1. Los artículos elegidos son los siguientes.
   - Ancient human genome sequence of an extinct Palaeo-Eskimo.
   - Population genetics of fruit bat reservoir informs the dynamics, distribution and diversity of Nipah virus
   - Analysis of changes in microbiome compositions related to the prognosis of colorectal cancer patients based on tissue-derived 16S rRNA sequences
   - Taxonomic and functional analyses of intact microbial communities thriving in extreme, astrobiology-relevant, anoxic sites.
   - Novel bacterial taxa in a minimal lignocellulolytic consortium and their potential for lignin and plastics transformation.
   - Phylogenomic analysis of the complete sequence of a gastroenteritis-associated cetacean adenovirus (bottlenose dolphin adenovirus 1) reveals a high degree of genetic divergence

2. Las tablas son las siguientes.
   `https://docs.google.com/spreadsheets/d/1tWvlbX8Hb2cHxauwjZcS3kPFxVR2dw-h/edit?usp=share_link&ouid=107305295184100115600&rtpof=true&sd=true`

3. El artículo elegido es *Ancient human genome sequence of an extinct Palaeo-Eskimo.*

   a. De manera general, ¿De qué trata éste artículo?
   Solución. De conocer más acerca del humano antiguo, en este caso de los humanos que existieron hace aproximadamente 4000 años en el noreste de Asia.

   b. ¿Cuáles son los objetivos de la investigación?
   Solución. Conocer más acerca de las personas antiguas y especificar el primer genoma humano antiguo.

   c. ¿Qué tipo de estudio es? (Metagenoma, genómico o metabarcoding)¿Por qué?
   Solución. Se trata de un estudio genómico ya que se estudia el conjunto completo del DNA de las personas antiguas para conocer más acerca de ellos.

   d. Busca los archivos crudos que se utilizaron en la investigación en una base de datos, y coloca el link.
   Solución. Los links relacionados a la investigación son los siguientes.
      - `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3951495/bin/NIHMS467171-supplement-SupportInfo_.pdf`
      - `https://gold.jgi.doe.gov/projects?id=Gp0008201`
      - `https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=46213`

   e. Investiga: ¿Cómo puedes saber el número de secuencias que hay en un archivo *FASTQ*?
   Solución. Hay que tomar en cuenta que en los ficheros *FASTQ* habitualmente existen cuatro líneas por secuencia en donde.
      - La primera línea comienza con un @(arroba) e incluye el nombre de la secuencia.
      - La segunda es la secuencia de nucleótidos.

- La tercera es un símbolo de +.
- La última es la codificación de la calidad en código ASCII.

Para saber el número de secuencias podríamos contar cada arroba del archivo con ayuda de un programa implementando un regex o simplemente utilizar BioPython el cual es capaz de leer y escribir ficheros *FASTQ*.

4. Reporte de control de calidad.

   - Secuencias de algún artículo del apartado anterior.
   - Secuencias otorgadas en Drive.

   Los reportes de control de calidad en formato HTML están en sus respectivas carpetas dentro del directorio.

5. ¿Qué gráficos/módulos del reporte arrojado por *FASTQC* consideras que son los más importantes? Elige 4 y justifica.
   Solución. Considero que los siguientes son los más importantes.

   - *Basic Statistics* ya que proporciona la información básica de todo el archivo y nos da un gran panorama informativo.
   - *Per base sequence quality* debido a que nos brinda la calidad de cada base de todas las secuencias del archivo además de proporcionar una buena visualización de las secuencias con respecto a su calidad.
   - *Per sequence quality scores* porque nos brinda la calidad de todas las secuencias además de darnos el promedio de calidad de las secuencias del archivo.
   - *Sequence Length Distribution* ya que nos muestra la distribución de los tamaños de las secuencias del archivo lo que nos da una idea del tamaño de pares de bases las secuencias.

6. ¿Qué concluyes de las secuencias de cada muestra? Escribe una conclusión para las secuencias que tú elegiste y paras las secuencias que te otorgué como problema.
   Solución.

   - Secuencias elegidas.
     - *SRR031056* En general las secuencias se mantienen dentro de los valores deseados con algunas lecturas identificadas, sin embargo, la mayoría de secuencias no se lograron leer. Notamos que en cuanto al contenido GC éste se mantiene algo alejado en su pico máximo respecto a las lecturas con la distribución teórica, también notamos que el tamaño promedio de secuencias es de 75 pb.
     - *SRR031057* De forma general observamos que las secuencias leídas se mantienen de igual forma, dentro de los valores deseados. Igualmente tenemos aproximadamente sólo el 39% de secuencias identificadas, lo que nos provoca un contenido GC un poco alejado del promedio cuyo tamaño es de apoximadamente 75 pb.
   - Secuencias otorgadas.
     - *SRR8082143_1* Las secuencias de este archivo me parecen que tienen buena calidad gracias a los valores obtenidos en la mayoría de las gráficas, notamos que que la calidad de secuenciación por base es buena, manteniéndose dentro de los valores cualitativos; la calidad por secuencia de igual forma se mantiene en los valores deseados; y en general no hay mayor problema con respecto a la calidad que se encuentra en las secuencias del archivo.

- *SRR8082143_2* Considero que las secuencias del archivo son buenas, tampoco excelentes pero definitivamente no son de mala calidad, observamos que los valores de la secuenciación por base se mantienen dentro de lo deseado a excepción de los pares de bases por arriba de 200; la calidad por secuencia es buena, estando dentro de los valores; en general existe una buena calidad de las secuencias de menos de 200 pares de bases del archivo.

7. ¿Por qué es importante conocer el contenido de GC en un genoma? Adjunta una referencia para ésta respuesta.
   Solución. Es importante ya que se puede valorar la calidad de la secuenciación de la hebra de DNA, gracias a que este valor es relativamente constante en cada especie.

   Referencia. Contenido GC. (s/f). Google Arts & Culture. Recuperado el 15 de noviembre de 2022, de https://artsandculture.google.com/entity/m0426v8?hl=es