

Apellido, Nombre: _____,

Fecha: ____ / 05 / 2023



Para la resolución de este examen, utilice siempre GitBash o algún otro terminal adecuado. Realice el control de cambios cada vez que responda una pregunta.

Realice la clonación del directorio de GitHub. Realice un fork (bifurcación) del repositorio https://github.com/gualapuemoises/2023l_gbi6_exam1.git. Luego en su PC/laptop vaya a un directorio de su preferencia y clone el directorio “bifurcado”. Ingrese al directorio *exam1*. Nunca ingrese a ningún subdirectorio, todo debe hacer desde esta carpeta *exam1*.

Genere scripts tipo *.sh*, para realizar lo siguiente:

1. **[10 puntos]** La secuenciación de células individuales (single cell sequencing) es un método de secuenciación de próxima generación que permite obtener genomas o transcriptomas de células individuales y permite tener una visión de alta resolución de las variaciones celulares. En la carpeta *data* tiene la subcarpeta *singlecell* con muchas subcarpetas y archivos. Cree **singlecell.sh** y realice la búsqueda de todos los tipos de extensión de los archivos en todas las subcarpetas (.csv, .txt, .tsv, .gz, .png, otros). Guarde ese conteo en *extnsiones.txt*.
2. **[10 puntos]** Se ha realizado un estudio de [redes de interacción de proteínas de mosquitos, virus del dengue y humanos](#) para identificar las proteínas de interés para controlar la infección. La data de este estudio se encuentra en la carpeta *data* se tiene la subcarpeta *denvint*. Cree **denvint.sh** y realice el conteo del número de filas y columnas de los cuatro archivos .csv utilizando un bucle for.
3. **[10 puntos]** Los micro RNAs (miRNA) son secuencias cortas no codificantes de RNA que están implicadas en el silenciamiento del RNA y la regulación post-transcripcional, por lo tanto juegan un rol fundamental en las funciones celulares. En la carpeta *data* se tiene el archivo *miRNA.dat* de la base de datos [miRBase](#). En ella se tiene todos los registros de publicaciones asociados a miRNAs. Cree **mirbase.sh** para contar cuantos artículos se tiene a la fecha (busque la palabra PUBMED). En el mismo archivo .sh, indique cuántos estudios de micro RNA se han publicado en la revista Nature y cuántos de ellos fueron el organismo *C. elegans* (busque la palabra elegans). Finalmente indique cuántos micro RNA estudiados tiene una longitud 139 de pares de bases (busque 139 BP).
4. **[10 puntos]** Recientemente se ha generado un método basado en aprendizaje de máquina llamado MAHOMES para predecir si los sitios de unión de metales de las proteínas tienen actividad enzimática. En la carpeta *data* se tiene la subcarpeta *mahomes* y en ella el archivo *sites.csv*. Cree **mahomes.sh** y realice la separación de las columnas “resName1”, “PDB Classification” y “Uniprot Acc”, estas columnas guarde en el archivo *pdb.csv*. Calcule cuántos tipos únicos de elementos hay en cada una de las columnas extraídas y guarde ese conteo en *pdb_count.csv*.
5. **[10 puntos]** La metagenómica pretende identificar y clasificar la secuencia no específica (todos los organismos) de una muestra. Este tipo de estudio ha permitido conocer sobre los grupos de microbiota de cualquier sitio que uno se pueda imaginar y comprender su influencia en estos sistemas (por ejemplo la microbiota intestinal y su influencia en la salud). En la carpeta *data* se encuentra la subcarpeta *metagen*. Cree **metagen.sh** para analizar cuántos registros del archivo “*infants_metagenome.txt*” tienen registro hasta el nivel de especies. Luego vamos a suponer que se tiene fragmentos de interés como “TATA”, “GAGA” o “GATA” que son sitios de unión de alguna enzima; realice la búsqueda de esos fragmentos y guárdelos en el *gata.txt*.