

Bioinformática y análisis de datos ómicos

© Ignacio Varela Egocheaga

Este material se publica bajo licencia Creative Commons CC BY-NC-SA 4.0



EJERCICIO TEMA 6

Alumno:

Contesta a las preguntas de abajo, pegando los comandos que has necesitado ejecutar en unix sobre los archivos del siguiente enlace Tema6. Finalmente, convierte este documento a PDF y entrégalo a través del moodle.

1.- De acuerdo a los archivos en formato fastq descargados, ¿cuántas secuencias se han generado en la reacción de secuenciación masiva? Razona tu respuesta.

2.- ¿Qué plataforma de secuenciación (Illumina, LifeTechnologies, Nanopore, etc.) crees que se ha utilizado para generar estos datos? Razona tu respuesta.

3.- El protocolo de secuenciación en un equipo NGS se denomina con un número para indicar el tamaño de las secuencias generadas y la abreviatura SE (para lecturas no pareadas) o PE (para lecturas pareadas). Por ejemplo, un protocolo de secuenciación 75 PE indica que se han generado secuencias pareadas de 75 nucleótidos. ¿Qué tipo de protocolo de secuenciación se ha realizado para generar los datos descargados?

4.- Utiliza el programa de análisis de calidad fastqc para analizar la calidad de las secuencias en el archivo fastq. ¿Es la calidad de secuencia comparable a lo largo de todos los ciclos de la carrera? ¿Sería conveniente acortar las secuencias para evitar errores en el análisis? Si lo consideras adecuado ¿a partir de que tamaño cortarías? Razona tu respuesta e ilústrala con algún gráfico extraído del informe del control de calidad.

5.- ¿La calidad de secuencia está expresada en escala Phred - 32 o Phred - 64? Razona tu respuesta.

6.- Indica el undécimo nucleótido que se ha leído en la read 2 del cluster @A00496:62:H3GVJDRX2:1:2101:5050:36385. ¿Cual es su calidad en formato numérico?