RESPUESTA CORRECTAS SUPUESTO PRÁCTICO

Las funciones del puesto **OPL3BINF** se centran en el análisis de datos biológicos utilizando herramientas bioinformáticas, por lo que para su desarrollo es necesario tener conocimientos en el uso de supercomputación y análisis de datos. A cada persona aspirante se le proporciona un *usuario* y una *clave* que dan acceso al ordenador en el cual se realizará el ejercicio práctico (host: *picasso.scbi.uma.es*).

Planteamos la siguiente situación:

"Un grupo de investigación/empresa innovadora contacta con el Servicio de Bioinformática para **identificar filogenéticamente** una muestra biológica problema. El proceso de análisis parte de los datos de secuenciación automática obtenidos de la misma y almacenados en el Supercomputador Picasso, en la carpeta **"datos"** de su *usuario*. Para facilitar los análisis utilizaremos las siguientes **herramientas**: FastQC, Fastp, Prokka, Megahit, Eggnog."

En este ejercicio solicitamos a los aspirantes:

1.- Inspección de datos (12 puntos)

1.1- Dentro del directorio **"datos/ejercicio1"** se encuentran los resultados de secuenciación de la muestra problema. ¿Cuántas lecturas contienen cada uno de los ficheros? Inspeccione la calidad de los datos de secuenciación utilizando el programa correspondiente de la **lista de herramientas** arriba indicada, y guarde la salida resultante en una carpeta denominada *resultado_1_1*. Escriba todos los comandos utilizados para resolver estas preguntas y los resultados obtenidos. (Valor de la pregunta: 6 puntos).

387.568 lecturas cada uno de ellos.
module load fastqc
mkdir resultado_1_1
fastqc ERR486840_1.fastq.gz ERR486840_2.fastq.gz -o resultado_1_1

1.2.- Una vez inspeccionados la calidad de las lecturas, éstas deben ser sometidas a un preprocesamiento asumiendo qué únicamente filtraremos por calidad de secuenciación. ¿Qué herramienta **de la lista de herramientas** utilizaría? Escriba todos los comandos de utilizados y denomine a los resultados finales **_1_clean* y **_2_clean**. ¿Cuál es el número de lecturas finales? (Valor de la pregunta: 6 puntos).

module load fastp
fastp -i ERR486840_1.fastq.gz -o file_1_clean.fastq.gz -I ERR486840_2.fastq.gz -O file_2_clean.fastq.gz
387.333 lecturas cada uno de los ficheros

2.- Procesamiento de datos de secuenciación (13 puntos)

2.1.- Partiendo de una serie de ficheros de secuenciación procesados almacenados en el directorio "**datos/ejercicio2.1"**, se requiere realizar un ensamblaje de lecturas. ¿Qué programa **de la lista de herramientas** utilizaría? Utilice los comandos por defecto. ¿Cuántos contigs genera dicho ensamblaje? (Valor de la pregunta: 5 puntos).

module load megahit
megahit -1 file_1_clean.fastq.gz -2 file_2_clean.fastq.gz
22 contigs

2.2.- Realice una anotación funcional del fichero fasta de aminoácidos almacenado en el directorio **datos/ejercicio2.2**. ¿Qué software **de la lista de herramientas** utilizaría? Escriba el comando de uso. Cree un file.sh, utilizando de base el fichero *ejemplo_simple.sh* de ~/ejemplo, y ejecute este trabajo por sistema de cola utilizando 32 CPU. Escriba los comandos de uso. Escriba cuál es la especie filogenética problema de nuestra muestra. (Valor de la pregunta: 8 puntos).

module load eggnog-mapper
emapper.py -i PROKKA_12112023.faa --cpu 32 -o result_eggog
Tenericutes