Tarea 6.1

José Miguel Amaro Estrada

Instituto de Ecología, UNAM

Posgrado en Ciencias Biológicas, UNAM

**Introducción a la bioinformática e investigación reproducible para análisis genómicos, semestre 2020-2**

Tarea 6.1

Control de calidad de lecturas NGS

**Número de reads e información del archivo fastq**

1. Se utilizó la siguiente línea de comandos para contar el número de reads en los archivos fastq filtrados de la muestra 19:

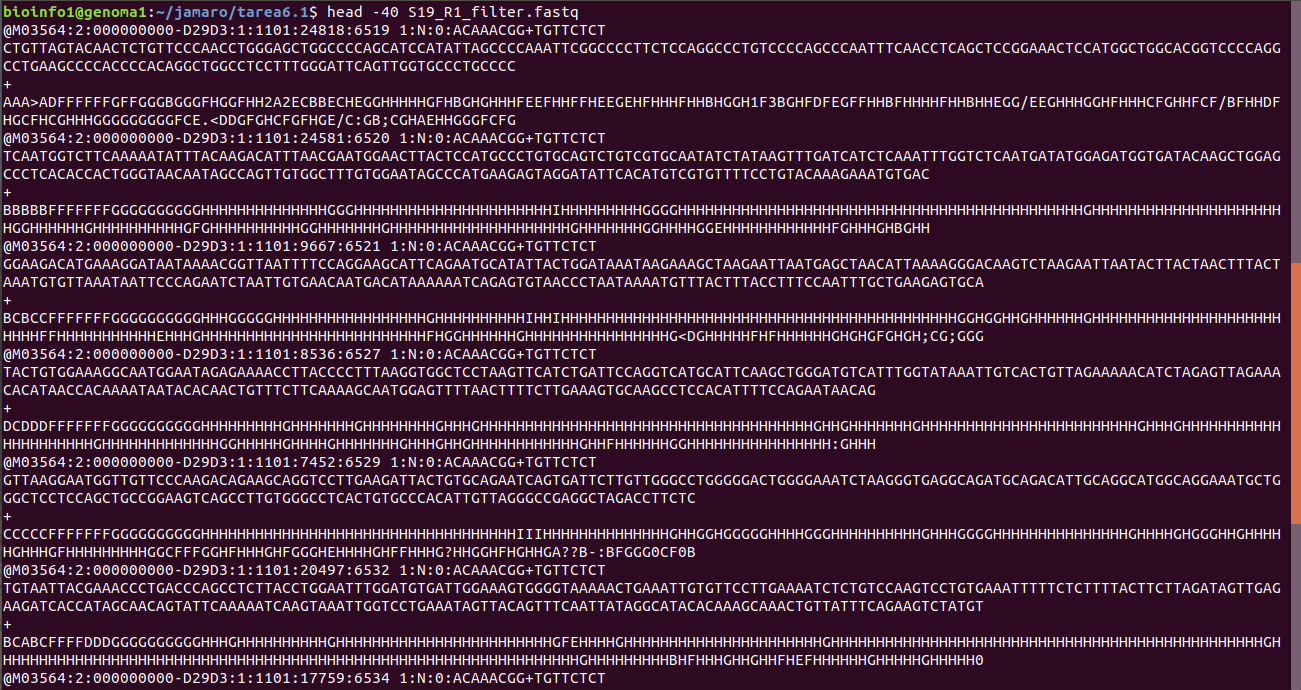
grep -o @M S19\_R1\_filter.fastq | wc

Se buscó el número de ocurrencias de los caracteres “@M”, los cuales marcan el inicio de los headers de cada uno de los reads, y por lo tanto deberían indicar el número de estos. El código arrojó un total de 20,426 ocurrencias y por lo tanto, este es el número de reads que contiene cada archivo fastq de la muestra 19. Es importante mencionar que, para poder realizar este paso, se descomprimió el archivo .gz.

1. Para la visualización de las 10 primeras lecutras, se aplicó el siguiente comando en el archivo fastq descomprimido:

head -40 S19\_R1\_filter.fastq

Es decir, se visualizaron las 40 primeras líneas del archivo fastq. A continuación se muestra el resultado de este comando y se indica la información proporcionada en el header de la tercera lectura.



Los valores de calidad de las 10 primeras bases son:

33,34,33,34,34,38,38,38,38,38

En amarillo se indica la zona donde se entrega la información sobre la calidad de la tercera lectura. Esta calidad se indica para cada base mediante un código de caracteres ASCII, cada uno de los cuales se corresponde con un valor particular de calidad. Cada uno de los caracteres ASCII corresponde a la base que se encuentra por arriba de él, es decir, el primer carácter ASCII corresponde a la calidad de la primera base, etc.

En anaranjado se resalta el identificador de la tercera lectura. Esta contiene los siguientes rubros de información, separado cada uno por dos puntos:

1. El número de serie del secuenciador que se utilizó
2. El número de corrida a la que corresponde la lectura en el secuenciador.
3. Idenficador del flowcell.
4. Número del lane
5. Número del tile
6. Las coordenadas dentro del tile (x,y)
7. La información de la lectura, incluido el número (si es lectura sencilla o pareada), si la lectura fue filtrada o no, y las secuencias de sus índices de secuenciación.

**Reporte de calidad de la secuenciación**

1. **Lecturas no filtradas**

De acuerdo con el reporte de fastqc, los archivos crudos contienen un total de 25,481 lecturas.

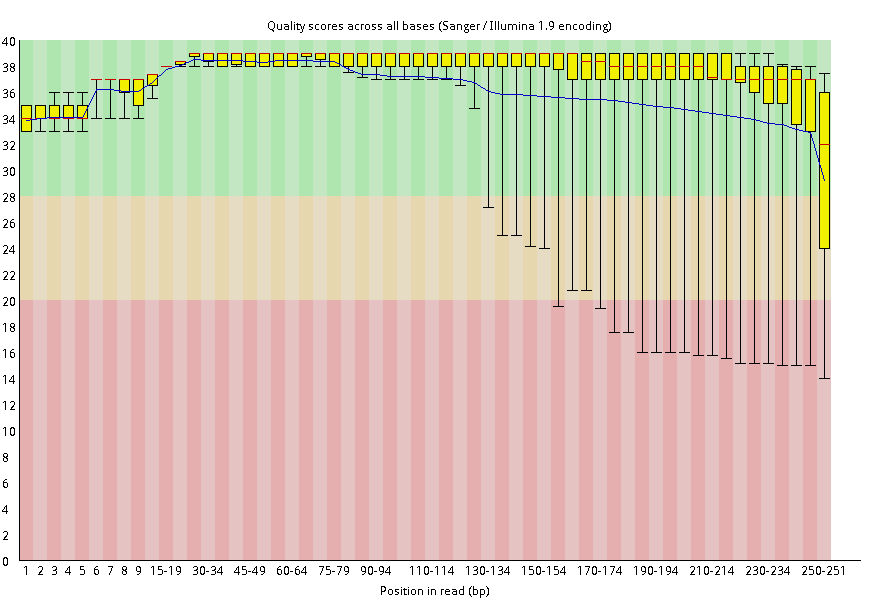
****

Figura 1. Calidad por base de las lecturas R1 sin filtrar de la muestra 19.

Como se observa en las imágenes 1 y 2, la calidad de las lecturas R1 y R2 sin filtrar tiene altos niveles de calidad durante la primera mitad de su longitud y la distribución de los niveles de calidad es homogénea. Por otro lado, la calidad se va volviendo más heterogénea, con niveles menores hacia el final de las secuencias. Esto es normal, y puede deberse a diferentes aspectos del proceso de secuenciación. En general, las calidades de las lecturas R1 y R2 son distintas hacia la segunda mitad del proceso de secuenciación, con las lecturas R2 mostrando una mayor proporción de lecturas con menor calidad comparada con las lecturas R1.

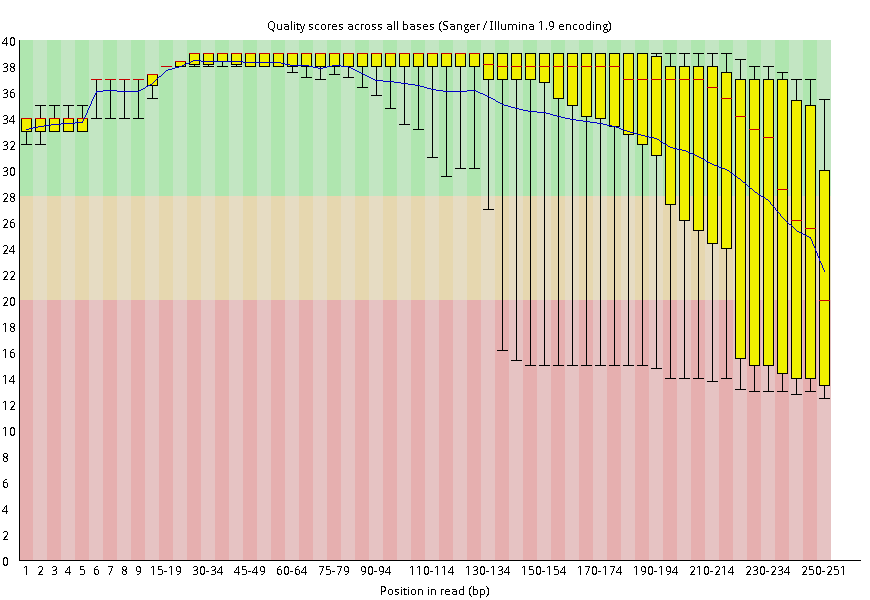
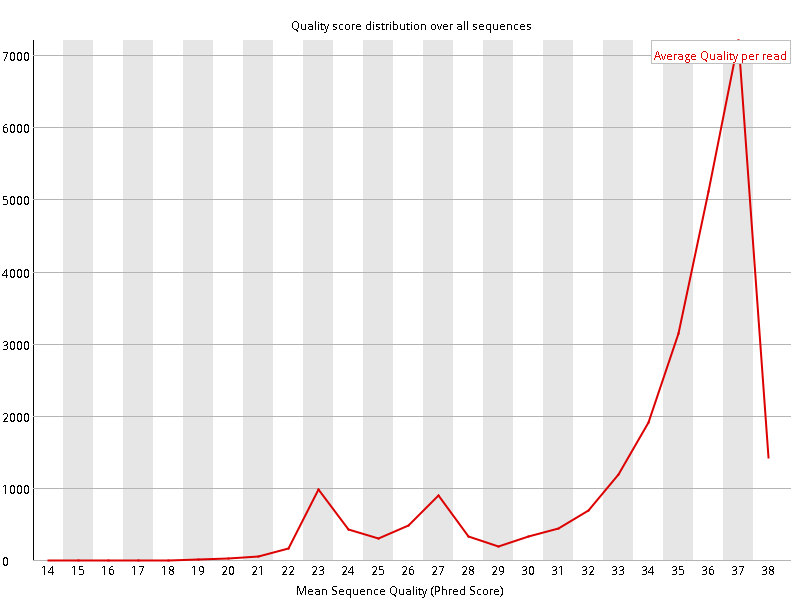
****

Figura 2. Calidad por base de las lecturas R2 sin filtrar de la muestra 19.

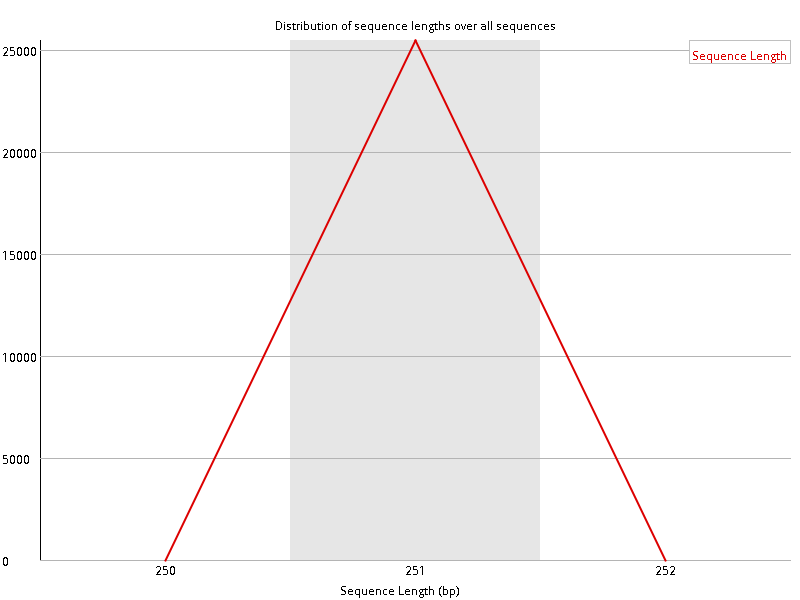
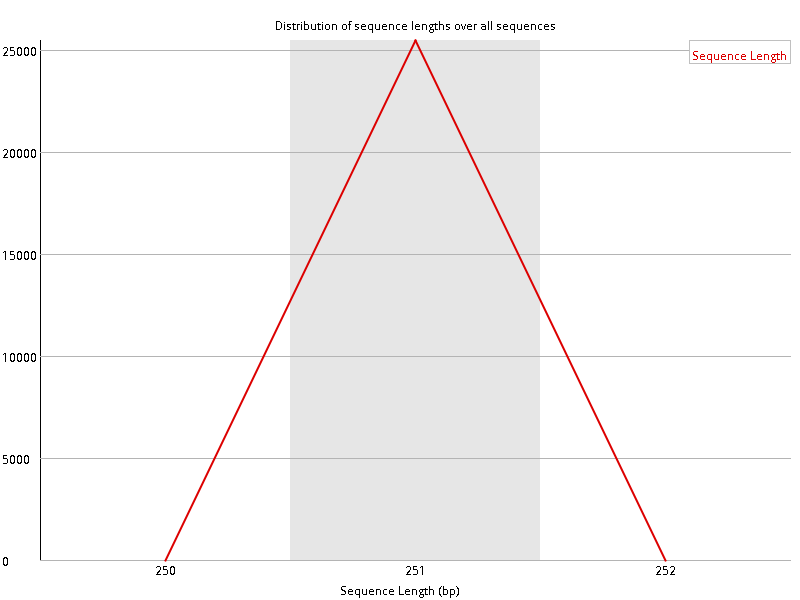
Esta diferencia entre la calidad de las lecturas R1 y R2 también puede apreciarse en la proporción de lecturas con distintos niveles de calidad. Como se observa en las figuras 3 y 4, las R2 contienen picos más altos en las lecturas con calidad 23 y 27, en comparación con las R1 que tiene una menor proporción de lecturas en estos picos. En ambos casos, el resto de las secuencias se encuentran en un pico con calidades por encima de 30.

En cuando a la longitud de las lecturas, puede observarse que tanto las lecturas R1 y R2 presentan longitudes iguales, de 251 bp (figuras 5 y 6). En cuanto a la representación de las bases en las lecturas, en ambos casos 8R1 y R2), estas se encuentran en proporciones equivalente, lo cual sugiere que no existe ningún tipo de contaminación o que no hay fragmentos sobre-representados (figs 7 y 8). Sin embargo, se observa una diferencia de proporción entre A’s y G’s con respecto las T’s y C’s hacia el final de las secuencias en las lecturas R2, lo cual quizá esté relacionado con las diferencias en calidad en estas lecturas hacia el final.

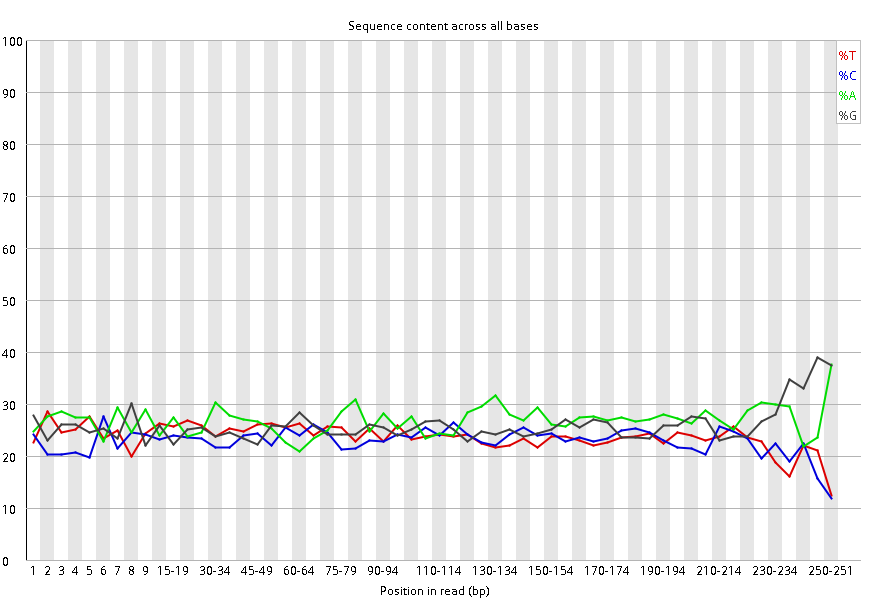
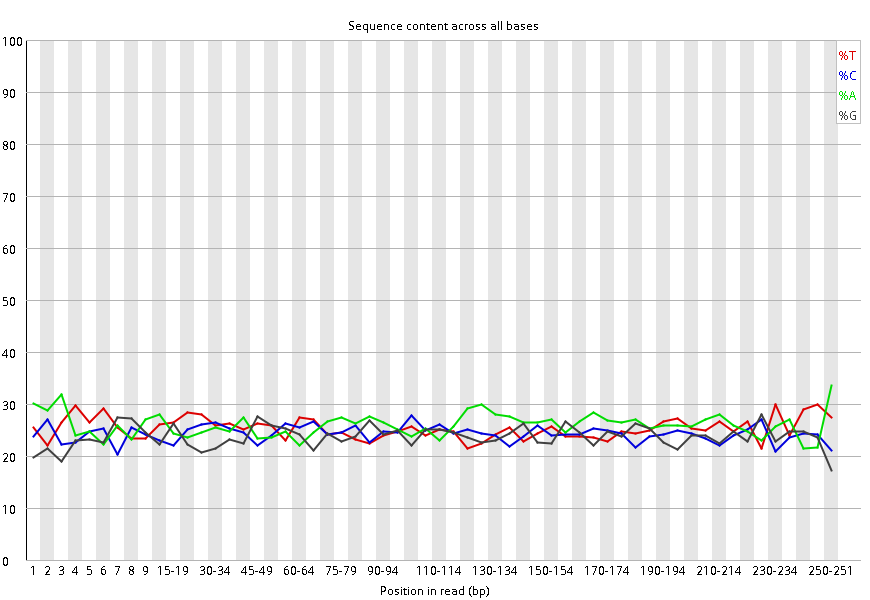
**Imagen que contiene competencia de atletismo

Descripción generada automáticamente **

Figuras 3 y 4. Distribución de las calidades a nivel de sequencia para las lecturas R1 (izquierda) y R2 (derecha).

****

Figuras 5 y 6. Distribución de los tamaños de las lecturas para las R1 (izquierda) y R2 (derecha).

****

Figuras 7 y 8. Contenido de bases a lo largo de las secuencias para las lecturas R1 (izquierda) y R2 (derecha).

1. **Lecturas filtradas**

De acuerdo con el reporte fastqc de las lecturas filtradas, los archivos contienen un total de 20,426 lecturas, lo cual coincide con lo encontrado utilizando los comandos Unix en el paso inicial de este reporte. Es decir, durante el proceso de filtrado, se eliminó un total de 5055 lecturas. Después del proceso de filtrado de las lecturas, se observa que en general hay una homogeneización en los valores de calidad, los cuales se mantienen más uniformes a lo largo de las secuencias. Los menores niveles de calidad solo decaen y se amplifican en los últimos ciclos de la secuenciación. Sin embargo, como se observa en la figura 2, se mantiene la tendencia de las lecturas R2 a tener una mayor heterogeneidad en sus valores de calidad comparadas con las lecturas R1 (figura 10).

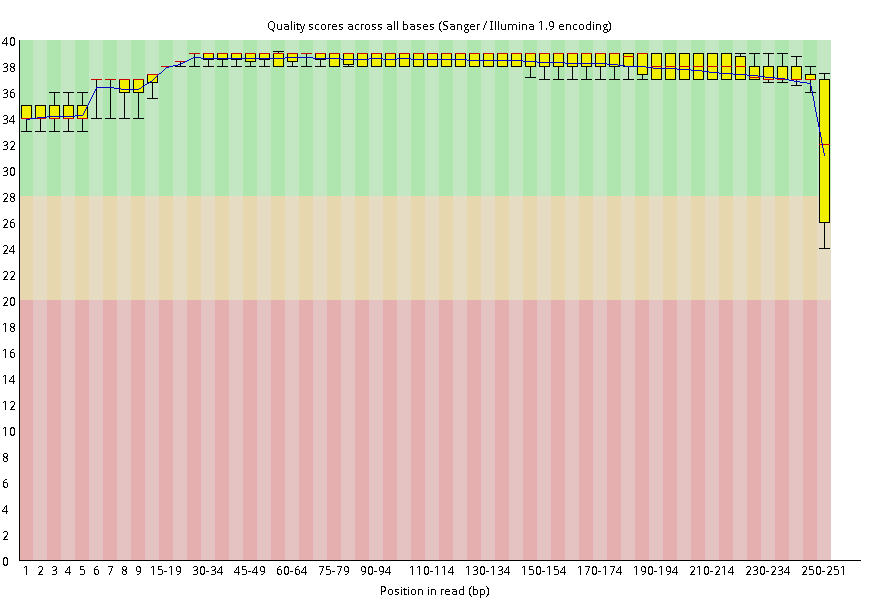


Figura 9. Calidad por base de las lecturas R1 filtradas de la muestra 19

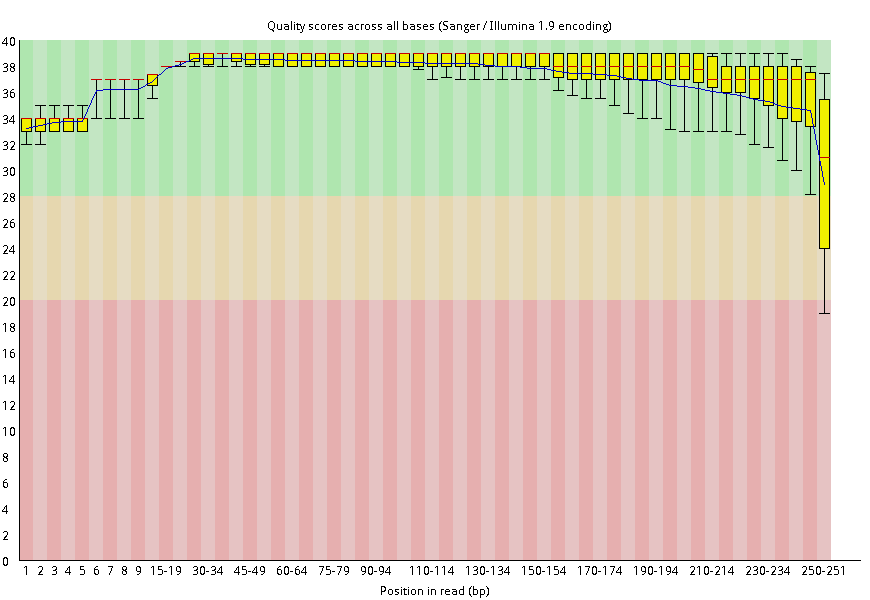
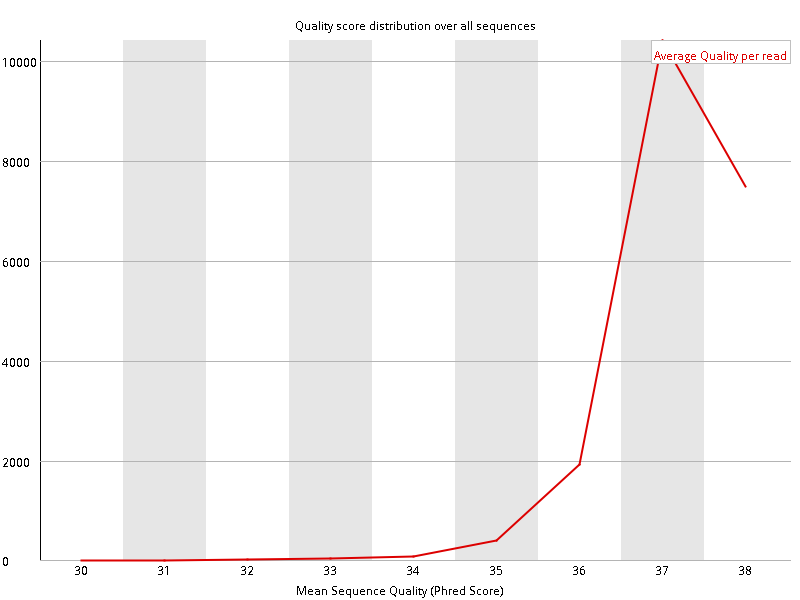
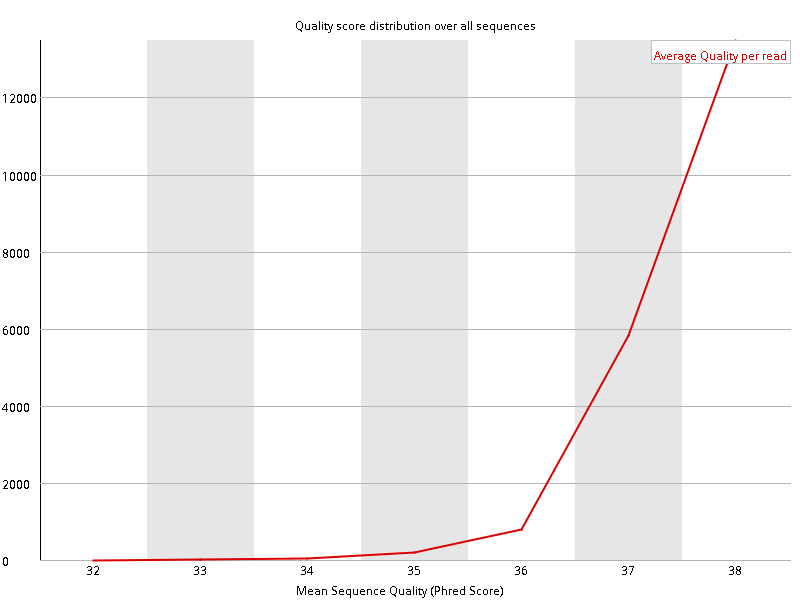


Figura 10. Calidad por base de las lecturas R2 filtradas de la muestra 19

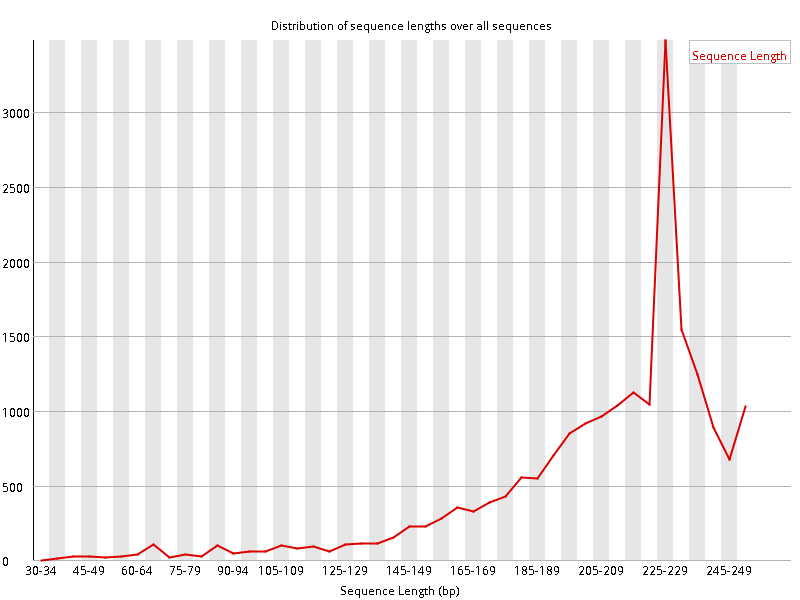
En cuando a la distribución de las calidades entre secuencias, se observa que después del filtrado, se eliminaron los dos picos observados de menor calidad en las secuencias crudas, y la mayoría de las lecturas conservadas se encuentran dentro de niveles de calidad por encima de 30 tanto en las R1 como en las R2 (figuras 11 y 12).



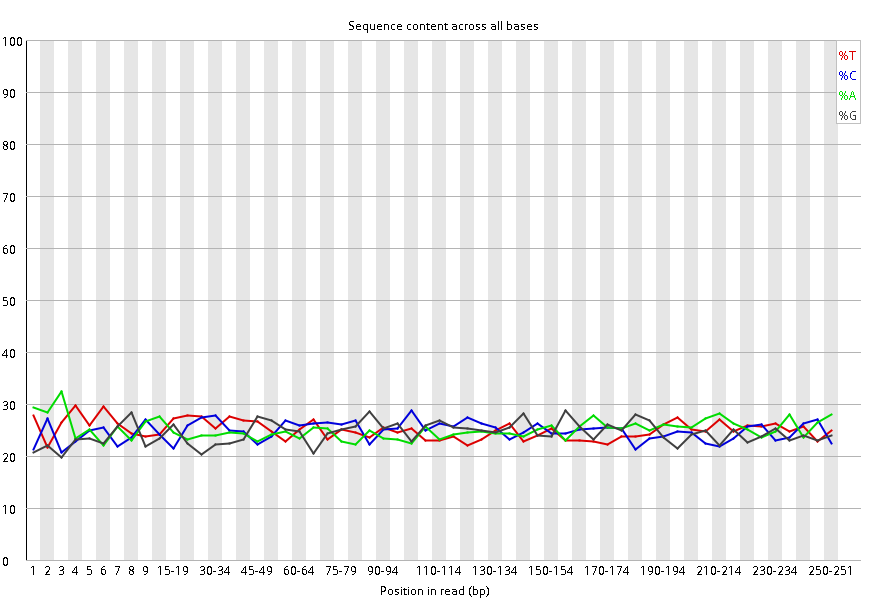
Figuras 11 y 12. Distribución de las calidades a nivel de sequencia para las lecturas R1 (izquierda) y R2 (derecha) después del filtrado.

La distribución de la longitud de las secuencias posterior al filtrado presenta ahora sus mayores picos alrededor de los 225-229 bp (figuras 13 y 14). Esto quizá debido a que esa es la longitud real de los fragmentos después de eliminar las secuencias de los adaptadores. Sin embargo, se observa una mayor proporción de fragmentos menores que 225 bp en las lecturas R2. Esto puede estar relacionado con la mayor proporción de lecturas con menor calidad que se encontró en estas lecturas, lo cual podría reflejarse en lecturas incomplentas o la presencia de otro tipo de fragmentos, los cuales presentan distintos tamaños. En cuando al contenido de bases en las secuencias, este se sigue manteniendo homogéneo a lo largo de las secuencias, excepto para las lecturas R2, donde se sigue observando una mayor proporción de A’s hacia el final de las lecturas (figuras 15 y 16).

Imagen que contiene biombo, competencia de atletismo

Descripción generada automáticamente

Figuras 13 y 14. Distribución de los tamaños de las lecturas para las R1 (izquierda) y R2 (derecha) después del filtrado.

Imagen que contiene reloj

Descripción generada automáticamente

Figuras 15 y 16. Contenido de bases a lo largo de las secuencias para las lecturas R1 (izquierda) y R2 (derecha) después del filtrado.