**Nombre y Apellido: Juan Carlos García Estupiñán**

**Actividad 2.- Manipulación y formateo de archivos: Formato FASTQ y FASTA**

**Objetivo**

El propósito de esta actividad es desarrollar un flujo de trabajo completo (denominado en inglés *pipeline*) bioinformático que nos permita procesar una serie de datos biológicos. Con esta actividad, se pretende que el estudiante adquiera destrezas para interactuar con el sistema operativo a través de la línea de comandos y que sea capaz de desarrollar Shell scripts propios para resolver diferentes retos bioinformáticos focalizados en dos tipos de formato de texto, el formato FASTQ y en formato FASTA.

**Parte I: Creación del directorio de trabajo.**

Para inicializar este ejercicio, deberán crear y organizar un nuevo directorio de trabajo que contenga los puntos más importantes que vamos a estudiar. La idea de esta organización es que alguien que no esté familiarizado con su proyecto debería poder mirar los archivos almacenados en el ordenador y comprender en detalle lo que hicimos y por qué lo hicimos. Este «alguien» podría ser cualquiera de una variedad de personas: alguien que leyó el artículo que hemos publicado y quiere intentar reproducir nuestro trabajo, un colaborador que quiere comprender los detalles de los experimentos realizados, un futuro estudiante que trabaja con ustedes en el laboratorio y deseamos ampliar el trabajo hecho hasta el momento, un asesor de investigación que puede estar interesado en comprender nuestro trabajo o que puede estar evaluando nuestras habilidades de investigación. Sin embargo, lo más común es que ese "alguien" seamos nosotros mismos.

Usando comandos de Linux a través de un terminal, deberá crear la siguiente estructura de directorios para este proyecto que identificará de la siguiente manera:

***User\_proyect\_year\_publication*** donde:

* User corresponde a su apellido.
* project corresponde al nombre de su proyecto hipotético en el cual se encuentra trabajando (por ejemplo: humanTonsils)
* year es el año de la publicación o investigación que está realizando.
* Publication corresponde a una revista o conferencia en donde quiere publicar sus resultados.

Este directorio, en mi caso, ***Soler\_humanTonsils\_2022\_Nature*** será creado dentro del directorio ***Documents*** que a su vez contendrá 5 directorios (***analysis****,* ***code****,* ***data, tools*** *y* ***submission***) y dos archivos (***README*** *y* ***LICENSE***). El directorio ***data*** tiene a su vez los directorios ***raw*** y **processed**. El directorio ***submission*** contendrá a su vez dos directorios (***version1*** y ***version2***).

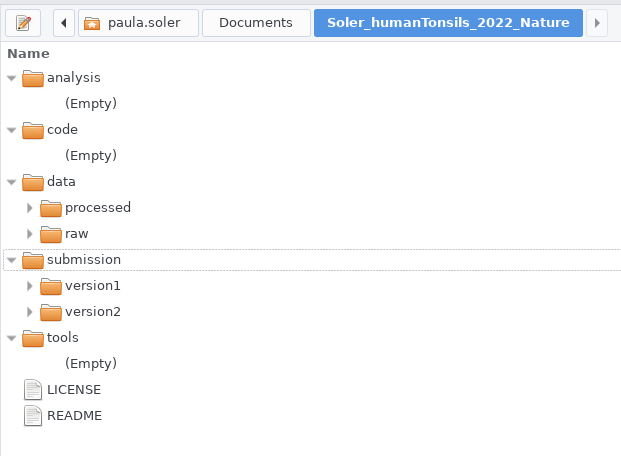


Figura 1. Estructura de directorio vista desde la interfaz gráfica

**Después de crear la estructura de directorios antes indicada, incluya una captura de pantalla usando el comando *tree* desde el terminal. También puede incluir una captura de pantalla desde la interfaz gráfica. (0,25 pts)**

Interfaz de usuario gráfica, Texto, Correo electrónico

Descripción generada automáticamente

**Parte II: Obtención de datos**

Ahora, va a obtener el conjunto de datos con los cuales va a trabajar. En este caso, se trabajará con un conjunto de datos generados mediante la conocida técnica de secuenciación “shotgun” del hígado de una *Boa constrictor* a la que se le diagnosticó la enfermedad del cuerpo de inclusión o IBD como consecuencia de una infección viral.

Para obtener el conjunto de datos, abra un determinado navegador y acceda al enlace que se le determina a continuación: [**https://www.ncbi.nlm.nih.gov/pubmed/25993603**](https://www.ncbi.nlm.nih.gov/pubmed/25993603). Desplácese hacia abajo y busque la sección “*Related Information*” (esquina inferior derecha) y haga clic en el enlace “SRA” (Short Read Archive). Esto muestra los conjuntos de datos SRA asociados a esta publicación.

**Responde a la siguiente pregunta relacionado con el SRA. ¿Qué es y para qué se utiliza? (0,25 pts)**

* El SRA se trata de los archivos principales del NIH de datos de secuenciación de alto rendimiento, formando parte de la Colaboración internacional de bases de datos de secuencias de nucleótidos (INSDC) que incluye el Archivo de lectura de secuencias (SRA) del NCBI, el instituto Europeo de Bioinformática (EBI) y la Base de datos de ADN de Japón. Los datos de Sequence Read Archive (SRA, por lo visto antes se les llamaba **short read archives**), disponibles a través de múltiples proveedores en la nube y servidores NCBI, son el repositorio público más grande disponible de datos de secuenciación de alto rendimiento. El archivo acepta datos de todas las ramas de la vida. SRA almacena datos de secuenciación sin procesar e información de alineación para mejorar la reproducibilidad y facilitar nuevos descubrimientos a través del análisis de estos datos (1).

Se le abrirá una ventana nueva donde podrá observar múltiples resultados asociados a todas las muestras procesada. De todas ellas, busque y acceda a “**shotgun sequencing of tissue from snake\_6\_viral**”. Observe en la parte inferior de la página el número de *run* (SRR #) para este conjunto de datos ([**SRR1984406**](https://trace.ncbi.nlm.nih.gov/Traces?run=SRR1984406)). **Añada una captura de pantalla que verifique que ha encontrado este conjunto de datos (0,25 pts)**

Escala de tiempo

Descripción generada automáticamente

Figura 2: Conjunto de datos de snake\_6\_viral.

Con este número, ya tiene la posibilidad de descargar los datos de archivo de SRA. Para realizarlo, deberá emplear una toolkit específica para descargar, forma automática, los datos asociados a este experimento en su entorno virtual. Para ello deberán:

### **Instalar SRA toolkit**

A continuación, se le expone los pasos claves para la instalación de esta herramienta. Todos ellos han sido extraídos de: *https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit*

* Descargue la herramienta **SRA toolkit** en el directorio creado anteriormente llamado ***tools*** (que utilizará para guardar las herramientas empleadas) utilizando el siguiente comando:

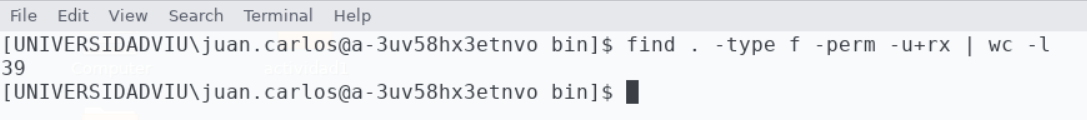
wget --output-document sratoolkit.tar.gz <https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-centos_linux64.tar.gz>

* Extraiga el contenido del archivo comprimido con el siguiente comando:

tar -vxzf sratoolkit.tar.gz

* Una vez descomprimido, puede eliminar el archivo descargado. Dentro del directorio generado llamado **sratoolkit.3.0.0-centos\_linux64,** podrá encontrar otro directorio llamada **bin** y dentro de éste podrá observar una gran cantidad archivos ejecutables.

**Responde a la siguiente pregunta: ¿Cuántos archivos ejecutables puede diferenciar dentro del directorio bin? Recuerde añadir una captura de pantalla con los comandos empleados para contar el número de archivos ejecutables y repórtelo por pantalla (0,25 pts)**

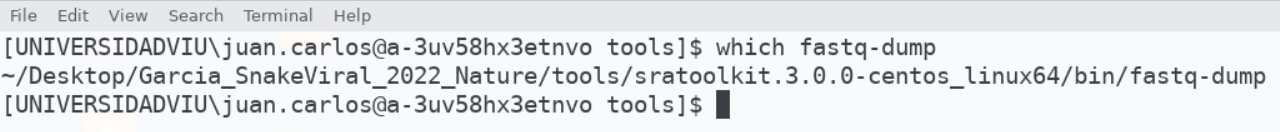


* Dentro del directorio **bin**, localice el ejecutable **fastq-dump** que será aquel que utilizará para descargar de forma automática los archivos asociados al SRA. Para ejecutarlo de forma sencilla, sin necesidad de que cada vez que lo quiera ejecutar tenga que determinar toda la ruta completa, puede añadir la ruta completa del mismo a la variable de entorno PATH de la siguiente manera:

export PATH=$PATH:$PWD/sratoolkit.3.0.0-centos\_linux64/bin

* **Verifique que fastq-dump puede ser encontrado por el Shell con el siguiente comando y adjunte una captura de pantalla del resultado del mismo (0.5 pts).**

which fastq-dump



* Finalmente, escriba el comando fastq-dump en la terminal. Cuando esto lo realice por primera vez, le aparecerá un texto que le indicará que la instalación no está configurada y para ello deberá correr el siguiente comando:

vdb-config --interactive

* Verá una pantalla azul con diferentes opciones, donde deberá habilitar la opción *Enable Remote Access* (quizá por defecto la vea marcada) y darle a la tecla X para salir de este menú interactivo.

### **Empleo de SRA toolkit para descargar las lecturas**

Finalmente, descargue el conjunto de datos utilizando el comando “fastq-dump” seguido del número SRR # del conjunto de datos seleccionados. Añada la opción --split-files para crear 2 archivos sincronizados por cada una de las lecturas emparejadas (paired-end reads). Aquí se muestra el comando completo a ejecutar:

fastq-dump SRR1984406 --split-files

**Confirme que descargó los archivos correctamente añadiendo una captura de pantalla del comando ls -l en el directorio data/raw** **(donde deberá localizar estos dos archivos) (0.25 pts).**

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

* En su momento me faltó hacer el ***ls -l***, lo muestro por separado:

Texto

Descripción generada automáticamente

Seguidamente, responda a cada una de las preguntas que se le indican, adicionando siempre los comandos empleados en cada caso y una captura de pantalla de ellos junto a su resultado.

* **¿Qué tipo de formato tienen los archivos descargados y para qué se utiliza dicho formato? (0.25 pts).**
  + EL formato FASTQ se trata de una secuenciación basada en texto que almacena archivos de datos con los contenidos de datos crudos de secuencia y el puntuaje de calidad (2).
  + A su vez se encuentra formado por los siguientes componentes (3):
    - Un identificador de secuencia.
    - La secuencia (los nucleótidos ACTG y N en caso de valores no identificados).
    - Un separador, al cual simplemente se le añade el signo “+”.
    - El puntuaje de calidad de la base. Estos están codificados por Phred +33, usando caracteres de ASCII que representan los valores numéricos de los puntuajes de calidad.
* **¿Cuántas líneas en total tienen cada uno de los archivos? (0,5 pts)**
  + **Script: 1nrow.sh.**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* + **Stdout: 1nrow.sh. (además creo las variables fuera del script srr1 y srr2 para que se vean mejor las imágenes y no sean tan larga).**

Texto

Descripción generada automáticamente

* **¿Cuántas secuencias incluye el archivo SRR1984406\_1.fastq? (0,5 pts)**
  + **Script: 2nsecuencias.sh**

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

* + **Stdout: 2nsecuencias.sh**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

* **¿Cuántas veces aparece la secuencia ATGATG en cada uno de los archivos descargados? (0,5 pts)**
  + **Script: 3seq\_identificar.sh.**

Texto, Escala de tiempo

Descripción generada automáticamente

* + **Stdout: 3seq\_identificar.sh.**

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

A continuación, va a convertir el archivo **SRR1984406\_1.fastq** que se presenta en formato fastq**,** en otro archivo con un formato un tanto distinto, el formato FASTA.

**¿Qué diferencias hay entre un archivo en formato FASTQ y un archivo en formato FASTA? (0,5 pts)**

* El formato FASTA estaría únicamente formado por el encabezado con la identificación de la secuencia y la secuencia en sí. No se le añade la tercera y cuarta línea del formato fastq (es decir, el separador con el signo “+” y los puntuajes de calidad). Además, el encabezado en vez de empezar con un "@" empezaría con un ">".

Para ello, deberá seleccionar la primera y la segunda línea de cada una de las 4 líneas que conforman cada una de las lecturas o *reads* del archivo **SRR1984406\_1.fastq** y así confeccionar un archivo final llamado ***all\_sequences.fasta*** que contenga el *header* o cabecera y la secuencia asociada a cada lectura. Recuerda guardar este archivo en el directorio *data/processed*.

**Visualice las 5 primeras líneas del archivo creado en el directorio determinado. Incluya una captura de pantalla con el código empleado y el resultado solicitado (0,5 pts)**

* + **Script: fasta\_format.sh.**

Texto

Descripción generada automáticamente

* + **Stdout: fasta\_format.sh.**

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Una vez creado, seleccione aleatoriamente 5 lecturas (con su cabecera y su secuencia asociada) del archivo all\_sequences.fasta y seguidamente guárdelas en 5 archivos de texto distintos denominados secuencia1.fasta, secuencia2.fasta y así sucesivamente. Cada uno de ellos contendrá lo siguiente:

>SRR1984406.1 1 length=135

GACGACTGCCATCTGAACGTGTGGAATCAACGGAGCCACATCTGACTTCCAGTATCCATCCGAAGTTCTCCATTCAATAGTGAGGAATCTGACGACTGCCATCTGAACGTGTGGAATCAACGGAGCCACATCTG

**Muestre los archivos creados adjuntado una captura de pantalla (0,5 pts)**

* + **Script: 5randomseq.sh.**

Texto

Descripción generada automáticamente

* + **Stdout: 5randomseq.sh.**

Texto

Descripción generada automáticamente

Con uno de estos archivos creados (puede seleccionar el que prefiera de los 5), deberá realizar un script en BASH (esté deberá estar almacenado en el directorio *code*) que sea capaz de leer dicho archivo, específicamente su secuencia, y deberá reportar por la salida estándar si se trata de una secuencia nucleotídica o aminoacídica.

**Pegue el script generado y adicione una captura de pantalla de su ejecución por terminal (2 pts)**

* + **Script: 6seq1.sh.**

Texto

Descripción generada automáticamente

* + **Stdout: 6seq1.sh.**

Texto

Descripción generada automáticamente

A continuación, implemente otro script en BASH (esté deberá estar almacenado en el directorio *code*) que lea cada uno de los 5 archivos generados (secuencia1.fasta, secuencia2.fasta, etc.) y transforme su secuencia de ADN a ARN (recuerde que base nitrogenada cambia entre estas dos moléculas). Una vez traducidas, deberá contar la frecuencia de aparición de cada una de las 4 bases nitrogenadas y reportarlas por la salida estándar para cada una de las 5 secuencias analizadas.

**Pegue el script generado y adicione una captura de pantalla de su ejecución por terminal (2 pts)**

* + **Script: 7dna\_to\_rna.sh.**

Texto

Descripción generada automáticamente

* + **Stdout: 7dna\_to\_rna.sh.**

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Seguidamente debe completar el archivo de texto llamado LICENSE, donde deberá indicar que tipo de licencia quiere utilizar para limitar o no el uso, la modificación y la distribución de su código BASH desarrollado. Es muy importante comenzar a concienciarse sobre las posibles licencias que podemos otorgarles a nuestros scripts, por ello, busque información en la red sobre los distintos tipos y cuál se adaptaría mejor a sus preferencias. **Adjunte una captura de pantalla con el contenido de este archivo (0,5 pts).**

* Yo en mi cuenta de GitHub uso siempre, o al menos hasta el momento, la licencia **MIT** (***Massachusetts Institute of Technology***).
  + Consiste en una licencia muy permisiva, que permite el uso de mi código, para muchas cosas (reproducir el experimento, sacar rédito económico...), con las limitaciones de garantía y responsabilidad con este y como única condición, el aviso de derechos de autor del código y hasta donde tengo entendido, el código modificado también tiene que presentar la licencia MIT.

* + En GitHub a la hora de crear un repositorio, es una de las opciones de serie de la página.

* + Adjunto la información que GitHub pone por defecto:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Figura 3. Información resumida de la Licencia MIT.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Figura 4. Contenido de la licencia MIT.

Además, deberá completar el archivo README, explicando en él, el contenido de cada uno de los directorios creados en el proyecto, especificando las características más relevantes de cada uno de ellos. En este archivo puede añadir la información más relevante que considere. **Adjunte una captura de pantalla con el contenido de este archivo (0,25 pts).**

En este apartado lo que he hecho es, añadir el título, mis credenciales, así como copiar el objetivo de la actividad y de explicar cada uno de los directorios, en formato Markdown.

Como no puedo añadir una captura de pantalla con todo lo que he puesto, adjunto su link (<https://github.com/Juankkar/Programacion_Shell_Scripting_VIU/blob/main/Garcia_SnakeViral_2022_Nature/README.md>).

**Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente**

Figura 5. Contenido del README en mi cuenta de Github.

Finalmente, genere una copia de esta plantilla cumplimentada y trasládela al directorio submission/version1 donde se almacenará la primera versión de su trabajo realizado. **Para acabar este proyecto, vuelva a incluir una captura de pantalla con la estructura de directorios actual usando de nuevo el comando *tree* desde el terminal (0,25 pts).**

* Como mi intención es subir el contenido de este pipeline a mi cuenta de GitHub y sratoolkit es muy pesado, incluso después de comprimirlo con el siguiente comando tar -zcvf sratoolkit.tar.gz sratoolkit.3.0.0-centos\_linux64 (me daba ~83 Mb de tamaño por si solo) lo he sustituido por un README con la cuenta de Github para descargar el programa. Igualmente adjunto un ls -l del escritorio donde he depositado sratoolkit para mostrar que lo he dejado fuera.

Texto

Descripción generada automáticamente con confianza media

* Los scripts con el código para los análisis se pueden ver en este enlace <https://github.com/Juankkar/Programacion_Shell_Scripting_VIU/tree/main/Garcia_SnakeViral_2022_Nature/code>.
* Finalmente adjunto el árbol final (En análisis tuve que añadir un README, no tiene nada en particular, pero si no añadía algo no me enviaba este directorio a mi repositorio con git push).

Texto

Descripción generada automáticamente con confianza media

## **Referencias:**

1. The Sequence Read Archive (SRA): Getting Started [Internet]. [citado 25 de noviembre de 2022]. Disponible en: https://www.ncbi.nlm.nih.gov/sra/docs/

2. Sequence File Formats | FASTQ & BCL formats for Illumina sequencing [Internet]. [citado 25 de noviembre de 2022]. Disponible en: https://emea.illumina.com/informatics/sequencing-data-analysis/sequence-file-formats.html

3. FASTQ files explained [Internet]. [citado 25 de noviembre de 2022]. Disponible en: https://emea.support.illumina.com/bulletins/2016/04/fastq-files-explained.html