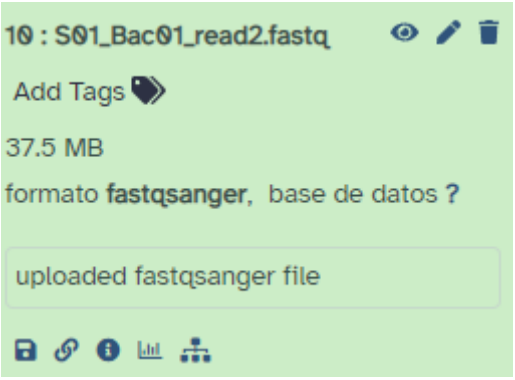
PEC 1 LAS ÓMICAS

Nombre: Diego Vázquez Zambrano

PREGUNTA 1

Podemos conocer los detalles del dataset con la herramienta Galaxy.



Pinchamos en la “i” de información y accedemos a los detalles de información.

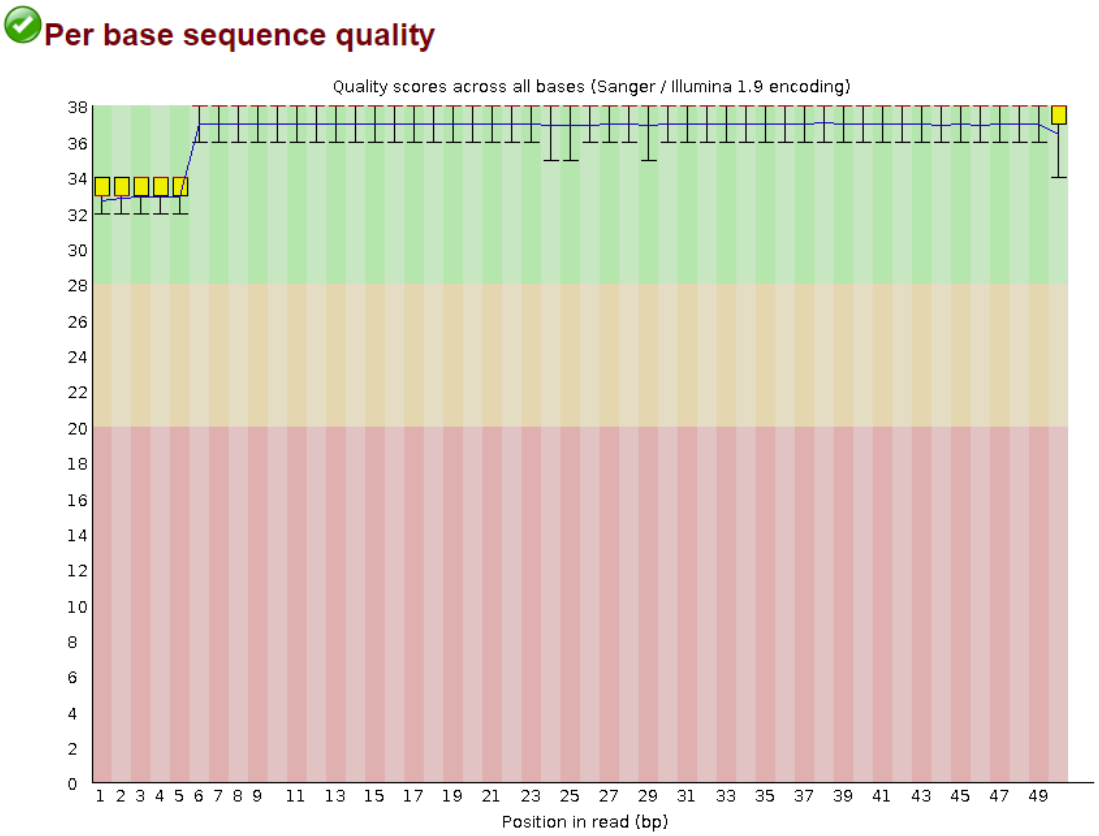
Format	fastqsanger

Según los datos proporcionados en la salida de FastQC, hay un total de 250,000 secuencias. No se ha identificado ninguna secuencia como de mala calidad (Sequences flagged as poor quality: 0). La longitud de las secuencias es de 50 bases, y el porcentaje de contenido de guanina-citosina (%GC) es del 55%. Por lo tanto, hay 250,000 secuencias en total. El tipo de secuenciación utilizado es Sanger/Illumina 1.9.

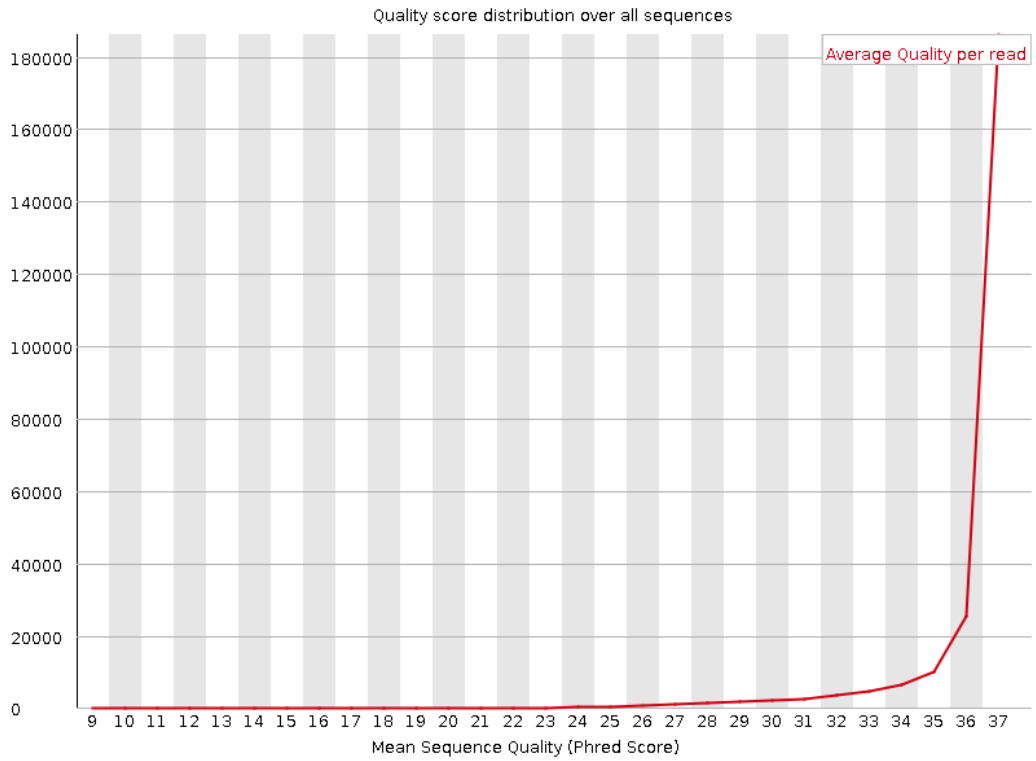
```
##FastQC        0.11.9
>>Basic Statistics      pass
#Measure        Value
Filename        S01_Bac01_read2_fastq
File type       Conventional base calls
Encoding        Sanger / Illumina 1.9
Total Sequences 250000
Sequences flagged as poor quality       0
Sequence length 50
```

Observando la gráfica, vemos que la calidad de la secuencia por base es óptima. Esto significa que las secuencias obtenidas tienen una alta probabilidad de ser correctas en términos de la identificación precisa de las bases nucleotídicas en cada posición de la secuencia. Una calidad de secuencia por base considerada "buena" significa que la probabilidad de error en la

determinación de la base nucleotídica es baja, lo que indica que las secuencias son confiables y precisas en términos de la identificación de las bases. Esto es importante para el análisis posterior de los datos, ya que una calidad de secuencia por base baja puede afectar la precisión del ensamblaje, la identificación de variantes, la detección de mutaciones y otros análisis bioinformáticos.



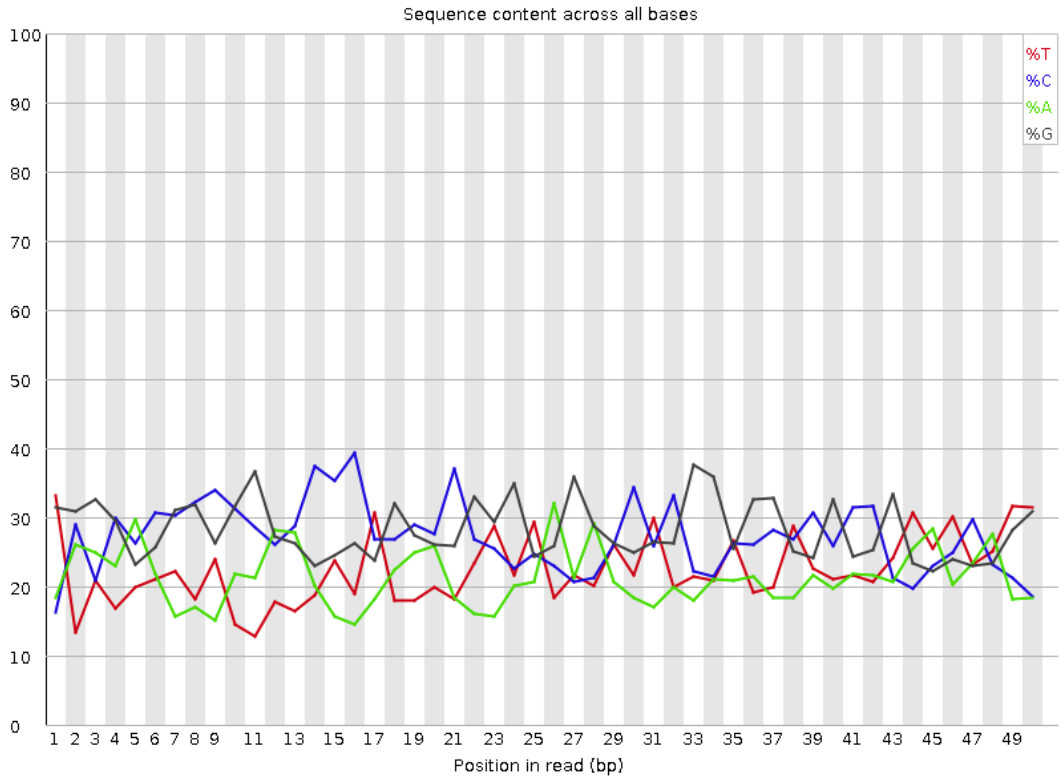
Per sequence quality scores



El "contenido de secuencia por base" se refiere al porcentaje de cada uno de los cuatro nucleótidos (T, C, A, G) en cada posición a lo largo de todas las secuencias de ADN o ARN en el archivo de secuencia de entrada. Esta información se representa en forma de gráficos o trazados que muestran cómo varía el porcentaje de cada nucleótido en cada posición de las secuencias en el archivo. Estos trazados son útiles para evaluar la composición de nucleótidos en diferentes regiones de la secuencia y pueden ayudar a identificar patrones o características específicas de la secuencia, como regiones ricas en GC o AT, o posibles errores de secuenciación. El contenido de secuencia por base es una medida importante de la calidad y confiabilidad de los datos de secuenciación, ya que una distribución equilibrada de los nucleótidos en todas las posiciones indica una mayor confiabilidad en la identificación precisa de las bases nucleotídicas en las secuencias.

En este caso hay poca alineación, lo que puede llevar a resultados inexactos o confusos en el análisis del contenido de secuencia por base.

Per base sequence content



El "contenido de GC por secuencia" en Galaxy se refiere al porcentaje de nucleótidos de Guanina (G) y Citosina (C) en una secuencia de ADN o ARN específica. Este análisis permite evaluar la composición de nucleótidos ricos en GC en una secuencia, lo que puede proporcionar información sobre la estructura y características de la secuencia.

El contenido de GC es una medida importante en la genómica y bioinformática, ya que la proporción de nucleótidos ricos en GC en una secuencia puede influir en su estabilidad, estructura y función.

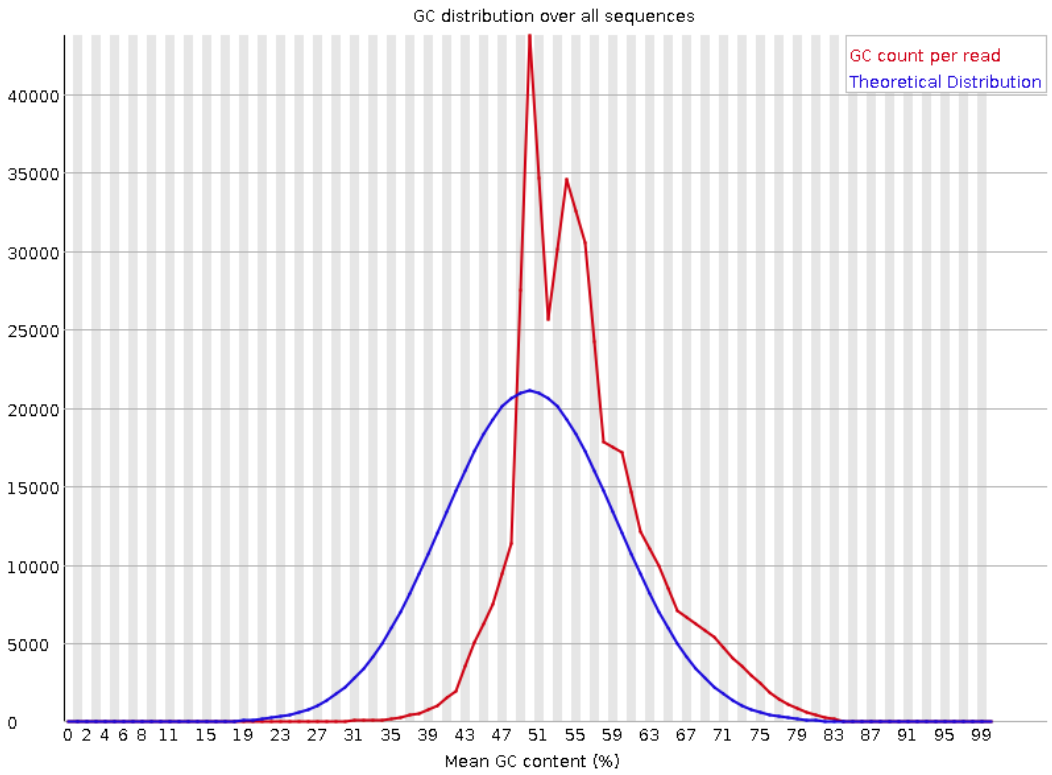
Nuestro análisis del contenido de GC por secuencia en Galaxy muestra un pico muy alto en el medio del gráfico, esto puede indicar la presencia de una región específica en la secuencia que tiene un contenido de GC inusualmente alto en comparación con el resto de la secuencia. Este pico puede deberse a diversas razones, y su interpretación puede depender del contexto del estudio y el tipo de secuencia que se esté analizando.

No necesariamente se puede determinar si un pico alto de contenido de GC en el medio del gráfico en Galaxy es bueno o malo sin más contexto y análisis adicional. La interpretación de si es bueno o malo dependerá del tipo de secuencia que se esté analizando, los objetivos del estudio y la pregunta de investigación.

Por ejemplo, en algunos casos, un pico alto de contenido de GC en el medio del gráfico puede ser indicativo de la presencia de una región codificante de un gen o una región de interés con características específicas, lo cual podría ser relevante para el estudio en curso. En otros casos, un pico alto de contenido de GC puede ser causado por artefactos técnicos o errores de secuenciación, lo cual podría afectar la calidad de los datos y la interpretación de los resultados.

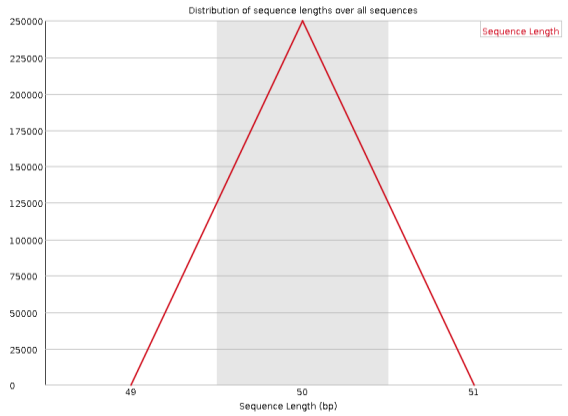
Lo más probable es que en este caso se trate de una contaminación de los datos.

Per sequence GC content



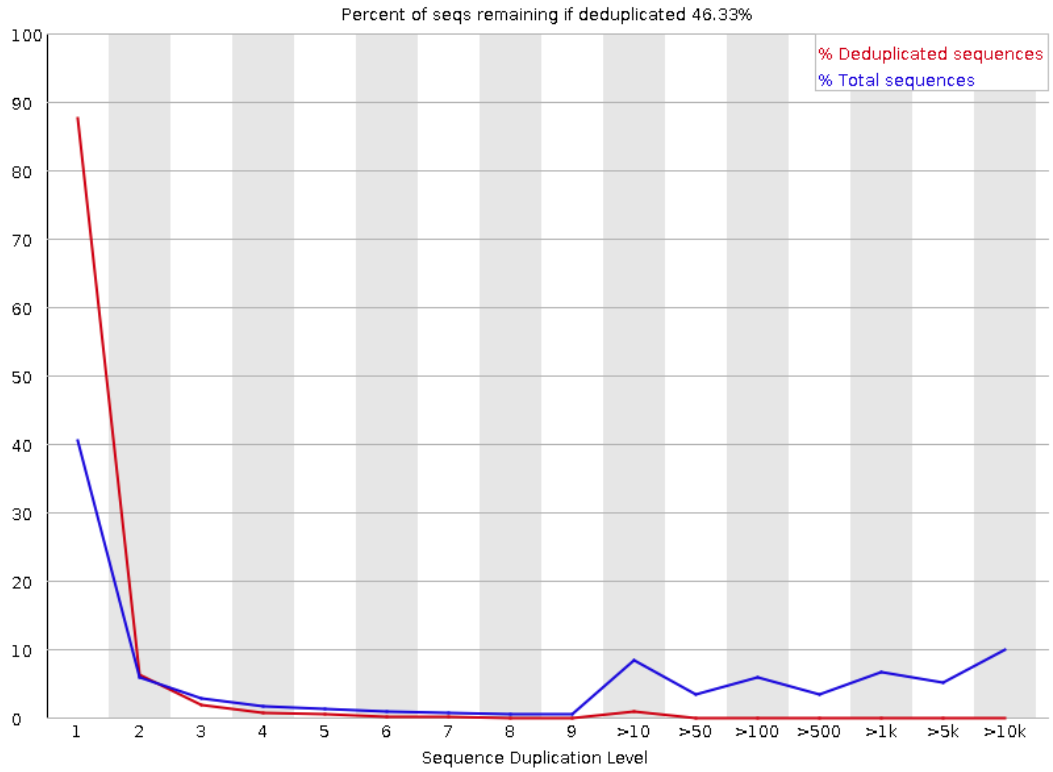
Indica una buena distribución.

Sequence Length Distribution



Se llevó a cabo una revisión de la calidad de la secuencia utilizando el paquete Rqc de R. El código utilizado fue el siguiente:

Sequence Duplication Levels




```
folderx <- system.file(package="ShortRead", "extdata/PEC1")
rqc(path = folderx, pattern = "fastq")
```

El cual generó un informe HTML con el control de calidad.

La primera tabla nos muestra las características de los datos, formato FASTQ con 250000 reads.

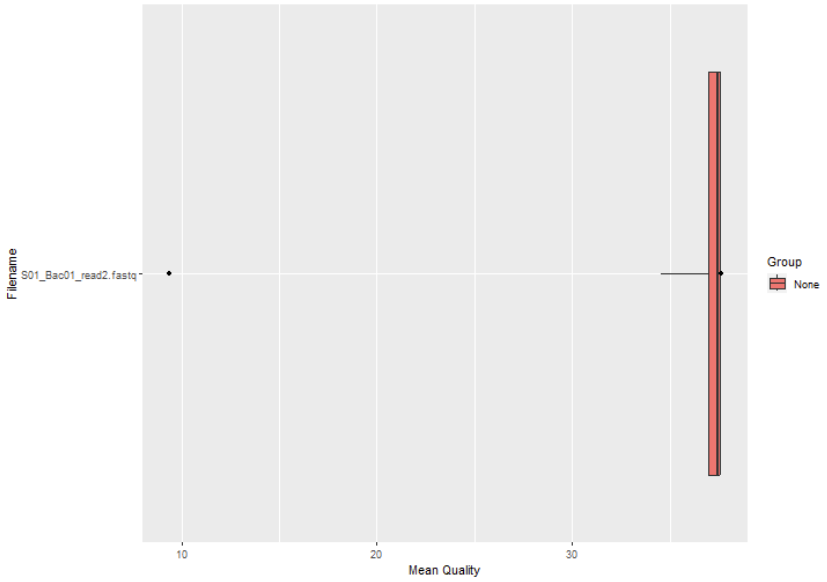
filename	pair	format	group	reads	total.reads	path
S01_Bac01_read2.fastq	1	FASTQ	None	250000	250000	C:/Users/diego/AppData/Local/R/win-library/4.2/ShortRead/extdata/PEC1

La segunda tabla nos muestra el grupo. En este caso solo tenemos un grupo por lo que aparece así, width = 50 ciclos.

Per Read Mean Quality Distribution of Files

This plot describe an overview of per read mean quality distribution of all files

```
rqcReadQualityBoxPlot(rqcResultSet)
```

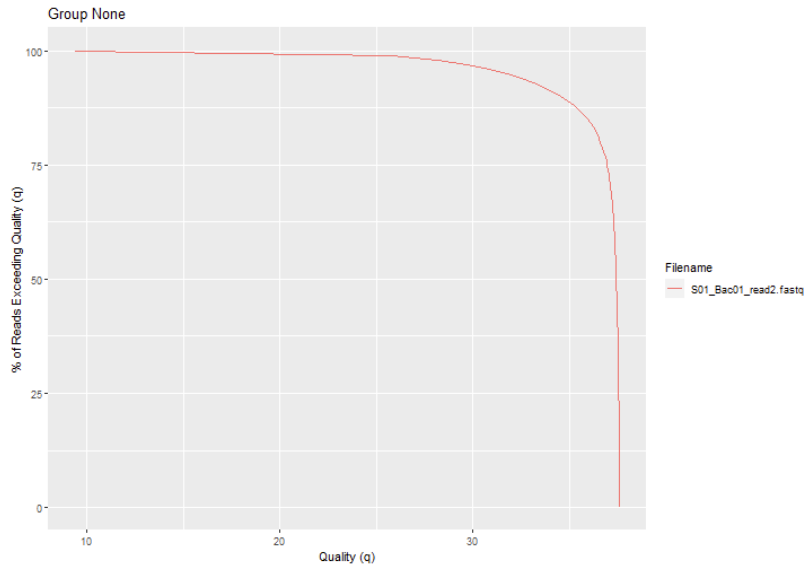


El gráfico representa el patrón de calidad promedio de las lecturas de un conjunto de datos (denominado rqcResultSet) en función de los umbrales de calidad en el eje X y el porcentaje de lecturas que superan ese nivel de calidad en el eje Y.

Average Quality

This plot describes the average quality pattern by showing on the X-axis quality thresholds and on the Y-axis the percentage of reads that exceed that quality level.

```
groups <- unique(perFileInformation(rqcResultSet)$group)
for (group in groups) {
  rqcResultSet.sub <- subsetByGroup(rqcResultSet, group)
  print(rqcReadQualityPlot(rqcResultSet.sub) +
          ggtitle(paste("Group", group)))
}
```

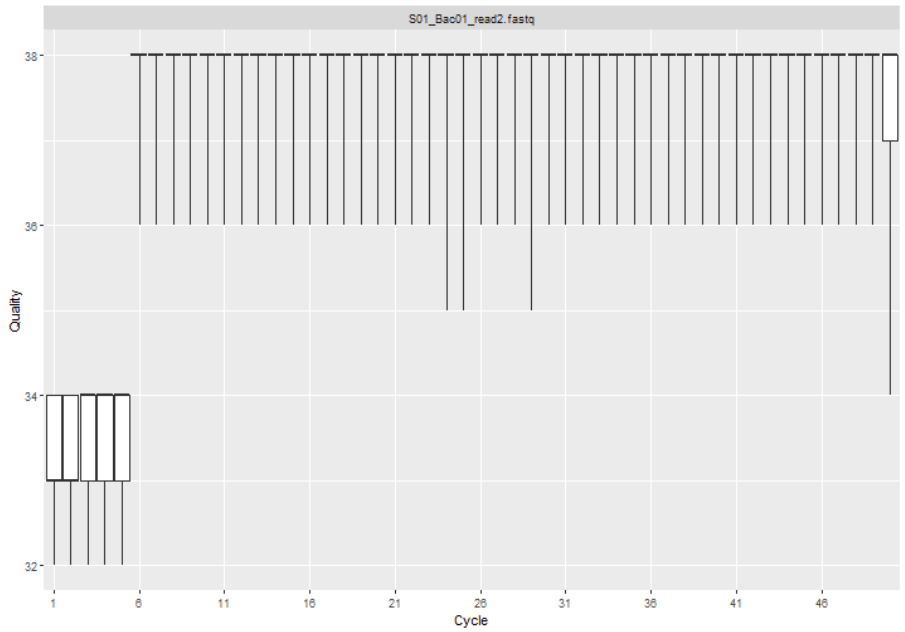


Este gráfico es el equivalente a (Per base sequence quality) generado por Galaxy. En este gráfico se observa el mismo resultado. La diferencia con el gráfico generado por Galaxy es que este último te reconoce si la distribución es buena o mala.

Cycle-specific Quality Distribution - Boxplot

Boxplots describing empirical patterns of quality distribution on each cycle of sequencing.

```
for(pair in pairs) {
  rqcResultSet.sub <- subsetByPair(rqcResultSet, pair)
  print(rqcCycleQualityBoxPlot(rqcResultSet.sub))
}
```

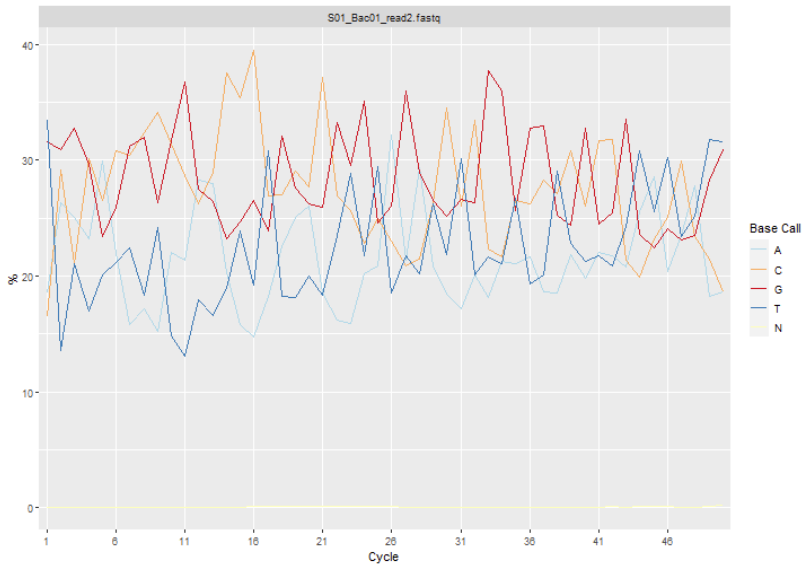


Este gráfico es el equivalente a “Per base sequence content” generado por Galaxy. Vemos una distribución similar en ambos gráficos. En este caso hay poca alineación, lo que puede llevar a resultados inexactos o confusos en el análisis del contenido de secuencia por base.

Cycle

The line plot shown below contains the same information as the plot above. However, some may find this easier to read when comparing the calling rates for each of the nucleotides.

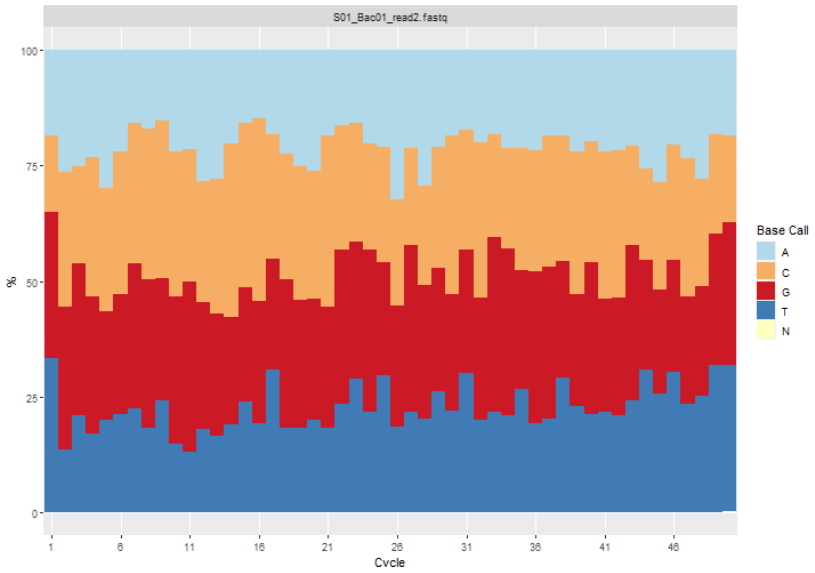
```
for(pair in pairs) {
  rqcResultSet.sub <- subsetByPair(rqcResultSet, pair)
  print(rqcCycleBaseCallsLinePlot(rqcResultSet.sub))
}
```



Esta gráfica es la misma que la anterior pero representada de otra forma.

This stacked bar plot describes the proportion of each nucleotide called for every cycle of sequencing.

```
for(pair in pairs) {
  rqcResultSet.sub <- subsetByPair(rqcResultSet, pair)
  print(rqcCycleBaseCallsPlot(rqcResultSet.sub))
}
```

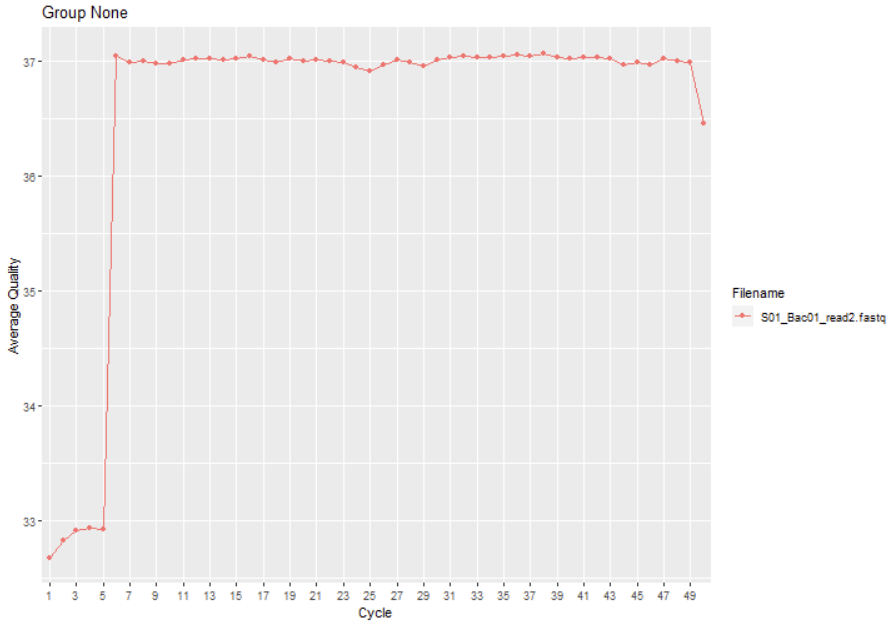


Esta gráfica describe los puntajes promedio de calidad para cada ciclo de secuenciación.

Cycle-specific Average Quality

This plot describes the average quality scores for each cycle of sequencing.

```
for (group in groups) {
  rqcResultSet.sub <- subsetByGroup(rqcResultSet, group)
  print(rqcCycleAverageQualityPlot(rqcResultSet.sub) +
          ggtitle(paste("Group", group)))
}
```

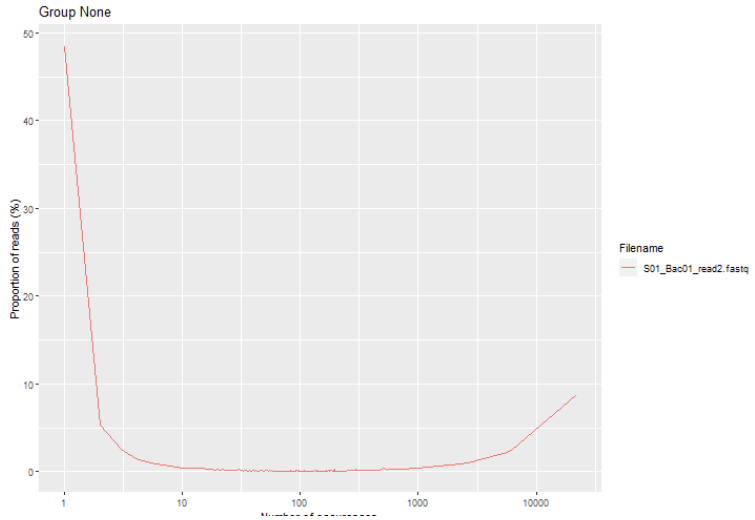


Esta gráfica muestra la proporción de lecturas que aparecieron muchas veces. A menor proporción a lo largo de las ocurrencias, más estable y de más calidad es la muestra.

Read Frequency

This plot shows the proportion of reads that appeared many times.

```
for (group in groups) {
  rqcResultSet.sub <- subsetByGroup(rqcResultSet, group)
  print(rqcReadFrequencyPlot(rqcResultSet.sub) +
          ggtitle(paste("Group", group)))
}
```



Esta gráfica nos muestra el contenido promedio de GC para cada ciclo de secuenciación. Se observa bastante variación (variación máxima de 15-20%).

```
for (group in groups) {
  rqcResultSet.sub <- subsetByGroup(rqcResultSet, group)
  print(rqcCycleGCPlot(rqcResultSet.sub) +
          ggtitle(paste("Group", group)))
}
```

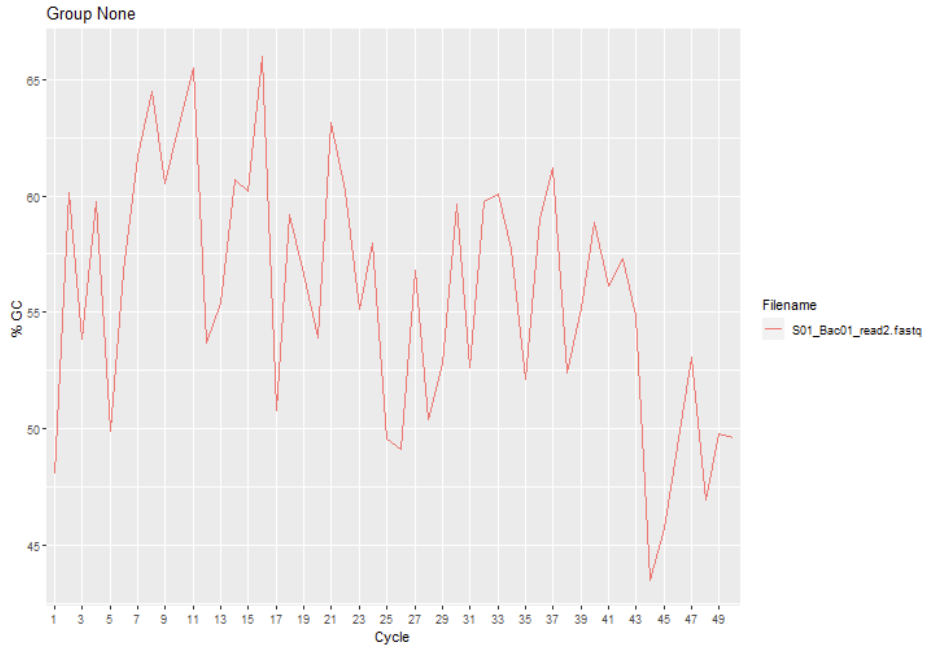
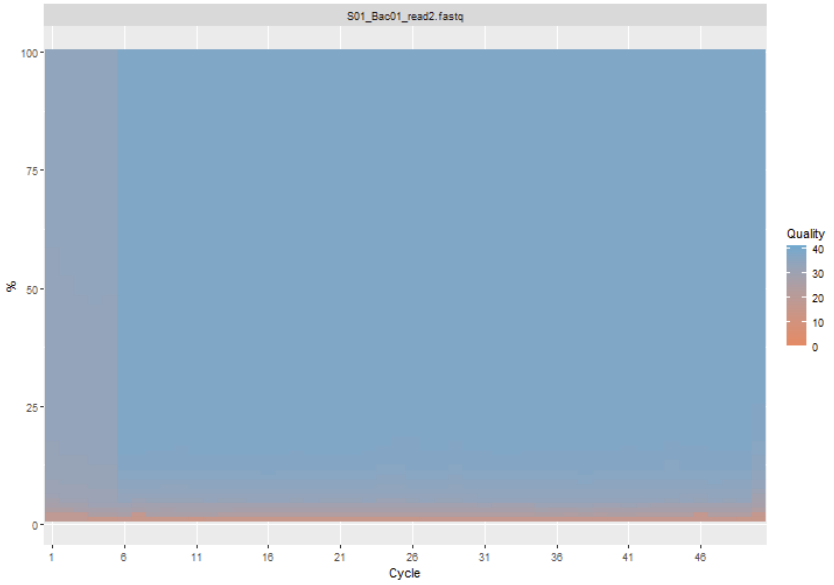


Gráfico de barras que muestra la proporción de llamadas de calidad por ciclo. Los colores se presentan en un degradado rojo-azul, donde el rojo identifica las llamadas de menor calidad.

```
for(pair in pairs) {
  rqcResultSet.sub <- subsetByPair(rqcResultSet, pair)
  print(rqcCycleQualityPlot(rqcResultSet.sub))
}
```



Comparación de ambos reportes

Comparando los resultados obtenidos en Galaxy con la herramienta FastQC y el informe generado con R con el paquete Rqc de Bioconductor, se observan unos resultados similares. No obstante, cabe decir que los resultados son más claros si observamos los obtenidos con la herramienta FastQC. Además, el paquete Rqc no nos proporciona una gráfica de “Per sequence GC content”, el cual me resulta más sencillo de interpretar que el generado por Rqc que se llama “Cycle-specific GC content”.