Universidad de los Andes

BCOM4102-Ecologia Microbiana

Programa de Maestría en Biología Computacional

Taller 3: Pre-procesamiento de secuencias

Febrero 11 de 2021

**Taller 3**

Carlos Andrés Díaz **- código: 202010343**

David León **- código: 201615216**

Cesar Patiño **-código: 201924259**

Inicie una sesión en el cluster (magnus.uniandes.edu.co), recuerde hacerlo vía ssh (desde Terminal o MobaXterm), su login es bcom4102 y el password es 202110bcom4102.

**Objetivos:**

Después de estudiar este tutorial, usted debería poder:

**1.** Describir los pasos involucrados en el pre procesamiento y limpieza de los datos de secuenciación.

**2.** Distinguir entre una buena y una mala corrida de secuenciación de Illumina.

**3.** Evaluar y mejorar la calidad de los datos de secuencia de un experimento de secuenciación.

**4.** Extraer secuencias de ITS de datos crudos de secuenciación para grupos eucariotas de interés.

**Sección I: Calidad**

Los datos para usar en esta sección se depositaron en el European Nucleotide Archive. Puede ir al [sitio web de ENA](https://www.ebi.ac.uk/ena/browser/home) y buscar los archivos con el número de accesión SRR957824. Estos datos son "casi" datos brutos, ya que se han eliminado todas las secuencias que se identificaron como pertenecientes al genoma PhiX. Este proceso requiere algunas habilidades que se aprenderán más adelante en el curso.

1. ¿A qué experimento corresponden los datos?

**Los datos corresponden a un experimento realizado por Slayton y colaboradores en el año 2013, y cuyo objetivo era caracterizar el patógeno causado por la ingesta de lechuga romana en personas de diferentes estados de Estados Unidos, el código de acceso de para este experimento en el archivo europeo de nucleótidos ENA por sus siglas en inglés es PRJNA215830**

¿Cuál fue el tipo de secuenciación utilizada para esta corrida?

**El tipo de secuenciación utilizada fue ILLUMINA**

¿A qué corresponden los archivos disponibles?

**Los archivos adjuntos al proyecto con código de acceso** **SRR957824 son :**

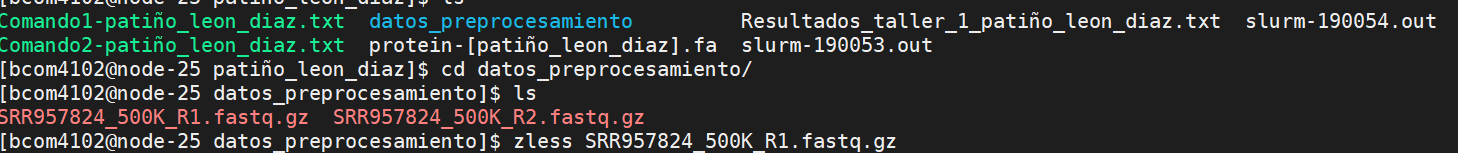
Como ha podido observar, estos archivos contienen alrededor de 3 millones de lecturas y, por lo tanto, son bastante grandes. Solo usaremos un subconjunto del conjunto de datos original para este taller.

**2.** Cree un directorio de datos en su carpeta de grupo que tenga el nombre “datos\_preprocesamiento”

**3.** Haga una copia en su directorio de los datos a para trabajar que están presentes en el directorio “Datasets/preprocessing/” **y** que tengan la extensión “.fastq.gz” únicamente.

**4**. Los archivos que estamos usando son archivos FASTQ. Inspeccione uno de ellos sin descomprimirlo.

**Para realizar la inspección de los archivos sin descomprimir los archivos se usará el comando zless en cada uno de los dos archivos tal y como se muestra:**



**Por lo que, al visualizar la información del archivo se obtiene:**



**5.** ¿Cuántas lecturas pareadas se tomaron aleatoriamente de los datos originales? Realice el conteo tomando en cuenta las cuatro líneas por lectura, de acuerdo con lo visto en la sesión de preparación para el taller.

**Para realizar el cálculo del número de lecturas generadas a partir de los datos originales, es muy importante tener presente que cada archivo cuenta con cuatro líneas por lectura, por lo que al realizar el conteo de líneas con el comando wc -l se debe dividir 4, en donde se obtienes 500000 lecturas pareadas tal y como se muestra:**



**6.** ¿Qué le indican los primeros cinco nombres de las lecturas en ambos archivos?

Para verificar la calidad de los datos de la secuencia usaremos una herramienta llamada FastQC.

**7**. Cargue el módulo más reciente de FastQC disponible en magnus.



**8.** Describa los archivos que ha producido FastQC.

**9.** Descargue y abra los archivos html con su navegador web de preferencia. 

**10**.

¿A qué indicadores debe prestar especial atención en los reportes obtenidos? Descripciones de los distintos campos indicadores de calidad están disponibles [aquí](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/).

**11.** Compare su reporte con estos ejemplos de un juego de datos de [buena calidad](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html) y otro de [mala calidad](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html).

**Sección II: Limpieza de secuencias**

En esta sección se realizarán tres pasos, se recortarán los adaptadores, se recortarán las lecturas de acuerdo a la calidad, y se realizará una nueva evaluación de la calidad de las secuencias.

**12.** Identifique en el reporte de FastQC la presencia de adaptador(es) y a qué tipo corresponde(n).

**13.** Cargue el módulo de *Trimmomatic* disponible en magnus. Busque los archivos de adaptadores que están incluidos en la instalación de este programa.

**14.** Ejecute el programa en modo pareado y asegúrese de retener ambas lecturas. Utilice un *sliding window* de 4 para remover las bases con puntuación *Phred* inferior a 15. También elimine las lecturas que tengan menos de 40 bases después de este paso.

**15.** ¿Qué pasa si utiliza un *sliding window mayor*?

**16.** Obtiene cuatro archivos de salida de *Trimmomatic*. ¿Puedes explicar qué son?

**17**. De las lecturas “*single*” resultantes de Trimmomatic, ¿cuál tiene más lecturas? ¿Por qué?

**18.** ¿Cuántos registros pareados se mantienen después de la corrida?

**19**. Vuelva a correr FastQC en las secuencias después del *trimming*. ¿Qué observa? ¿Hubiera sido necesario incluir una remoción de las bases iniciales o finales de cada lectura? ¿Cómo debería modificarse el comando de *trimming* para incluir ese paso? (no se debe ejecutar un nuevo *trimming*)

**Sección III: Extracción de secuencias ITS**

ITSx (por *e****X****tractor*) es una herramienta basada en Perl para extraer ITS1, 5.8S e ITS2, así como secuencias ITS completas, de datos de secuenciación de Sanger y nueva generación. ITSx utiliza modelos ocultos de Markov calculados a partir de grandes alineamientos de un total de 20 grupos de eucariotas, incluidos hongos, metazoarios y plantas, y la extracción de la secuencia se basa en las posiciones predichas de los genes ribosomales en las secuencias.

**20.** Ahora copie en su directorio “datos\_preprocesamiento” el archivo “ITS\_dataset\_untrimmed.fasta” del directorio compartido de datos “Datasets/preprocessing/”. Inspeccione el archivo. ¿Cuántas secuencias tiene?

**21.** Cargue el módulo de Seqtk en magnus. Esta es otra herramienta rápida para procesar secuencias en formato FASTA o FASTQ. Para realizar la práctica, vamos a trabajar con una muestra del 10% de las secuencias de este archivo. Busque el comando en [Seqtk](https://github.com/lh3/seqtk) para realizar esta labor y nombre el archivo de salida como “ITS\_dataset\_untrimmed\_0.1.fasta”

**22.** Con ayuda del [tutorial de ITSx](https://microbiology.se/publ/itsx_users_guide.pdf), genere archivos de salida para solo ITS1 y solo ITS2 del conjunto de datos submuestreado. ¿Qué comando utilizó? ¿Cuántos genes obtiene para cada uno?

**23.** Inspeccione el archivo de salida “.extraction.results”, que contiene los resultados detallados de la extracción de estas regiones. ¿Cuántas entradas están marcadas como potenciales quimeras?

**24.** Finalmente, las secuencias en estos archivos están marcadas según su origen putativo (tipo de organismo). Indique para qué organismos pudo extraer secuencias ITS y las cantidades correspondientes.