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 6.2

Filtrado y alineamiento de lecturas NGS

1. Filtrado de los reads

Se aplicaron tres filtros a las lecturas usando el paquete bbduk. El primer filtró se aplicó para eliminar las secuencias de los adaptadores de secuenciación. El segundo, para eliminar las secuencias del fago phix. El tercero, para eliminar las lecturas de baja calidad, conservando solamente aquellas con valores mínimos de Q20 y una longitud mínima de 30 bp.

1. Alineamiento de las lecturas

Posteriormente se alinearon las lecturas contra un genoma de referencia, utilizando el alineador bwa. Posteriormente se convirtió el archivo SAM resultante del alineamiento, en un archivo BAM (binario). Para esto se utilizó el programa Picard. Con este mismo programa, se ordenaron las lecturas de acuerdo con su posición, y se agregó el ReadGroup al archivo BAM. Finalmente, utilizando SAMTools, se indexó el alineamiento.

1. Reporte de calidad

Finalmente se evaluó la calidad del alineamiento utilizando el programa Qualimap.

**Resultado**

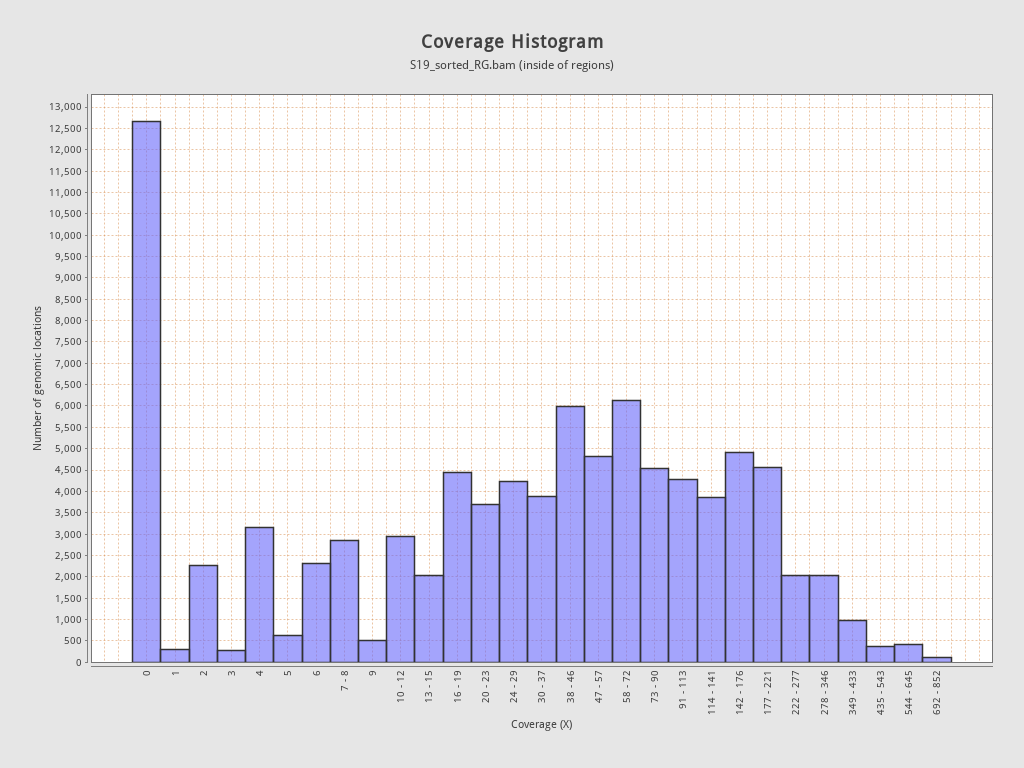


Figura . Histograma del número de regiones genómicas con una determinada cobertura.

De acuerdo con el análisis de calidad, la profundidad de la secuenciación en las regiones blanco es altamente variable, con muchas lecturas representando una cobertura de cero. Sin embargo, la mayoría de las lecturas se concentran en profundidades de entre 10X y 200X en las regiones blanco (figura 1).

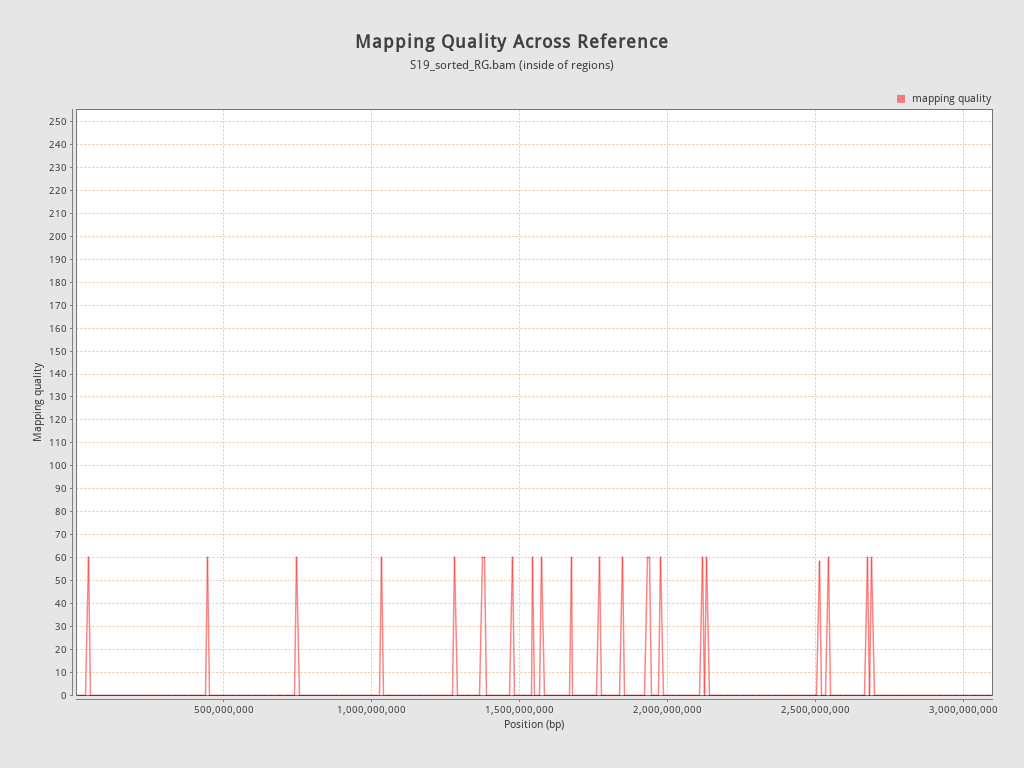


Figura . Distribución de las calidades de mapeo en las regiones genómicas

En cuando a la calidad de mapeo, se puede observar en la figura 2 que, para todas las regiones blanco, se obtuvieron valores de 60, lo cual representa una tasa de error de 1 por cada millón y, por lo tanto, puede considerarse una buena calidad de alineamiento.

Por otro lado, el análisis de qualimap indica que la mayoría de las lecturas mapeadas tiene un tamaño de entre 220 y 270 bp (figura 3).

Finalmente, el contenido de bases por cada posición mapeada se observa relativamente uniforme (figura 4).

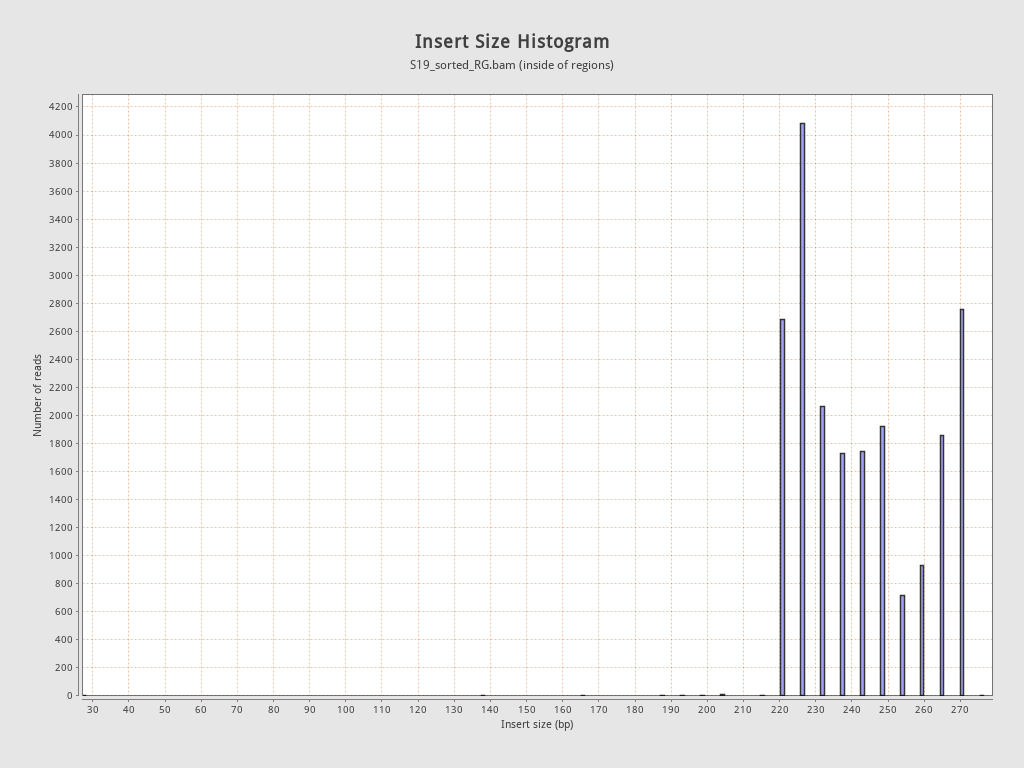


Figura . Distribución de los tamaños de los insertos mapeados.

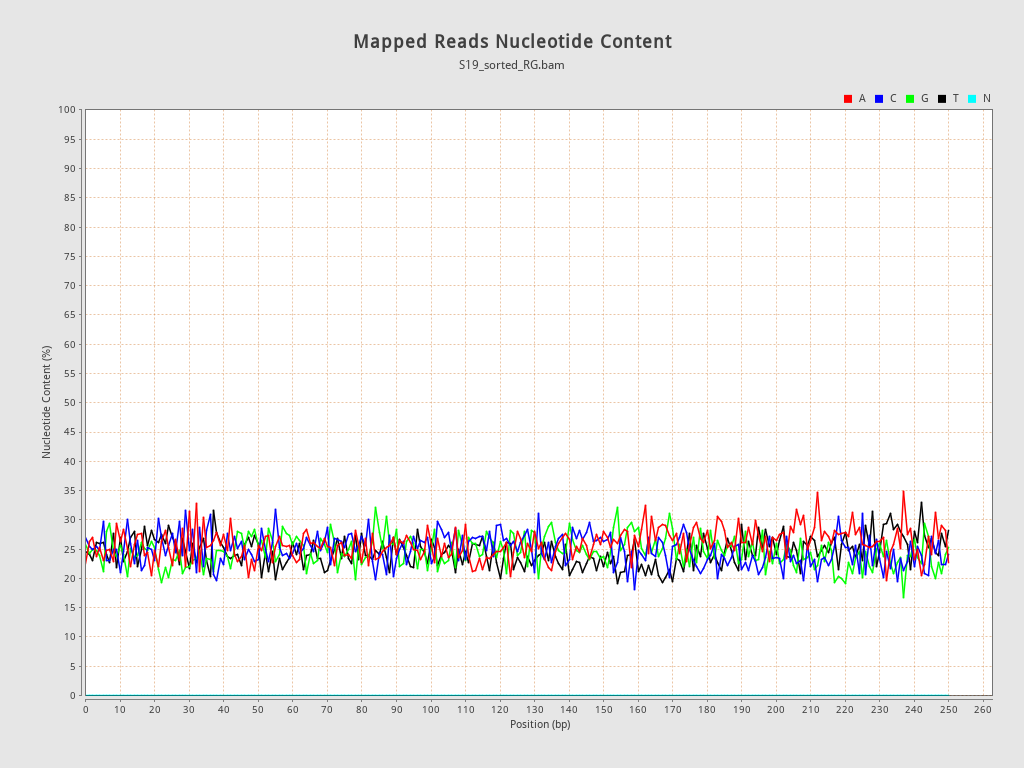


Figura . Contenido de bases por cada posición mapeada

**Conclusión**

La alta proporción de sitios con cobertura entre 1 y 0 quizá se deba al número de regiones genómicas que no contenían regiones blanco para el alineamiento y que por lo tanto no fueron cubiertas por ninguna lectura. Sin embargo, esto no quiere decir que el alineamiento sea malo, ya que como se mencionó en los resultados, todas las regiones mapeadas tienen una calidad de 60, y adicionalmente, según el reporte de qualimap, el 99.96% de las lecturas pudo mapearse exitosamente (ver Anexo).

Por otro lado, el tamaño de los insertos mapeados según qualimap, corresponde con lo previamente encontrado durante el proceso de análisis de calidad y filtrado de las lecturas. El contenido de bases en los sitios mapeados también sugiere que no existen secuencias sobrerrepresentadas ni enriquecidas.