Actividad 2 – 1a convocatoria.

Edición abril 2024

Introducción a la actividad

Esta actividad corresponde al ejercicio de evaluación del Tema 5. Tendréis que descargar los datos crudos de secuenciación Illumina pareados de un genoma completo de *Salmonella enterica*, aislado de una muestra clínica. El ejercicio debe realizarse dentro de lo establecido en clase. Para ello debéis hacer uso del entorno de trabajo proporcionado 04MBIF_bacteriano, utilizado en clase, y que contiene todos los programas necesarios para la ejecución. Atentos a las notas particulares de cada apartado.

Nota: recordad que los entornos están disponibles en el apartado Recursos y materiales / 01. Materiales docentes / Environments Conda de la web de la asignatura.

Link de descarga de datos

Actividad2_1C_alumnos

https://alumnosviu-my.sharepoint.com/:f:/g/personal/laura_gutierrez_m_professor_universidadviu_com/EjPOsrWiqmZMicwryS3tndwB8cuyA_bxMXHlXXeWe-mDWA?e=EKkU5r

Objetivo

Realizar el análisis completo de esta muestra, incluyendo los **pasos** siguientes:

1. Realiza el análisis de calidad de las lecturas iniciales.
2. Realiza la limpieza de calidad de las lecturas, junto con el análisis de calidad de las lecturas resultantes
3. Realizar el ensamblaje *de novo* a partir de las lecturas limpias obtenidas en pasos anteriores.
4. Realizar la evaluación del ensamblaje.
5. Anotar el genoma ensamblado, utilizando las proteínas de referencia del genoma indicado (debéis descargarlas) como base de datos primaria, y los contigs ensamblados como elemento para anotar.
6. Realizar el análisis de genes de resistencia a antibióticos, metales pesados y biocidas, así como de genes de virulencia con los parámetros especificados.

Formato de entrega

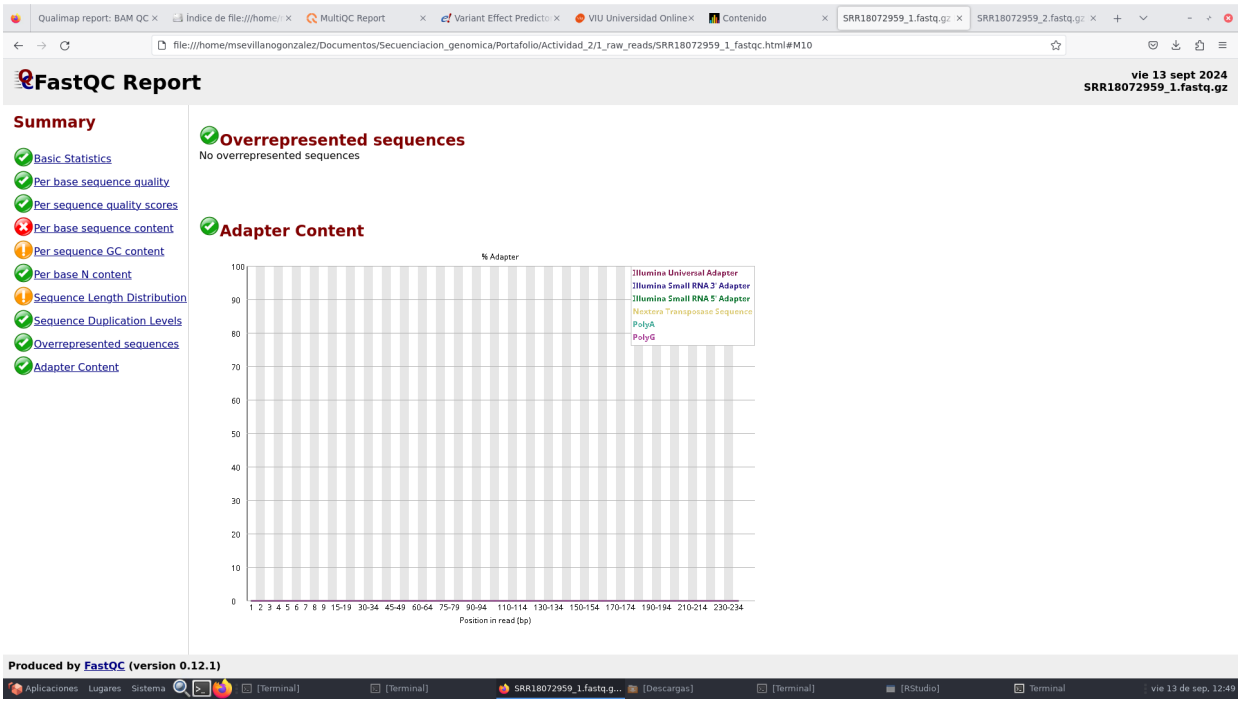
Importante: Leed atentamente y contestad a lo que se pregunta

1. **Descarga las lecturas de trabajo desde el link provisto y realiza un análisis de calidad de las lecturas. Contesta a las siguientes preguntas (1 punto):**
    a. **¿Cuántas lecturas se han secuenciado?**
    Se han secuenciado 1.626.620 lecturas

    b. **¿Qué longitud de secuencia tienes?**
    La longitud de las distintas secuencias es de 35 a 250 pb.

    c. **Incluye el/los comandos utilizados para obtener la información**
    fastqc *.fastq.gz
    firefox SRR18072959_1_fastqc.html
    firefox SRR18072959_2_fastqc.html
    zgrep -c "^@" *.fastq.gz

    d. **Incluye el pantallazo de la gráfica "Adapter Content" obtenida para las lecturas R1 e indica si consideras necesario realizar la limpieza de adaptadores o de algún tipo de elemento.**



    Realmente atendiendo a la gráfica no parece muy necesario, pero siempre conveniente porque Fastq no muestra todos los adaptadores posibles.

2. **Realiza la limpieza de calidad de las lecturas iniciales, junto con el análisis de calidad de las lecturas resultantes (1 punto). Parámetros para considerar: longitud mínima 151 bp, calidad media, inicial y final de lectura = Q25. Entrega:**
    a. **Comando/s necesarios para realizar la tarea.**
    Comando para realizar la limpieza de las lecturas:

    fastp -i 1_raw_reads/SRR18072959_1.fastq.gz -I 1_raw_reads/SRR18072959_2.fastq.gz -o 2_trimming/out_R1.fastq.gz -O 2_trimming/out_R2.fastq.gz --cut_tail 25 --cut_front 25 --cut_mean_quality 25 -l 151 --detect_adapter_for_pe -h report_fastp.html

    b. **¿Cuántas lecturas finales han pasado filtros? ¿Cuántas lecturas se han eliminado por ser demasiado cortas? ¿Y por baja calidad? ¿En cuántas lecturas se han eliminado adaptadores?**
    ¿y polyX?
    3155758 lecturas han pasado los filtros.
    Se han eliminado 70344 por ser demasiado cortas.

26234 lecturas se han eliminado por baja calidad.
En 5004 lecturas se han eliminado los adaptadores.

c. **Indica cuál es el tamaño medio de las lecturas resultantes R1 y R2, así como el porcentaje de duplicación.**
El tamaño medio de las lecturas de R1 y R2 es de 151-250 pares de bases.
El porcentaje de duplicación para R1 es de 82,93%, y para R2 es de 83,57%.

d. **Incluye el pantallazo de la gráfica "Per base sequence quality" de las lecturas R2 de este apartado.**

3. **Realiza el ensamblaje de novo con el programa SPAdes sobre las lecturas limpias obtenidas en el paso anterior. Activa las tres fases de análisis (corrección de errores, ensamblaje y corrección de mismatches).** ***(2 puntos)***

a. **Indica el comando ejecutado**
spades.py -1 ../2_trimming/out_R1.fastq.gz -2 ../2_trimming/out_R2.fastq.gz --careful -k auto -o spades_out

b. **Indica cuáles han sido los kmeros testados**
Ha testeado los kmeros: 21, 33, 55, 77, 99, y 127.

c. **Indica cuál ha sido el kmero seleccionado como óptimo por SPAdes**
El kmero seleccionado como óptimo es el K127.

d. **Indica el número de contigs que has obtenido del ensamblaje.**
Se han obtenido 4047 contigs.

4. **Analiza la calidad del ensamblaje sobre todos los contigs del ensamblaje anterior. (2 puntos).**

a. I**ndica el comando utilizado y muestra un pantallazo del resultado obtenido**
quast.py -o ../../4_quast_result/ -m 0 -t 2 -k --k-mer-size 127 --circos --pe1 ../../2_trimming/out_R1.fastq.gz --pe2 ../../2_trimming/out_R2.fastq.gz contigs.fasta scaffolds.fasta

b. **Indica el valor de N50, L50, N90 y L90 obtenidos**
N50=106.030.
N90=503
L50=13
L90=2303

c. **Indica el número de contigs que has obtenido mayores de 5000 bp, así como la longitud total ensamblada en este tipo de contigs**
34 contigs mayores de 5000 pb, cuya longitud total ensamblada es de 4.828.231.

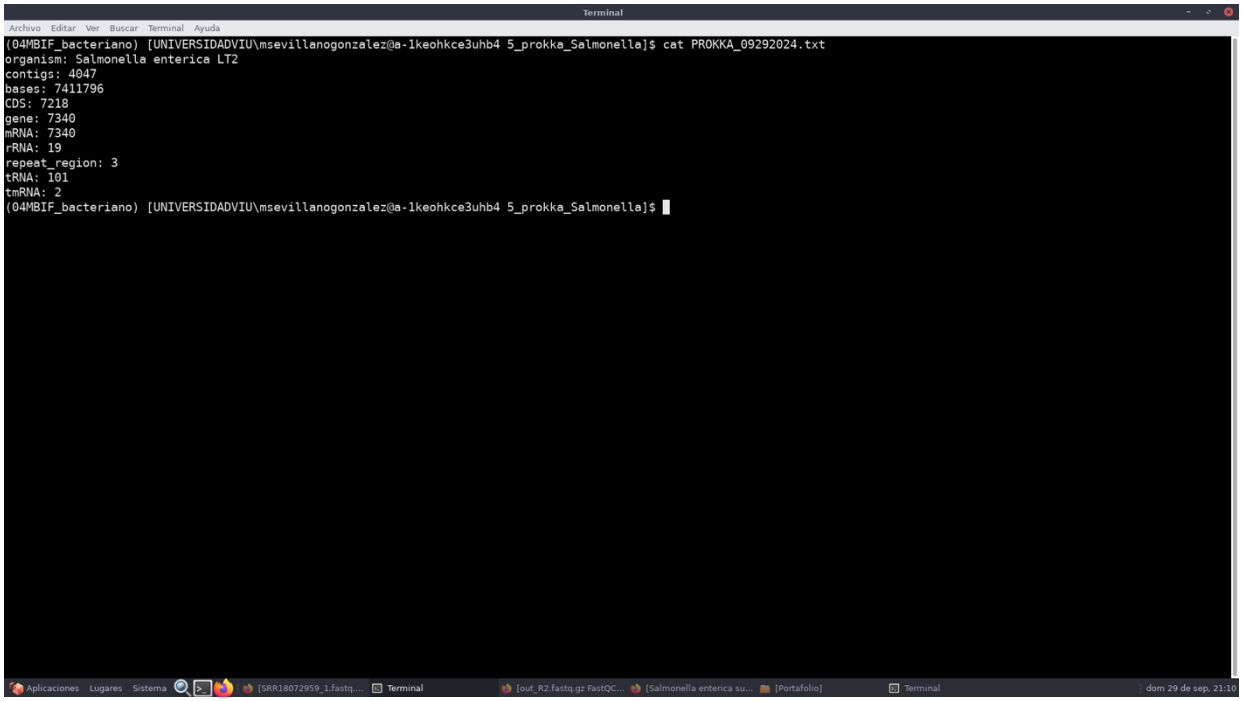
d. **Indica el número de lecturas mapeadas sobre los contigs ensamblados, así como la cobertura media estimada.**
El número de lecturas mapeadas sobre los contigs ensamblados es de 3.195.979 y la cobertura media estimada es de 104 contigs.

5. **Realiza la anotación de los contigs obtenidos en SPAdes. Para ello debes utilizar como proteínas de referencia las del genoma Salmonella enterica serovar Typhimurium LT2 (GCF_000006945.2), que podrás obtener del link indicado a continuación. Descarga el genoma y obtén el proteoma para poder utilizarlo, tal y como vimos en clase. Link: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000006945.2/ (2 puntos)**
   a. **Indica el comando necesario para obtener el proteoma de la referencia para su uso en la anotación.**
   prokka-genbank_to_fasta_db ncbi_dataset/data/GCA_000006945.2/genomic.gbff > ref_prots.faa

   b. **¿Cuántas proteínas de referencia tiene este genoma descargado?**
   Tiene 4.544 proteínas de referencia

   c. **Indica el comando que has utilizado para la anotación de los contigs ensamblados.**
   prokka --outdir 5_prokka_Salmonella --addgenes --addmrna --genus Salmonella --species enterica --strain LT2 --kingdom Bacteria --usegenus --proteins ref_prots.faa --mincontiglen 0 3_SPAdes/spades_out/contigs.fasta

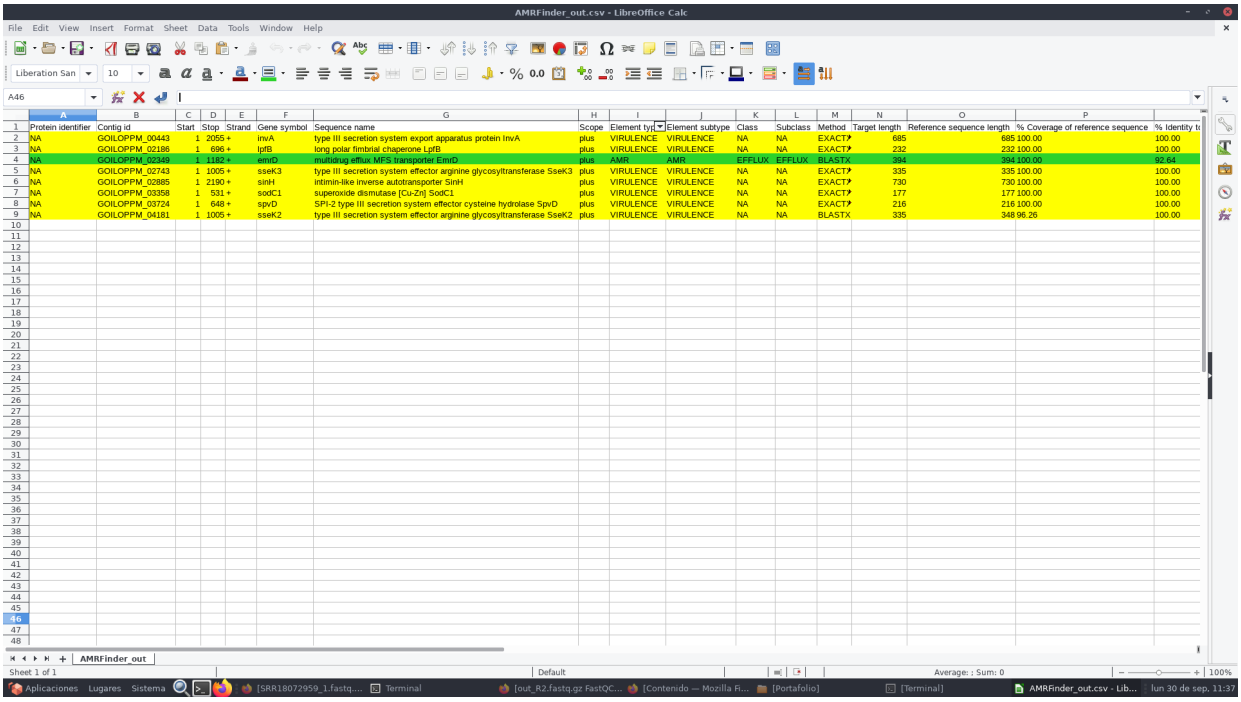
   d. **Revisa los archivos de salida, incluye un pantallazo del archivo donde te indica esta información, e indica explícitamente: el número total de CDS detectados, el número de mRNA, el número de rRNA y de tRNA anotados.**
   CDS: 7218
   mRNA:7340
   rRNA:19
   tRNA:101



6. **Realiza el análisis de genes de resistencia a antibióticos, metales pesados, biocidas y factores de virulencia. Utiliza los valores de 90% de cobertura y 90% de identidad para el análisis. (2 puntos)**
    a. **Indica la versión del software y de la base de datos que estás utilizando.**
    La versión del software es 3.12.8, y la de la base de datos es 2024-07-22.1 del programa ARMFinder

    b. **Indica el comando/s utilizados para realizar el análisis.**
    amrfinder --organism Salmonella -n ../5_prokka_Salmonella/PROKKA_09292024.ffn -c 0.9 -i 0.9 --plus --log amrfinder.log > AMRFinder_out.csv

    c. **Abre el archivo de salida y muestra un pantallazo donde indiques coloreados sobre la tabla los genes de resistencia a antibióticos (verde) y genes de virulencia (amarillo).**



    d. **¿A qué tipo de antibióticos es este genoma resistente?**
    Es resistente a algunos antibióticos β-lactámicos y tetraciclinas.

Instrucciones entrega

Cada captura debe contener la hora y fecha del ordenador (captura de pantalla completa). Cualquier captura que no cumpla este requisito no será válida para la entrega del ejercicio.

La entrega se debe realizar en un único documento en formato PDF.

Límite de entrega:

18/10/2024 a las 23:59

Criterios de evaluación.

La actividad tiene 6 apartados, con valor variable (indicado en el apartado FORMATO DE ENTREGA).

Debe numerarse apropiadamente cada uno de los apartados para su entrega. La entrega en un apartado que no

corresponde, no se puntuará.

Se valorará el correcto seguimiento de las instrucciones de la actividad.