César Emiliano Escalona Prado

Tarea 1 Unidad 6

Los incisos 1 a 3 de la tarea están adjuntos en un .md intitulado “ CEEP6\_1ClMx2020”

4. Informe de calidad fastqc para las lecturas R1 y R2.

S16\_R1) La calidad de las secuencias se encuentra por encima de 36 en la mayoría de los casos, a pesar de su calidad, se nota que a partir de la mitad de las lecturas la calidad se ve mermada paulatinamente. Existe un numero amplio de secuencias repetidas, así pues en función al contenido GC se sugiere que existe una posible contaminación en las librerías

S16\_R1\_filter) El filtro homogeneiza la calidad calidad de las bases, sin embargo al observar el contenido GC respecto a la distribución normal sugiere que es posible que exista contaminación pese a la aplicación del filtro. De igual manera el filtro resulta poco eficaz para deshacerse del exceso de duplicados.

S16\_R2) La calidad de la secuencia base por base se mantiene arriba de 32, a partir de la segunda mitad del las bases se disminuye hasta llegar a valores cercanos a 20. El error en el contenido de GC es más del 30% de las lecturas lo que sugiere contaminación en la librería. La diferencias significativa (20%) entre A-T y G-C, sugiere una sobre representación en alguna de las secuencias o un sesgo en la secuenciación

S16\_R2\_filter) El filtro homogeniza la calidad de la secuencia de manera considerable elevando los valores de calidad por arroba de 28 tanto para la media como par la mediana, el error en el contenido de GC se mantiene por arriba del 30% de las lecturas, sin embargo la diferencias A-T y G-C se disminuyen considerablemente, lo que indica que el filtro disminuyó sesgo en las secuencias.

5. Valores calculados en el punto 1 respecto a lo entregado en el informe de calidad FastQc

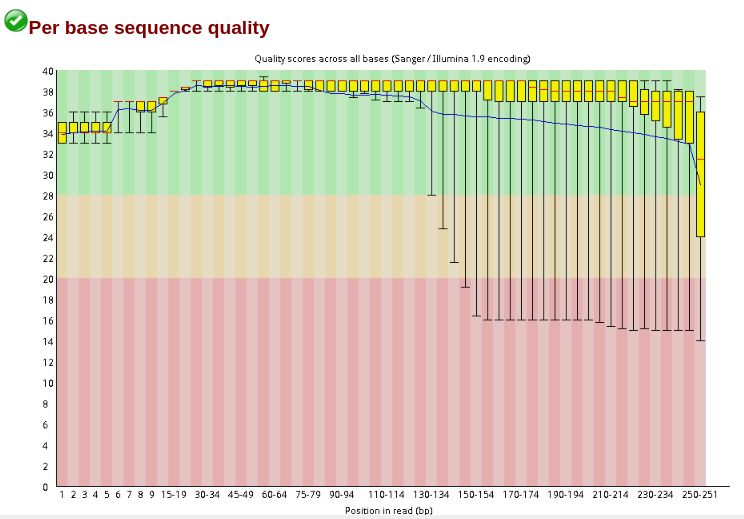
En cuanto a la comparación respecto de la información obtenida en los informes de calidad FastQc con respecto a la generada a partir de lineas de comando (codigo unix), la mayor ventaja de los informes es que te dan un acceso más visual a la información contenida en los archivos fastq.

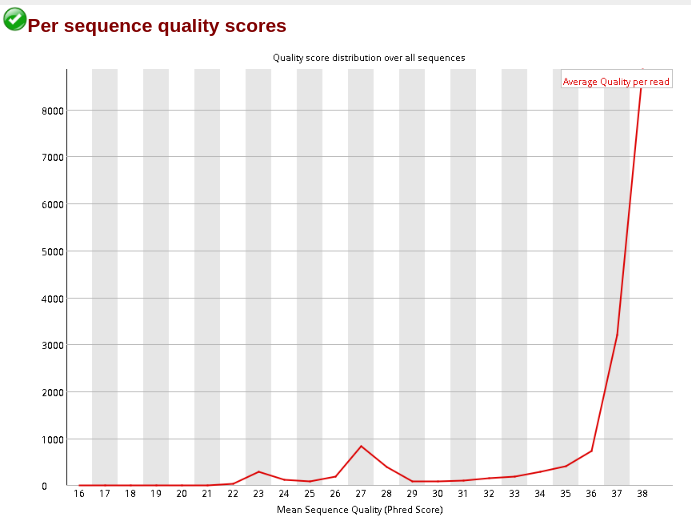
Por ejemplo, es posible observar ciertas tendencias respecto a los valores de calidad por base en cada una de las lecturas mediante la aplicación de comandos unix, sin embargo el soporte grafico que genera el informe facilita de manera sustancial la noción de calidad respecto a las secuencias y facilita el análisis de los datos de manera considerable.

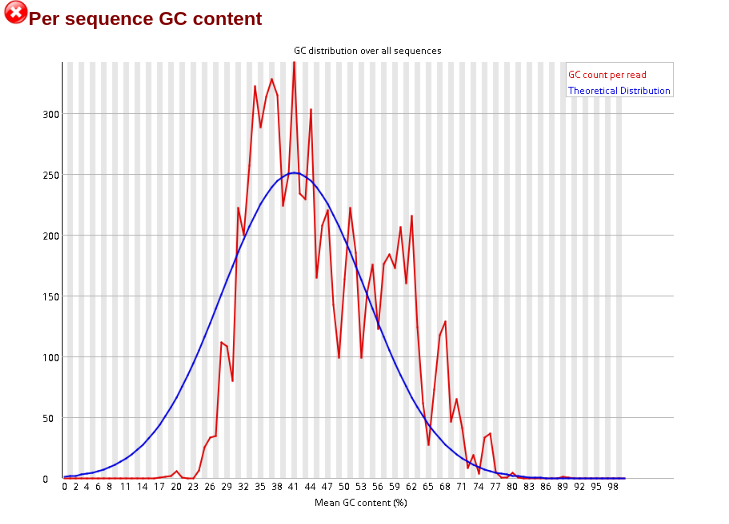
En resumen, la linea de comandos puede ser de utilidad para conocer a bote pronto algunas medidas, como por ejmplo el número de reads, o bien para familiarizarse con el funcionamiento de programas como FastQc, sin embargo, en la práctica, el uso de herramientas que procesan/decodifican información y la expresan de forma gráfica es sumamente apreciable.

6. Figuras (4) más importantes para calidad de la corrida

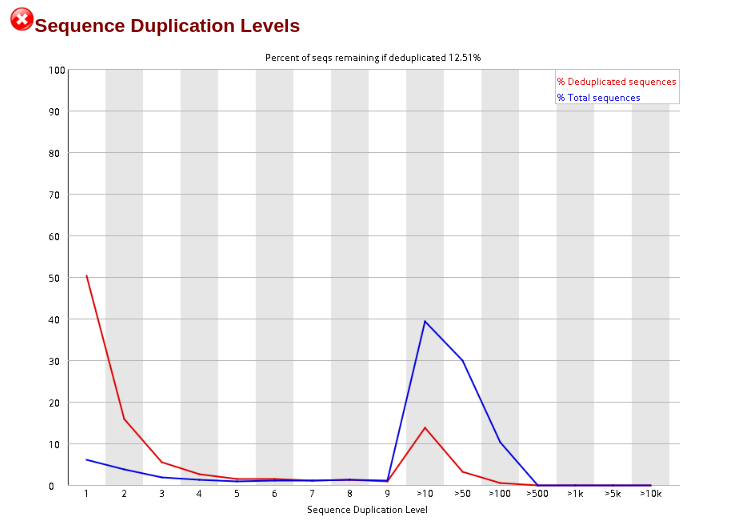
S16\_R1

La imagen muestra la calidad de la secuencia base por base, según muestra la linea roja (mediana) y la azul (media), a partir de la segunda mitad del las bases se disminuye paulatinamente la calidad, y en las ultimas bases (las barras amarillas) se muestra que se amplía el rango intercultural, lo que hace suponer dispersión y menor calidad.

En éste caso se ve que la mayoría de las lecturas tiene una calidad por encima de 36, en cuanto al pico bajo cercano a 27 probablemente responde burbujas o residuos en la célula de flujo.

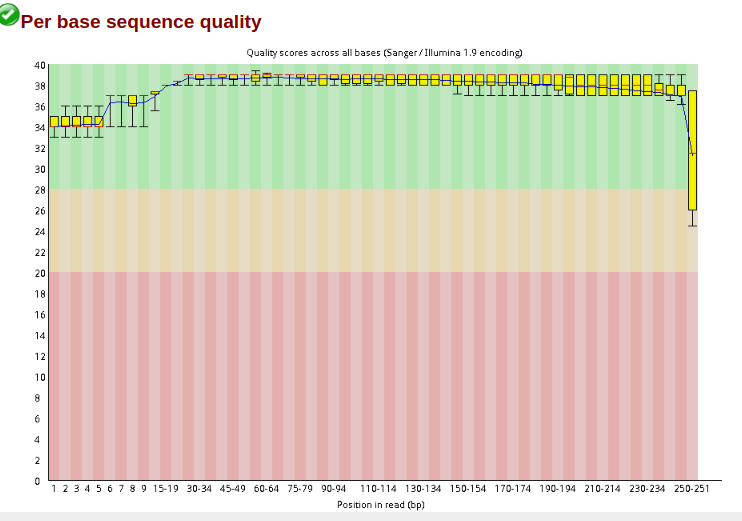
En éste caso lo esperado es una distribución normal, los picos en rojo podrían sugerir algún tipo de contaminación en la librería.

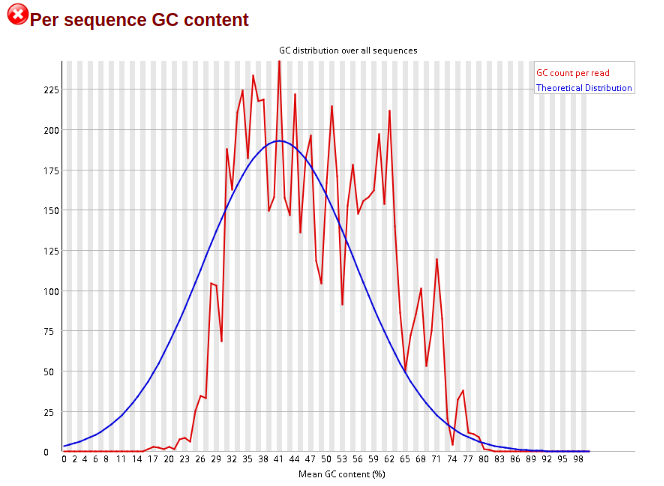
Error en más del 30% de las lecturas

Hay un alto numero de secuencias no únicas (más del 50% de la total), El error marca que hay secuencias no únicas que constituyen más del 50% de la total.

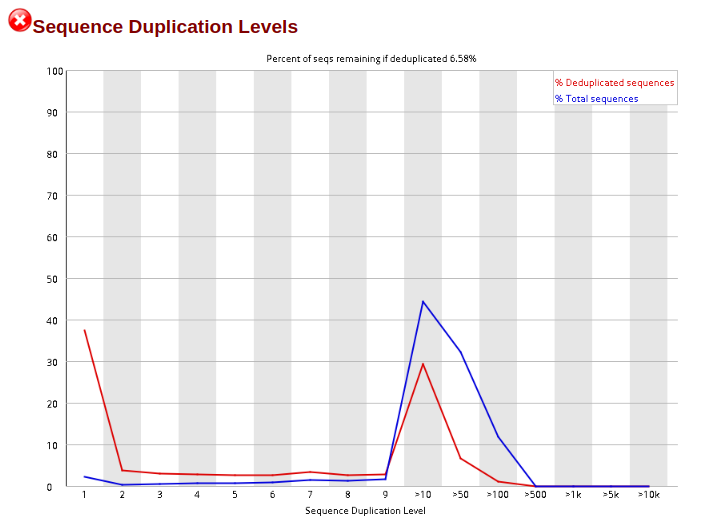
40% de las secuencias se repiten más de 10 veces y 30% más de 50 veces.

S16\_R1\_filter

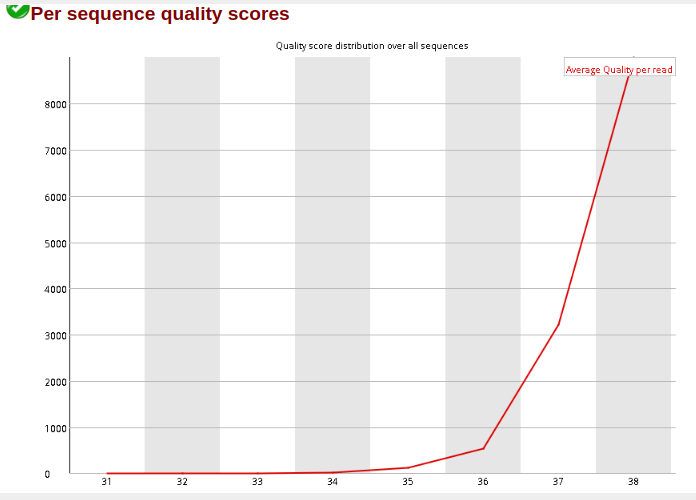
En este caso , el filtro permite que la calidad media y mediana se mantenga, salvo en en la ultima base, donde (las barras amarillas) se muestra que se amplía el rango intercultural, y que decaen las lineas azul y roja, lo que hace suponer dispersión y menor calidad en dicha base en específico.

Pese a la aplicación de filtro, los picos en rojo mantienen el indicativo de contaminación en la librería.

Se mantiene error en más del 30% de las lecturas.

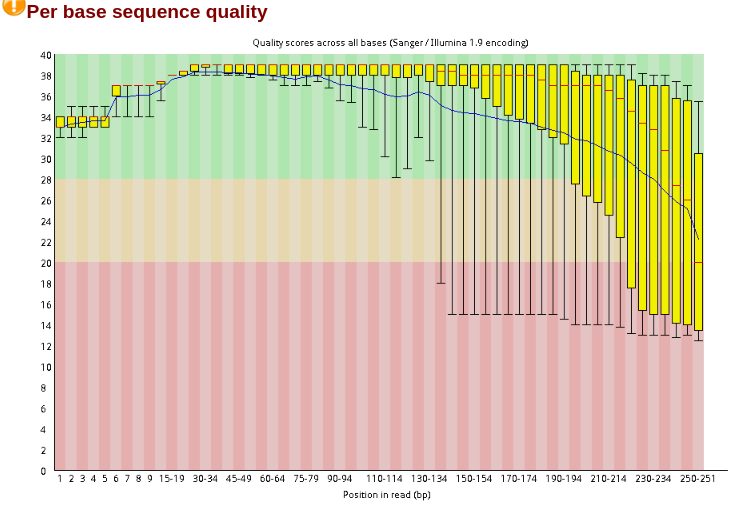


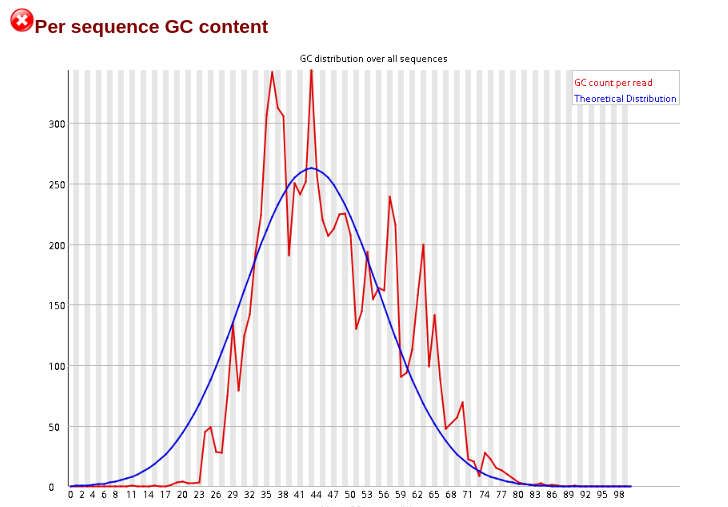
Se mantiene el error pese al filtro, secuencias no únicas que constituyen más del 50% de la total.



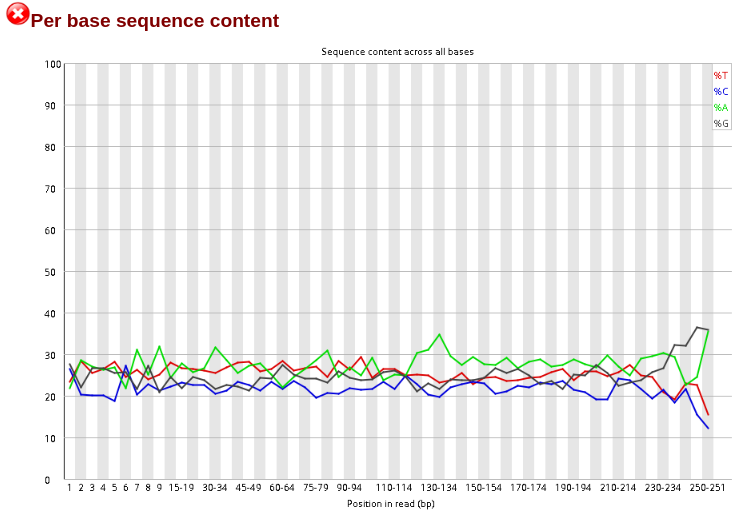
A partir del filtro se observa que se elimina el error (pico bajo cercano a 27) observado en el informe sin filtro.

S16\_R2

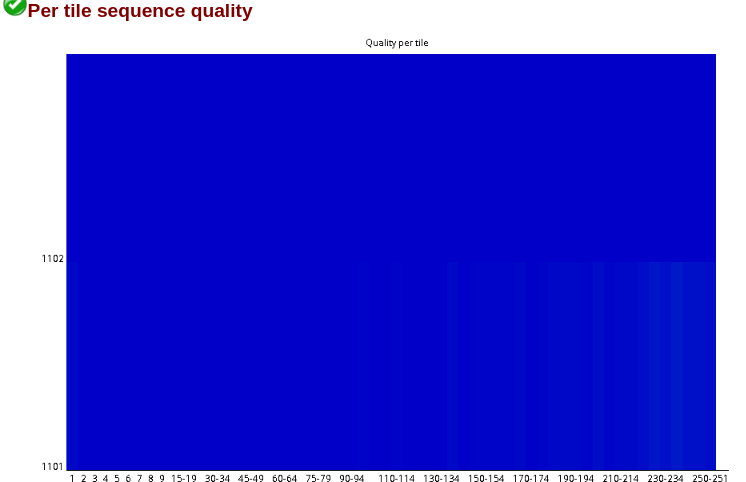
La calidad de la secuencia base por base se mantiene arriba de 32 según muestra la linea roja (mediana) y la azul (media), a partir de la segunda mitad del las bases se disminuye paulatinamente la calidad, hasta llegar a valores cercanos a 20.



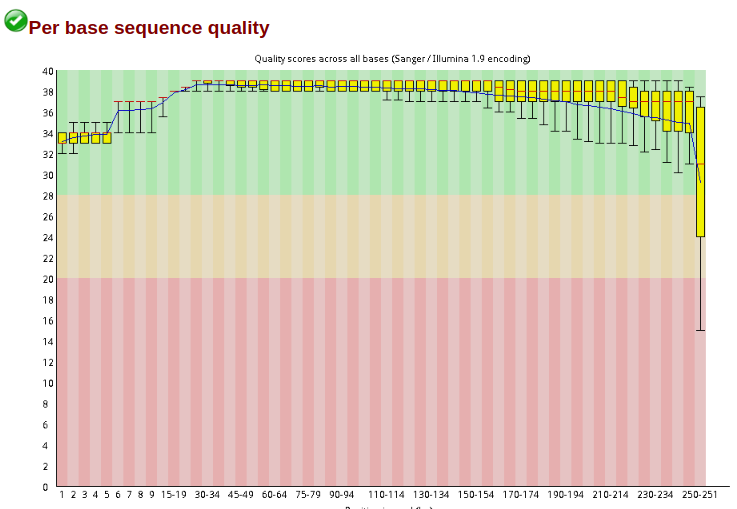
El error en más del 30% de las lecturas sugiere que existe algún error, posiblemente contaminación en la librería.

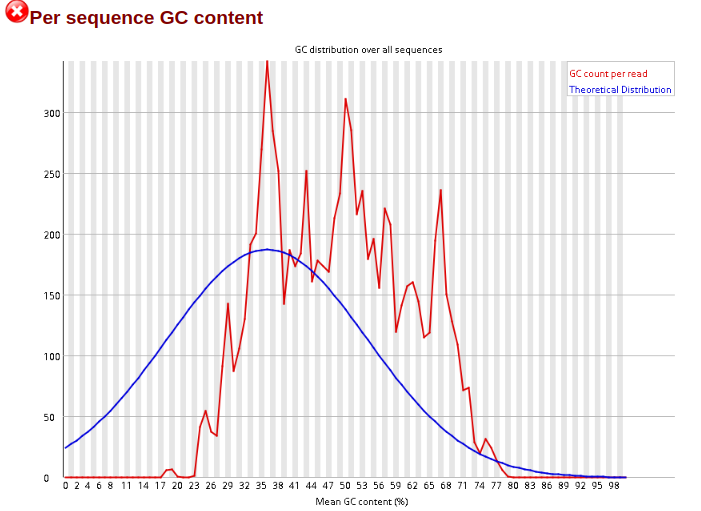


En este caso se observa que la librería muestra diferencias significativas (20%) entre A-T y G-C, lo que sugiere una sobre representación en alguna de las secuencias o un sesgo en la secuenciación

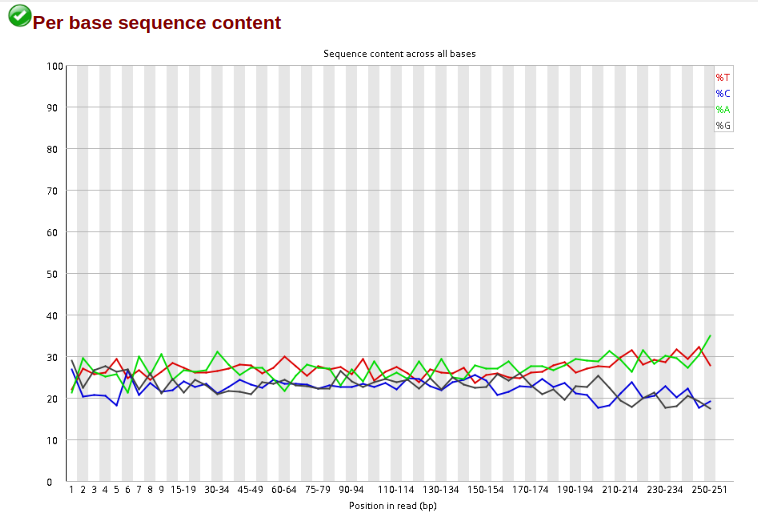
Se muestra el mosaico de la celda de flujo de donde viene cada lectura, los colores fríos del mosaico representan que en dichas posiciones la calidad es superior al promedio de esa base en ejecución (para ilumina). En mosaicos fríos como el presente hacen suponer que los errores por burbujas o referidos a las celdas de flujo son despreciables.

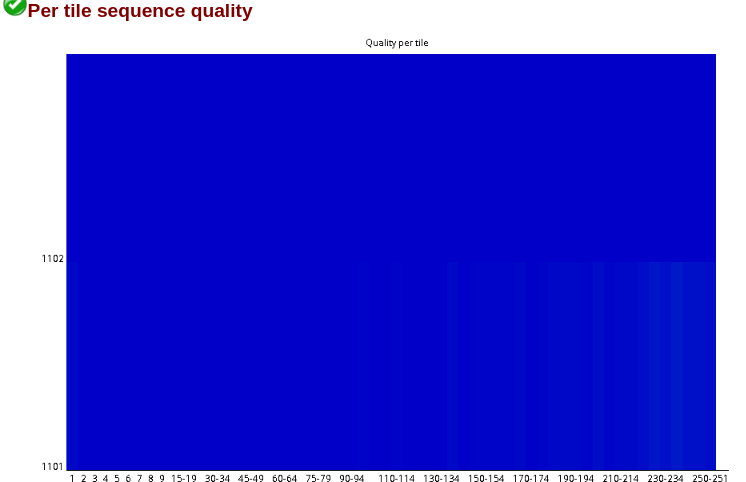
S16\_R2\_filter

El filtro corrige de manera sustancial la calidad de las secuencias, así mismo homogeneiza la mayoría de esta excepción de la ultima lectura.



Se mantiene error en más del 30% de las lecturas a pesar de la aplicación del filtro.

A partir de la aplicación del filtro se reduce la diferencia entre A-T y G-C (menor al 10%)

En mosaicos fríos como el presente hacen suponer que los errores por burbujas o referidos a las celdas de flujo son despreciables.