Análisis de datos ómicos - Primera prueba de evaluación continua

Aràntzazu Alonso Carrasco

15 de abril, 2023

Índice

1 Descarga del archivo

Primero, debemos descargar el archivo *fastq* con el que trabajaremos a lo largo de la Prueba de Evaluación Continua (PEC). Para hacerlo, debemos ir al *Google drive* de la asignatura y buscar el archivo que corresponda. Dado que el último dígito numérico de mi DNI es un 4, el archivo que debo descargar corresponde a la muestra "S04_Bac03_read1.fastq".

2 Descripción de los datos

Una vez descargados los datos, podemos proceder a describirlos. El formato del archivo es .fastq, que es un formato ampliamente utilizado en bioinformática para representar datos de secuenciación de ADN o ARN en bruto. Concretamente, los archivos en formato .fastq contienen secuencias de nucleótidos junto con información de calidad asociada a cada nucleótido de la secuencia. Los archivos .fastq son muy utilizados en la secuenciación de nueva generación (NGS) para representar datos de secuenciación de alta calidad generados por tecnologías como Illumina.

Particularmente, el archivo `S04_Bac03_read1.fastq` contiene 250000 secuencias.

FALTA COMPLETAR

3 Control de calidad con `Galaxy`

Para hacer el control de calidad en Galaxy, podemos empezar creando un nuevo historial para la PEC. A continuación, en la parte superior del panel `Tools`, hacemos clic en `Upload` y añadimos el fichero `S04_Bac03_read1.fastq`. Una vez lo tenemos subido, tecleamos `FastQC` en el cuadro de búsqueda del panel de herramientas y clicamos en la herramienta homónima. Seguidamente, seleccionamos el fichero que

acabamos de subir y lo analizamos. Para visualizar los resultados, le damos al icono del ojo en el archivo de salida `FastQC on data 1: Webpage`.

Una vez hemos desplegado el análisis de los resultados, lo primero que vemos es un gráfico titulado *Per base sequence quality* (Figura 1). En él, podemos ver un resumen del rango de valores de calidad en todas las bases de cada posición del fiquechero FastQ. Para interpretarlo, debemos saber que la línea central roja representa el valor medio, la cajita amarilla representa el rango inter-cuartil (25-75%), los bigotes inferior y exterior representan los puntos 10% y 90%, mientras que la línea azul representa el la media de la calidad. El eje y del gráfico muestra las puntuaciones de calidad de manera que, a mayor puntuación, mejor calidad.

Como podemos ver en la Figura 1, todas las secuencias tienen una calidad muy buena, ya que se sitúan principalmente en la región verde.

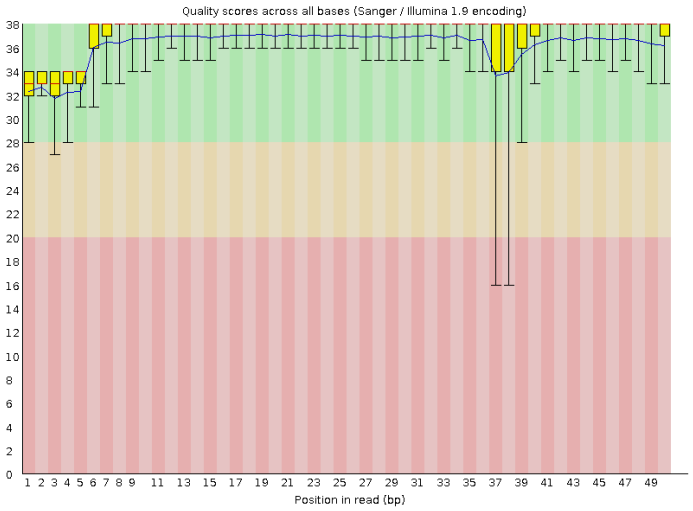

Figura 1: Per base sequence quality

4 Control de calidad con `Bioconductor`

```
library(Rqc)
rqc(path = "G:/Bioinformàtica/Curs 2022-2023/2n semestre/Anàlisi de dades òmiques/PAC1/Analisis_de_datos
    pattern = ".fastq")
```