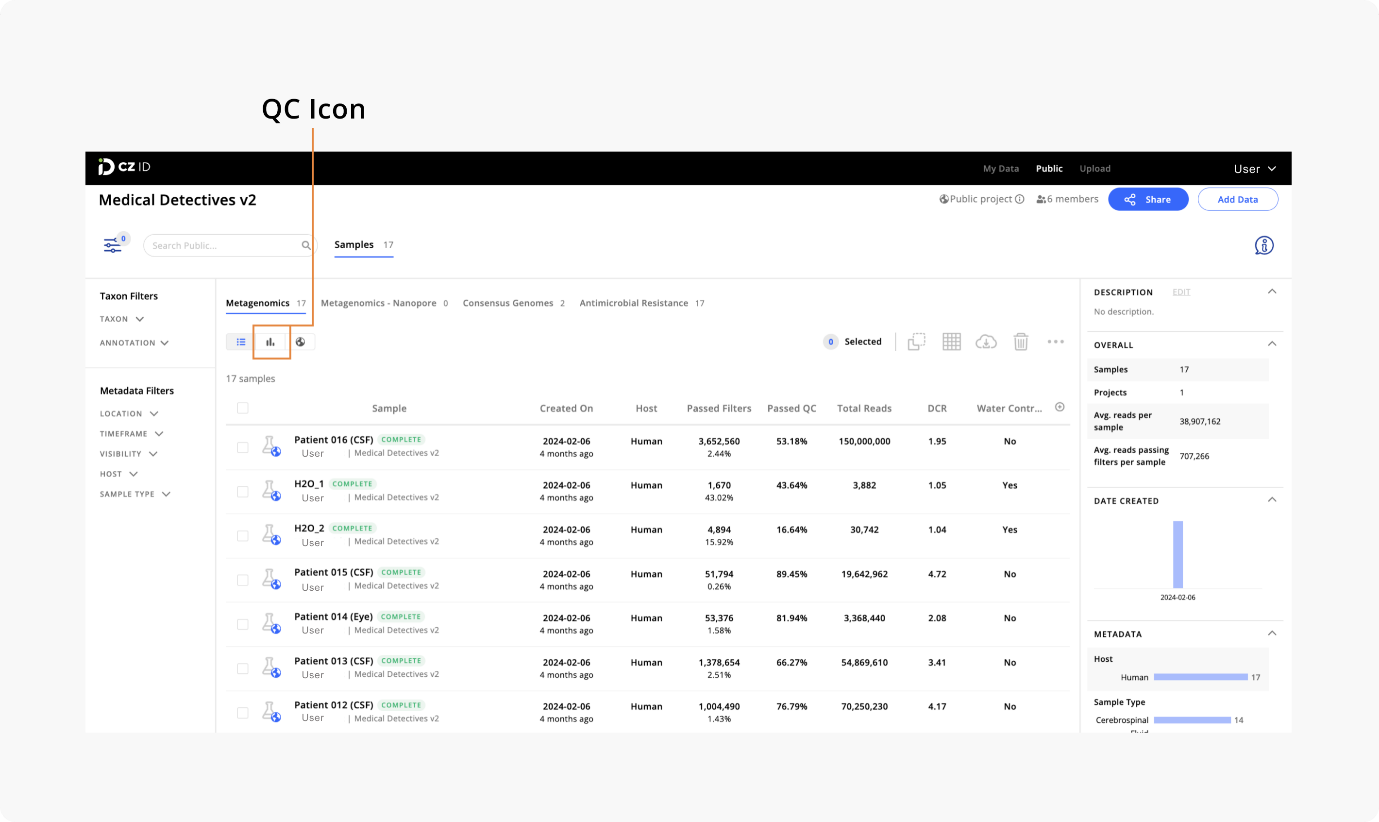
¿Mis muestras tienen suficientes lecturas de alta calidad?

¿Mis muestras tienen suficiente diversidad de secuenciación?

¿Mis muestras tienen longitudes de inserción suficientes?

¿Cómo se procesaron mis muestras a través del pipeline?

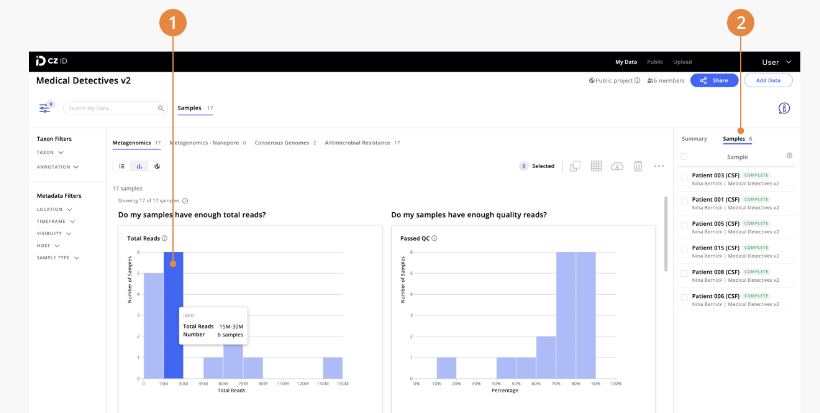
Para navegar a la página de visualización de PLQC, haga clic en el **ícono de control de calidad** (gráfico de barras) debajo de la pestaña "Metagenómica".



Los gráficos incluyen histogramas para lecturas totales, control de calidad aprobado, relación de compresión de duplicados (DCR) y tamaño de inserción medio. Además, verá un gráfico de barras apiladas para lecturas perdidas que muestra las lecturas filtradas a través de cada paso del proceso. Tenga en cuenta que cada gráfico es interactivo y podrá ver detalles sobre las muestras incluidas en cada barra al pasar el cursor sobre una barra determinada o hacer clic en ella.

**Para ver muestras representadas dentro de cada barra del histograma:**

1. Haga clic en la barra de su interés
2. Vea los detalles de muestra para el rango seleccionado en el panel de la derecha.



**¿Hay suficientes lecturas totales?**

El histograma de lecturas totales muestra la distribución de las lecturas totales en todas las muestras. Puede hacer clic en cada barra del histograma para ver las muestras asociadas con un rango determinado de cantidad total de lecturas.

Una forma de identificar rápidamente un problema es buscar valores atípicos. Si una muestra tiene muchas menos lecturas que las otras muestras, es posible que esa muestra haya experimentado un problema en la etapa de agrupamiento.

**Control de calidad**

La métrica de control de calidad aprobado representa el porcentaje de lecturas restantes después del filtrado de control de calidad utilizando fastp para eliminar bases de baja calidad, lecturas cortas (<35 pb) y lecturas de baja complejidad.

**¿Mis muestras tienen suficientes lecturas de alta calidad?**

El histograma de control de calidad aprobado (Passed QC) muestra la distribución de lecturas que pasaron el control de calidad en todas las muestras.

La métrica de control de calidad aprobado se refiere a la cantidad de lecturas que han pasado los filtros de control de calidad. Durante la secuenciación, se asignan puntajes de calidad de Phred (a cada base para indicar la probabilidad de un error de secuenciación. Cuanto mayor sea el puntaje Q, más confiable será la determinación de la base.

**Relación de compresión duplicada**

La relación de compresión de duplicados (DCR) es la relación entre la cantidad de lecturas que pasan el filtrado de control de calidad y la eliminación de lecturas del host/humano y la cantidad de lecturas únicas después de la eliminación de duplicados.

**¿Hay demasiadas lecturas duplicadas en mi biblioteca?**

La relación de compresión de duplicados (DCR) es un indicador de la diversidad de secuencias en su muestra. Si la muestra contiene muchas lecturas duplicadas, esto da como resultado una biblioteca menos diversa, lo que posiblemente indique un sesgo debido a la amplificación o la secuenciación. Las secuencias duplicadas podrían deberse a un enriquecimiento de PCR sesgado o podrían ser realmente un fenómeno biológico. Los métodos de laboratorio húmedo influirán en la DCR. Por ejemplo, si las muestras se prepararon con un método de enriquecimiento como MSSPE , la DCR sería alta porque se enriquecen secuencias específicas. Si se realiza metagenómica sin enriquecimiento, un valor de DCR inferior a 2 es ideal.

**¿Mis muestras tienen longitudes de inserción suficientes?**

El tamaño medio de inserción se puede utilizar como indicador de la calidad del ácido nucleico en la biblioteca final. Los tamaños de fragmentos cortos pueden indicar degradación de la muestra o sobrefragmentación durante la preparación de la biblioteca. Tenga en cuenta que el valor solo se calculará para muestras recolectadas de huéspedes humanos. Por lo tanto, el gráfico de tamaño medio de inserción estará en blanco para muestras de huéspedes no humanos.

En nuestro set de datos de estudio, aparecen 17 secuencias en el grafico de tamaño medio de inserción, a pasar de que ninguna de ellas fue tomada en huéspedes humano ¿A que cree que se pueda deber esto?

**¿Cómo se procesaron mis muestras a través del pipeline?**

El gráfico de lecturas perdidas muestra en qué etapa del proceso se filtran las lecturas. La parte del gráfico que indica “lecturas restantes” indica la cantidad de lecturas que pasaron los pasos de filtrado del host y control de calidad y pasaron a la parte de análisis e identificación de patógenos del proceso. El valor de “lecturas restantes” dependerá de numerosas variables, entre las que se incluyen (sin limitarse a ellas):

* **Tipo de material de muestra** : por ejemplo, las muestras de líquido cefalorraquídeo (LCR) contendrán más lecturas del huésped que una muestra fecal. Por lo tanto, las muestras de LCR darán como resultado menos lecturas restantes después del filtrado del huésped en comparación con las muestras fecales. Las muestras de LCR a menudo tienen > 99% de lecturas eliminadas durante los pasos de filtrado del huésped, ya que se espera que la fracción microbiana sea mínima en comparación con la fracción humana. Las muestras de heces, por otro lado, pueden tener un porcentaje mucho menor de lecturas eliminadas durante los pasos de filtrado del huésped, ya que se espera que la mayor parte del material genético sea microbiano en lugar de derivado del huésped. Recuerde que, independientemente del huésped seleccionado, las lecturas humanas se filtran.
* **Si se filtraron o no las lecturas del anfitrión**: si el anfitrión no tenía un genoma secuenciado y se seleccionó “Solo ERCC” para el genoma del anfitrión al cargar la muestra en CZ ID, no se perderá ninguna lectura durante el filtrado del anfitrión. Si no se filtra el anfitrión, quedará una mayor cantidad de lecturas.
* **Calidad de lectura**: si se filtraran muchas de las lecturas debido a su baja calidad, quedaría una menor cantidad de lecturas restantes.
* **Condiciones de almacenamiento de la muestra**: si la muestra no se almacenó correctamente, el ácido nucleico puede haberse degradado, lo que dio lugar a muchas secuencias cortas que pueden filtrarse durante el recorte del adaptador. Esto reduciría la cantidad de lecturas que pasan el filtrado de control de calidad.
* **Nivel de duplicación**: si se marcan muchas lecturas como duplicadas durante el paso de "identificación de duplicados", el DCR será alto. Los niveles altos de duplicación pueden indicar que hubo una gran cantidad de amplificación por PCR durante las etapas de preparación de la biblioteca, lo que dio como resultado muchas secuencias duplicadas. Tenga en cuenta que CZ ID no está diseñado para usarse con bibliotecas de amplicones (por ejemplo, secuenciación 16S).
* **Complejidad de lectura**: si se filtran muchas lecturas durante el filtrado de control de calidad, esto indicaría que las lecturas de secuenciación tenían una complejidad relativamente baja (por ejemplo, contienen una alta proporción de nucleótidos de homopolímero o repeticiones de secuencias simples).

Tomando en consideración el apartado de metadatos (costado izquierdo de la interfaz gráfica) y los datos de calidad analizados anteriormente ¿cree usted que existe relación entre el hospedero o tipo de muestra con la calidad de las secuencias?