Análisis de datos ómicos - Primera prueba de evaluación continua

Aràntzazu Alonso Carrasco

18 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 obtenidas a partir del trascriptoma completo de *Pseudomonas aeruginosa*. Se trata de muestras de RNA-seq que son *paired-end*. Sabiendo esto, podemos pasar a hacer un control de calidad del archivo con diferentes herramientas. Para este informe usaremos las herramientas `Galxay` y el paquete `Rqc` de `Bioconductor`.

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`.

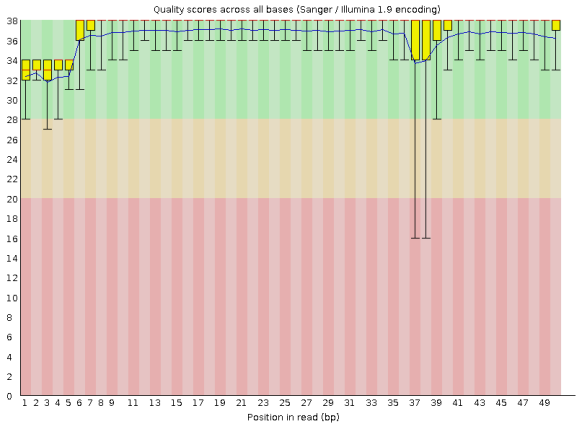

Figura 1: Per base sequence quality

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 fichero 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 superior 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. Sin embargo, podemos destacar que en la posición 37-39 del *read* hay una pequeña pérdida de calidad, ya que el bigote inferior del diagrama de cajas entra dentro de la región roja, que indica una baja calidad.

De todas formas, tal como podemos ver en la Figura 2, en la que se representa la puntuación de la calidad de cada una de las secuencias analizadas, la mayor parte de las secuencias tienen una calidad muy elevada, ya que se sitúan en el pico superior izquierdo de la figura.

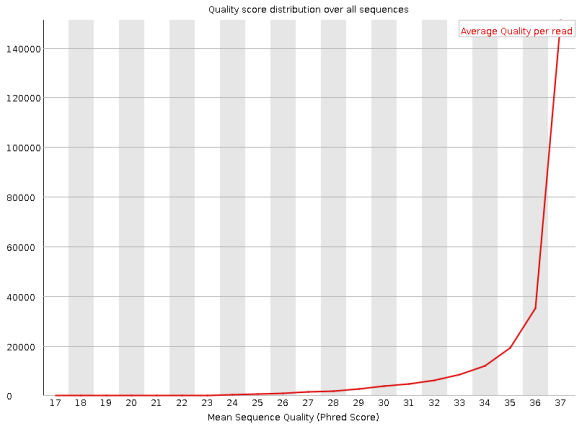

Figura 2: Per sequence quality score

Finalmente, podemos analizar el contenido de cada base nitrogenada en la secuencia. En general, esperamos encontrar poca o ninguna diferencia entre las diferentes bases de una *run*, de manera que las líneas del gráfico

deben ser relativamente paralelas entre sí. De hecho, la cantidad relativa de cada base debería reflejar la cantidad total de las bases del genoma, así que no deberíamos encontrar grandes desbalances entre ellas. Si vemos una tendencia considerable hacia una de las bases en particular, suele ocurrir por sobrerrepresentación de una secuencia que está contaminando la librería entera.

Volviendo al caso concreto de nuestros datos, en la Figura 3 podemos ver que no hay grandes desviaciones ni tendencias a favor de alguna de las bases en particular. Por tanto, parece que no hay ninguna secuencia contaminando los datos.

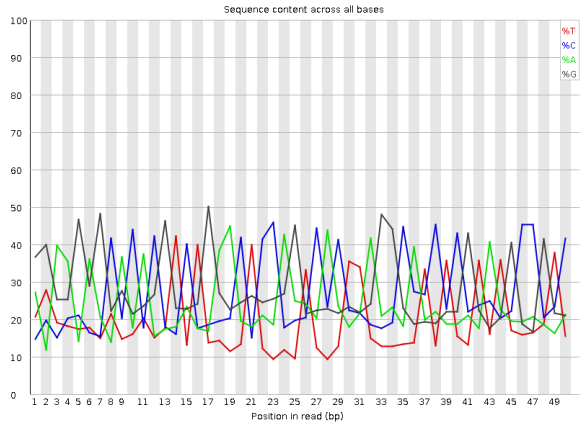

Figura 3: Per base sequence content

4 Control de calidad con `Bioconductor`

Para hacer el control de calidad con `Bioconductor`, lo primero que debemos hacer es seguir las instrucciones de instalación de la página principal. A continuación, debemos instalar y cargar el paquete `Rqc` usando los comandos `install.packages("Rqc")` y `library(Rqc)`, respectivamente. Finalmente, para realizar el análisis debemos usar el comando `rqc()` especificando la carpeta donde se encuentran los datos y el formato de los ficheros que queremos analizar. Esto generará un archivo `.htlm` con un análisis completo de los archivos indicados.

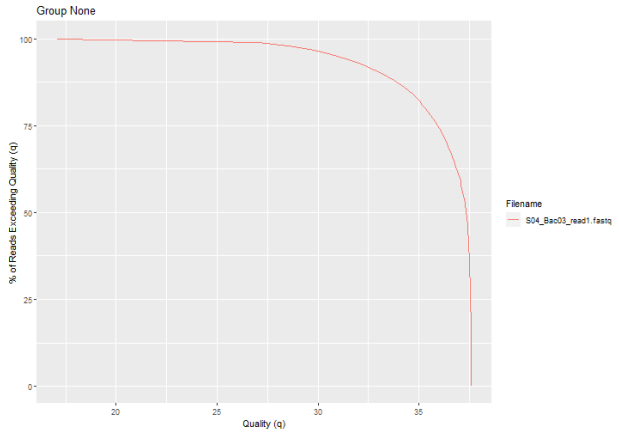

Figura 4: Average Quality

El primer gráfico que vamos a analizar del informe del control de calidad del fichero `S04_Bac03_read1.fastq` tiene por título *Average Quality* y se puede ver en la Figura 4. Este gráfico expone el patrón de calidad media

mostrando en el eje `x` los umbrales de calidad y en el eje `y` el porcentaje de lecturas que superan ese nivel de calidad. Por tanto, podemos ver que casi el 100% de las lecturas superan el umbral de calidad de 30 (Q30). Cuando la calidad de las secuencias llega a Q30, virtualmente todas las lecturas serán perfectas, sin errores o ambigüedades. Es por esto que Q30 se considera el valor de referencia para la calidad en la secuenciación de siguiente generación (NGS). Así pues, podemos asegurar que nuestros datos tienen una muy buena calidad media.

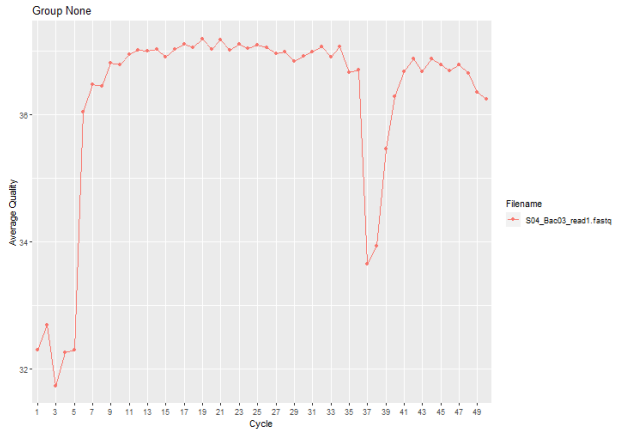

Figura 5: Cycle-specific Average Quality

No obstante, el paquete `Rqc` también nos proporciona un gráfico en el que podemos ver las puntuaciones de calidad medias para cada ciclo de secuenciación. En este caso tenemos la calidad media en el eje `y` y el número del ciclo en el eje `x` (Figura 5). Como podemos observar, los 5 primeros ciclos tienen una calidad media en torno a Q32, que es algo más baja que el Q38 que encontramos en el resto de ciclos. Excepcionalmente, en los ciclos 37 y 38, no obstante, observamos otra pequeña pérdida de calidad que se sitúa en torno a Q34. De todas formas, dado que, como hemos comentado antes, la calidad media de todas las secuencias es superior a Q30, podemos afirmar que en todos los ciclos la calidad de las secuencias es excepcionalmente elevada.

Con respecto al ciclo de secuenciación también podemos obtener la media del contenido de GC (Figura 6)

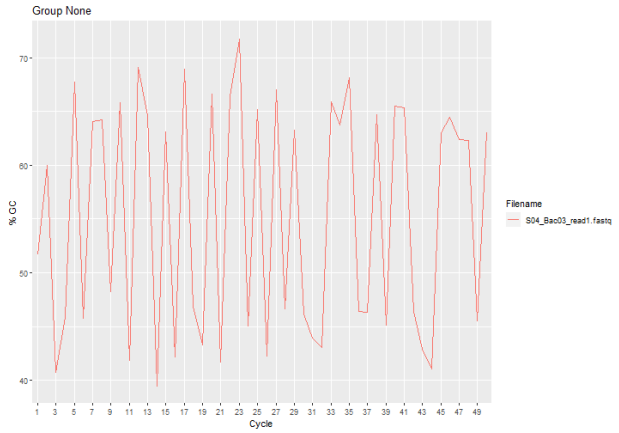

Figura 6: Cycle-specific GC Content