Análisis de datos ómicos - Primera prueba de evaluación continua

Aràntzazu Alonso Carrasco

18 de abril, 2023

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 `Galaxy` 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`.

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

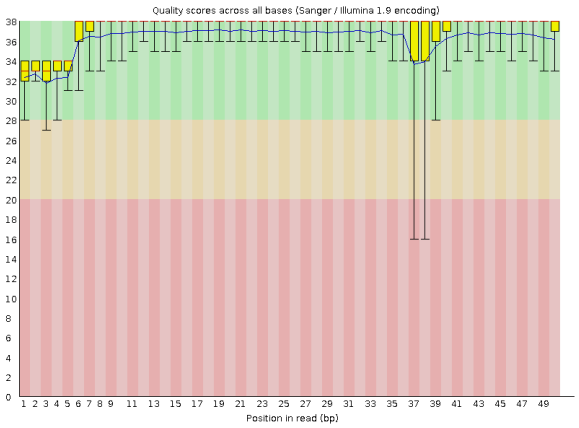

Figura 1: Per base sequence quality

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 un *Phred score* de 37. El *Phred score* es una medida de la calidad de la secuencia en la secuenciación de ADN y es una escala logarítmica que mide la probabilidad de que la base sea errónea. Cuando mayor sea el *Phred score*, mayor probabilidad de que la base sea correcta. Por ejemplo, un *Phred score* de 30 indica que la probabilidad de que la base sea incorrecta es de 1 en 1000, mientras que un *Phred score* de 40 indica que la probabilidad de que la base sea incorrecta es de 1 en 10000. Dado que el *Phred score* de la mayoría de secuencias es de 37, podemos afirmar que la probabilidad de que sean incorrectas es muy baja, es decir, la calidad es muy elevada.

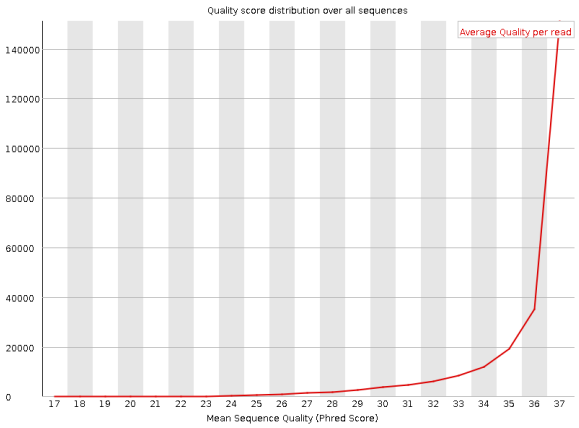

Figura 2: Per sequence quality score

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, que el informe FastQC nos indica con un error, podemos ver que no obtenemos líneas paralelas entre las bases, si no que hay grandes desviaciones entre

ellas. Por tanto, puede ser que haya secuencias contaminando los datos o que la fragmentación del ARN se haya producido de forma sesgada.

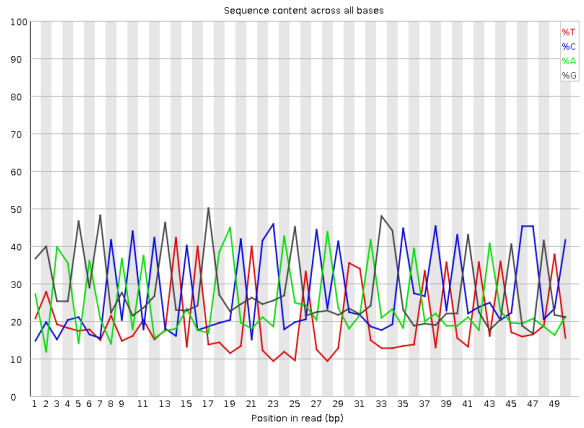

Figura 3: Per base sequence content

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.

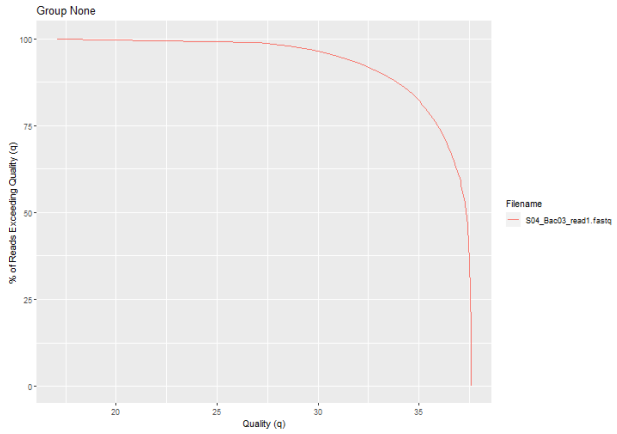

Figura 4: Average Quality

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.

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.

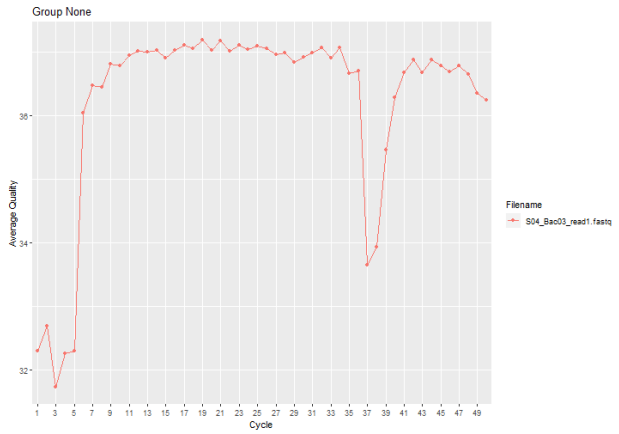

Figura 5: Cycle-specific Average Quality

Con respecto al ciclo de secuenciación también podemos obtener el porcentaje del contenido de GC (Figura 6). *A priori* esperamos que el contenido de GC sea uniforme a lo largo de los ciclos. No obstante, vemos que en nuestros datos hay una gran variabilidad de este porcentaje en función del ciclo, hecho que puede estar indicando sesgos en la secuenciación del ARN. Por tanto, sería conveniente volver a secuenciar los datos realizando los ajustes necesarios del protocolo de secuenciación.

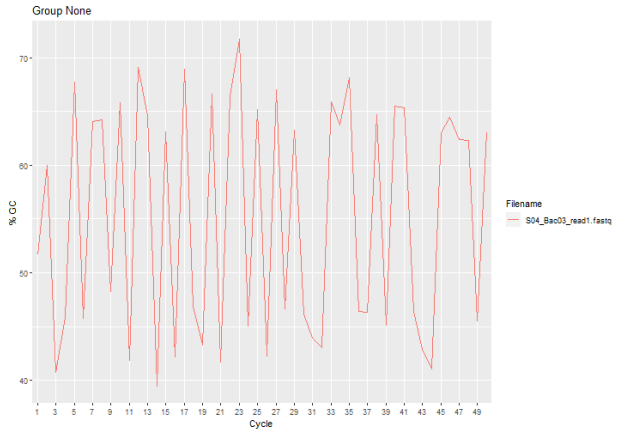

Figura 6: Cycle-specific GC Content

Otro gráfico que podemos analizar es del de la Figura 7. Como vemos, es análogo al de la Figura 1. Así pues, la interpretación también es la misma. Podemos ver que todas las secuencias tienen una puntuación mayor a Q30, hecho que indica la gran calidad de las secuencias. Sin embargo, volvemos a ver que en los ciclos 37 y 38 algunas secuencias tienen una calidad muy inferior a la del conjunto; probablemente se trata de alguna secuencia *outlier*.

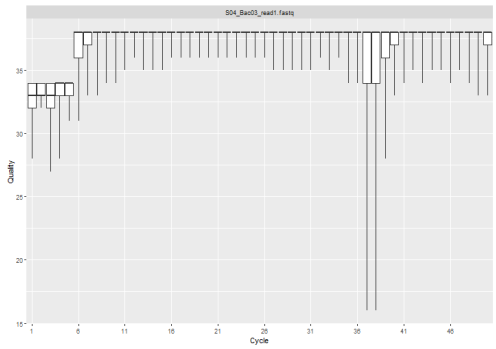

Figura 7: Cycle-specific Quality Distribution - Boxplot

Finalmente, podemos ver, tal como veíamos en la Figura 3, que usando `Bioconductor` obtenemos un gráfico muy similar (Figura 8). La gran dispersión entre las bases nitrogenadas nos hace sospechar que hay secuencias contaminando la muestra o que la fragmentación del ARN no se ha producido correctamente.

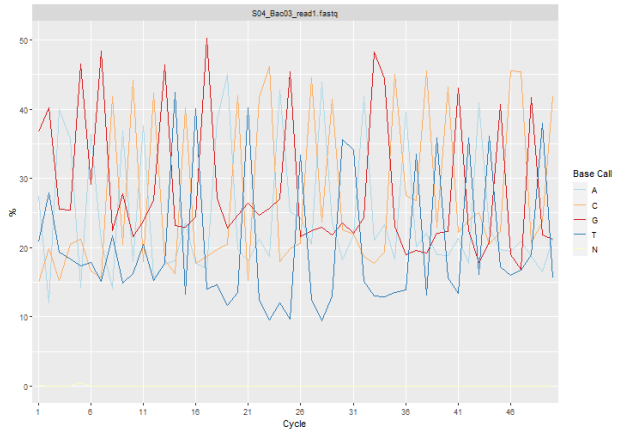

Figura 8: Cycle-specific Base Call Proportion

5 Conclusiones

Una vez hecho el control de calidad de las secuencias usando las dos herramientas, `Galaxy` y `Bioconductor`, podemos compararlas entre sí. Como hemos visto en las Figuras 1 y 7, así como en las Figuras 3 y 8, ambas herramientas nos permiten sacar las mismas conclusiones respecto a las muestras. En este caso, hemos visto que la secuenciación es de elevada calidad, dado que obtenemos puntuaciones superiores a Q30, pero hay contaminaciones o algún error en el tratamiento de las muestras, pues el contenido de GC y de bases nitrogenadas es muy dispar entre los diferentes ciclos. De hecho, los informes obtenidos con ambas herramientas contienen prácticamente los mismos gráficos y la misma información. Por tanto, ambas herramientas son igual de válidas para el análisis de resultados.

No obstante, es cierto que `Galaxy` da ciertas pistas de cómo interpretar los resultados, pues aparece un círculo verde con un tick en aquellos gráficos que muestran que la secuenciación se ha realizado con éxito, mientras que aparece un círculo rojo con una cruz en aquellos gráficos que muestran algún problema en la secuenciación. Es por esto que los resultados son más sencillos de interpretar. Podemos decir que `Galaxy` ofrece una cierta guía para aquellas personas que no somos expertas en este campo, mientras que `Bioconductor` deja en manos del investigador la interpretación del resultado.