Actividad 3 – 1a convocatoria.

Edición abril 2024

Introducción a la actividad

Esta actividad corresponde al ejercicio de evaluación del Tema 6. Tendréis que descargar los datos crudos de secuenciación Illumina MiSeq pareados de un genoma de SARS-CoV-2, cuyo RNA se ha obtenido a partir de un hisopo de un paciente. El ejercicio debe realizarse dentro de lo establecido en clase. Para ello debéis hacer uso del entorno de trabajo proporcionado 04MBIF_COVID, 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

En esta carpeta encontraréis:

- Carpeta 1_reads: tenéis las lecturas crudas (directas del secuenciador) en formato FASTQ correspondientes a la muestra a analizar. Se trata de una muestra de SARS-CoV-2, amplificada según el protocolo ARTIC v4 y secuenciada en un equipo MiSeq.
- Carpeta 2_refs: en esta carpeta tenéis los archivos de referencia del genoma Wuhan en formato fasta. También tenéis disponible el archivo en formato BED con los cebadores/primers utilizados en la preparación de la genoteca y la base de datos Kraken, con el genoma humano y los genomas de virus, lista para desempaquetar y utilizar.

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. Realiza la limpieza de lecturas de hospedador y extrae las lecturas útiles para el análisis.
4. Mapeo completo de las lecturas libres de humano y limpias sobre el genoma de referencia.
5. Realizar el enmascaramiento de primers y la llamada de variantes.
6. Extraer la secuencia consenso, utilizando para el programa iVar los parámetros mínimo score de calidad para contar una base: 30; mínima frecuencia para llamar a un consenso: 90%; mínimo profundidad para llamar a consenso: 50
7. Finalmente, realizar la determinación del linaje asociado a esta secuencia. Recordad actualizar la base de datos en el momento de realizar el ejercicio. Indicad las versiones de la base de datos y motores de búsqueda en la entrega.

Formato de entrega

Importante: Leed atentamente y contestad a lo que se pregunta. Recordad que es importante numerar cada apartado y entregar lo que se pide en el apartado numerado correctamente. Si no, no se evaluará positivamente.

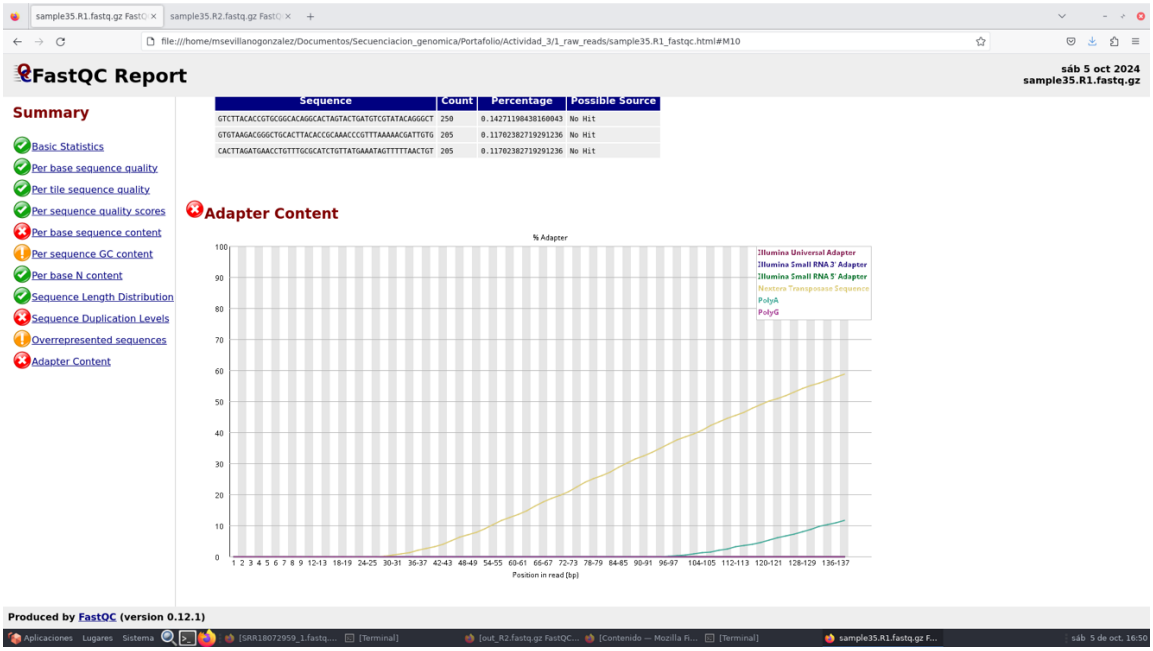
1. **Descarga las lecturas de trabajo desde los links provistos y realiza un análisis de calidad de las mismas.Contesta a las siguientes preguntas (1 punto):**
   a. **¿Cuántas lecturas se han secuenciado?**
   175.178

   b. **¿Qué longitud de secuencia tienes?**
   Una longitud de 150 pb.

   c. **Incluye el/los comandos utilizados para obtener la información**
   zgrep -c "@M04178" sample35.R*.fastq.gz

   fastqc sample35.R*.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.**



Sí es necesario realizar una limpieza, tanto de adaptadores, ya que claramente aparece en el gráfico Nextera transposase sequence, como de las secuencias polyA.

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 100 bp, calidad media, inicial y final de lectura = Q25. Entrega:**

   a. **Comando/s necesarios para realizar la tarea.**
   fastp -i sample35.R1.fastq.gz -I sample35.R2.fastq.gz -o ../2_clean_reads/outR1.fastq.gz -O ../2_clean_reads/outR2.fastq.gz --detect_adapter_for_pe --cut_tail 25 --cut_front 25 --cut_mean_quality 25 -l 100 --trim_poly_x --html out.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?
   El número de lecturas finales que han pasado los filtros es de 212.592.
   Se han eliminado 137.742 lecturas por ser demasiado cortas y 22 lecturas por bajada calidad.
   Se han eliminado adaptadores en 226.509 lecturas y polyX en 2.234 lecturas.

   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 está entre 100 y 150 pb.

   d. **Incluye el pantallazo de la gráfica "Per base sequence quality" de las lecturas R1 de este apartado.**

3. **Realiza la limpieza de lecturas de hospedador y extrae las lecturas útiles para el análisis. Indica el número de secuencias que no han sido clasificadas a genoma humano (1 punto)**
   El número que de secuencias que no han sido clasificadas a genoma humano es de 105.736.

4. **Realiza el mapeo completo de las lecturas libres de humano y limpias sobre el genoma de referencia que se ha provisto en el ejercicio. Entrega el/los comando/s utilizados para realizar este paso (2 puntos).**
   Comandos utilizados:

   bwa index GCF_009858895.2_ASM985889v3_genomic.fna

   bwa mem -Y -M -R '@RG\tID:\tSM:.' -t 2 GCF_009858895.2_ASM985889v3_genomic.fna ../3_kraken/sample1.R1.nonhuman.fq.gz ../3_kraken/sample1.R2.nonhuman.fq.gz | samtools sort | samtools view -F 4 -b -@ 2 -o sample1.sort.bam

5. **Realiza el enmascaramiento de primers y la llamada de variantes. (2 puntos).**

a. **Entrega los comandos utilizados para enmascarar primers. En qué porcentaje de reads se han recortado los primers?**

Comandos utilizados:

ivar trim -i sample1.sort.bam -p sample1.trim -m 30 -q 20 -s 4 -b nCoV-2019_v4.primer.bed

samtools sort sample1.trim.bam -o sample1.trim.sort.bam

samtools index sample1.trim.sort.bam

El porcentaje de reads en el que se han recortado los primers es del 23,58%

b. **Entrega los comandos utilizados para llamar variantes así como el número de variantes detectadas en el archivo iVar final (2 puntos).**

Comando para la llamada de variantes:

samtools mpileup -A -d 0 --reference GCF_009858895.2_ASM985889v3_genomic.fna -B -Q 0 sample1.trim.sort.bam | ivar variants -p sample1.ivar

Comando para contar el número de variantes detectadas en el archivo iVar final:

wc -l sample1.ivar.tsv, a lo que hay restarle uno, que es el encabezado.

89 es el número de variantes detectadas.

6. **Extraer la secuencia consenso, utilizando para el programa iVar los parámetros siguientes:**
mínimo score de calidad para contar una base: 30
mínima frecuencia para llamar a un consenso: 90%
mínimo profundidad para llamar a consenso: 50.
Entrega el comando escrito en el documento. (1.5 puntos).

Comando:

samtools mpileup -aa -A -d 0 -B -Q 0 sample1.trim.sort.bam | ivar consensus -p sample1.ivar -q 30 -t 0.9 -m 50 -n N

7. **Finalmente, realizar la determinación del linaje asociado a esta secuencia. Recordad actualizar la base de datos en el momento de realizar el ejercicio. Indicad las versiones de la base de datos y motores de búsqueda en la entrega. Debéis entregar escritos los campos "lineage", "scorpio call" y "scorpio support" del archivo resultante de la determinación del linaje. En este punto, utilizad un parámetro de ambigüedad del 10%. (1.5 puntos).**

Comando utilizado:

pangolin sample1.ivar.fa -o sample_Actividad3.pangolin --outfile sample_Actividad3.csv -t 2 --max-ambig 0.1

Versión de Pangolin, v4.3.1, versión de la base de datos Pangolin-data, v1.30. Versión de Scorpio, v0.3.19, y versión de Costellations, v0.1.12.

Lineage: no asignado
Scorpio call: Omicron
Scorpio support: 0.75

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 7 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.