# CLASE 10

## Introducción al análisis de secuencias NGS

## DBT 792
## GENÉTICA Y GENÓMICA EN PRODUCCIÓN ANIMAL

**Profesor**
**Dr. José Gallardo**

# PLAN DE LA CLASE

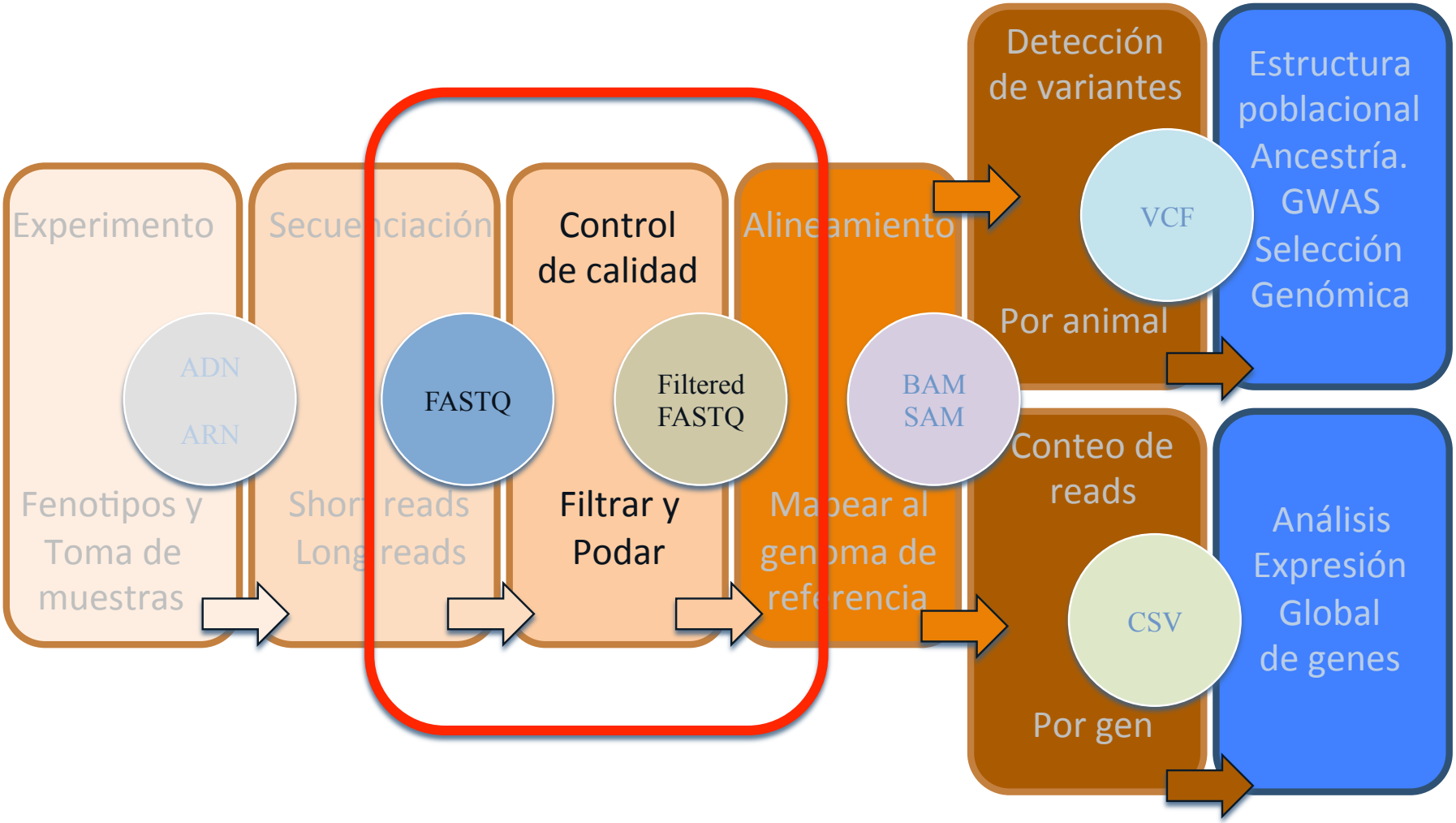
## Introducción

- Repaso flujo de trabajo proyecto genómica aplicada.
- Qfred: score de calidad de una secuencia.
- Formato de secuencias FASTQ.
- Control de calidad: Software fastQC
- Filtrado y poda: Software Trimmomatic.

## Práctica

- Linux para genómica: Análisis de secuencias FASTQ y control de calidad.

# FLUJO DE TRABAJO DE UN PROYECTO DE GENÓMICA APLICADA



# QFRED: SCORE DE CALIDAD DE UNA SECUENCIA

$$Q_{\mathrm{PHRED}} = -10 \times \log_{10}(P_e)$$

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |

# FORMATO FASTQ

## Codificación ASCHI del Score de calidad

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
   Quality score: 0.........10........20........30........40
```

## Formato fastq

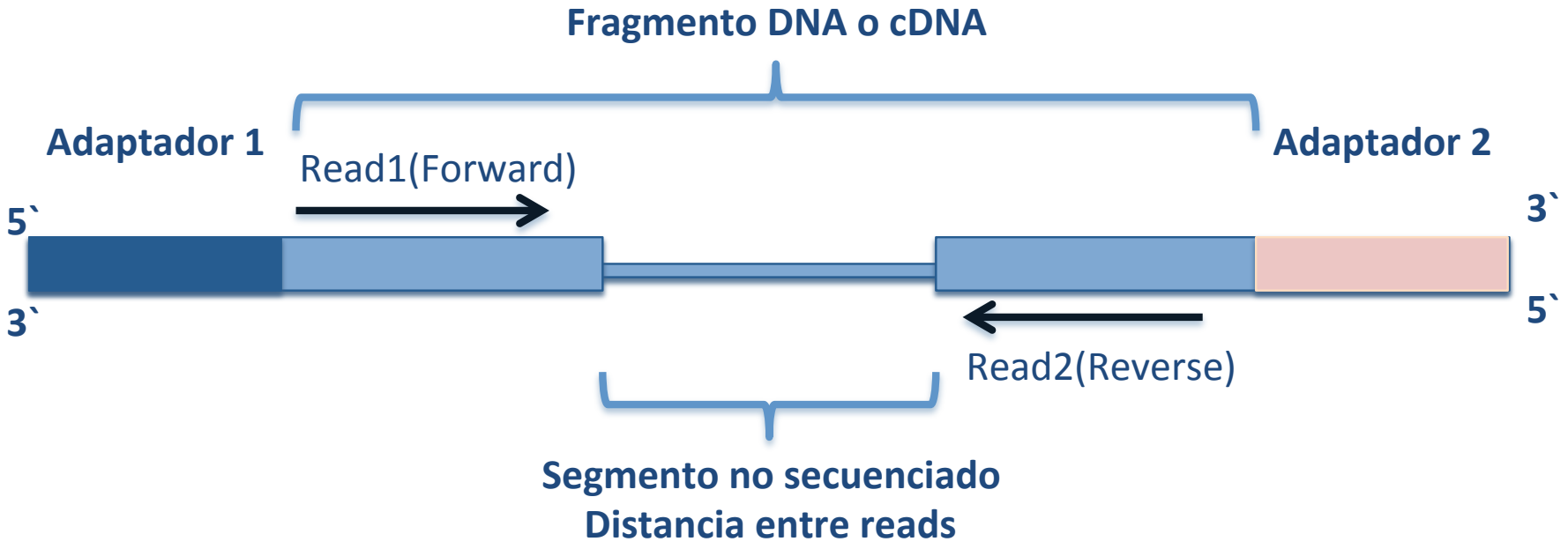
@SEQ_ID

**NN**TTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTGGGGGGG

+

**!+ABCF.BCDHGFEABC89BCFEFFFAAAACCF????????CC++..//5**

# SECUENCIAS PAIRED-END




When performing sequencing on an Illumina instrument, sequences corresponding to the library adapter can be present in the FASTQ files at the 3' end of the reads if the read length is greater than the insert size.

# EJEMPLO SECUENCIA FASTQ – DESDE NCBI

## Reads (separated)

**>gnl|SRA|SRR5585855.1.1** 1 *(Biological)*

```
GTACANGAATGCCATTAATGGAACAGTGTTTTCAGGGTTCATTGGTATTCTGGTTATGCT
GGGTGACAGTCAGAATCCCAATGTATCTTTCTGACATTATTTGATTGTTCTATATTTGAA
ATGCAGGCCTAGTTAATAAGGTTGGGAGGT
```

One channel quality score

```
32 32 32 32 32  2 36 36 36 36 36 36 36 36 36 36 36 36 36 36
36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36
36 36 36 36 36 36 36 36 36 36 36 36 36 32 36 36 36 36 36 36
36 36 36 36 36 14 36 36 36 36 36 32 36 36 21 36 36 36 36 36
36 36 36 36 36 36 36 36 36 36 36 27 36 32 36 36 27 36 36 36
36 36 36 14 27 36 36 27 36 36 36 36 36 36 36 36 36 36 32 36
36 36 36 36 32 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36
36 36 36 36 27 36 36 36 36 32
```

**>gnl|SRA|SRR5585855.1.2** 1 *(Biological)*

```
ATCATGATCATTAATCGCCCACTGCTCCGAGGGAGCCAGTGTAACGAAGATAGAAATGTC
AGGACAACCCTGATCACAGGTACACAGGTTATATNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

One channel quality score

```
32 32 32 32 32 36 36 36 36 36 32 36 36 36 36 36 36 36 36 36
36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 36 32 21 21
32 36 36 36 36 36 14 36 36 32 36 36 36 36 36 14 36 14 36 36
14 36 36 32 36 36 14 36 36 36 36 14 14 36 36 14 36 14 32 32
36 36 36 36 36 14 36 36 14 27 36 14 27 36  2  2  2  2  2  2
 2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
 2  2  2  2  2  2  2  2  2  2  2
```

# EJEMPLO SECUENCIA FASTQ – DESDE NCBI

## Reads (separated)

**>gnl|SRA|SRR5585855.131.1** 131 *(Biological)*

```
CTGATNCTAGTAGGCATGGTAACTAACTGATACTAGTAGGCATGGTAAATAACTGATACA
AGGTATGAGGAAAAATATCTAATACAATAAATCAAGCGAAAAAAATAATACATATAATAA
ATGAAAGAGAGTGGGGGGGGGGGGGGGGGGG
```

**One channel quality score**

```
32 32 32 32 32  2 36 36 36 36 36 36 36 36 36 36 36 36 36 36
36 36 36 36 36 36 36 36 36 36 36 21 36 36 32 32 36 36 36 36
36 36 36 36 36 32 36 32 14 36 36 32 36 36 14 36 36 36 14 14
14 14 14 14 14 14 14 36 14 14 14 14 32 14 14 36 14 14 14 14
14 21 36 14 36 14 14 14 14 36 14 36 14 14 14 14 14 14 14 32
14 14 14 21 14 14 14 32 27 14 14 14 14 14 14 14 14 14 14 14
14 14 27 14 36 32 14 14 14 14 14 14 36 14 27 36 27 14 27 32
21 32 32 32 32 36 32 32 36 14 32
```

**>gnl|SRA|SRR5585855.131.2** 131 *(Biological)*

```
ATAAGTTATTTACCATGCCTACTAGTATCAGTTATTTACCATGCCTACTAGTATCAGATA
TATGAAGAGCGTCGTTTTATGAAAGAGTTAAAGATATAAATTTNATCANTTANNTAANCT
NNNNATNAAT
```

**One channel quality score**

```
32 32 32 32 32 36 36 14 36 36 36 36 36 36 36 36 21 36 36 36
36 36 36 32 36 36 32 36 36 36 36 36 32 36 14 32 36 36 36 14
36 36 36 36 36 36 36 14 36 36 36 36 36 36 36 36 36 36 32 14 36
14 14 14 14 14 14 14 32 32 14 14 36 14 14 36 14 14 27 14 32
14 36 14 14 14 21 14 27 14 14 36 14 14 14 14 14 14 14 32 14
27 27 14  2 14 36 14 14  2 21 14 14  2  2 32 14 14  2 14 32
 2  2  2  2 14 32  2 14 14 27
```

# Control de calidad: Software fastQC.

**fastQC:** Software que permite realizar un análisis de control de calidad de secuencias fastq que provienen de secuenciadores NGS. Trabaja de manera modular analizando diversos parámetros predefinidos que resultan del análisis estadístico de millones de secuencias.


Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

# FastQC REPORT: Módulos

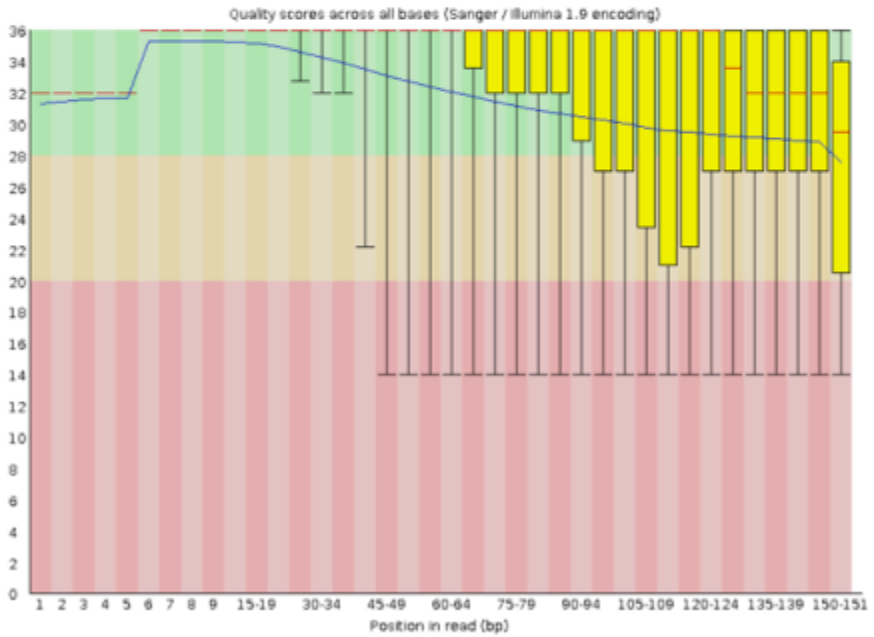
## FastQC Report

### Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

| | |
|---|---|
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 130838907 |
| Sequences flagged as poor quality | 0 |
| Sequence length | 35-151 |
| %GC | 43 |

### Per base sequence quality

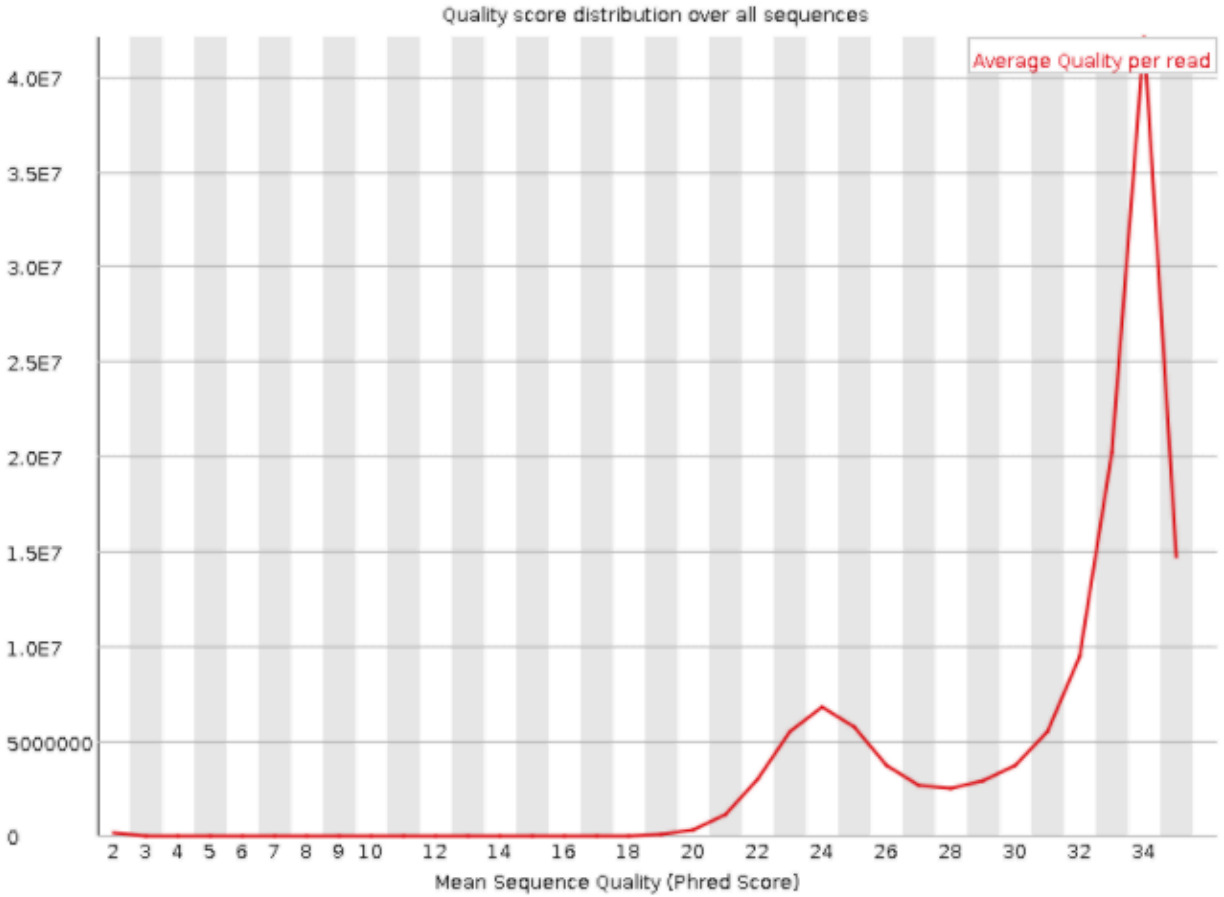


# FastQC REPORT: módulo 2

```
>>Per base sequence quality     pass
#Base    Mean    Median  Lower Quartile  Upper Quartile  10th Percentile 90th Percentile
1        31.341284934873713      32.0    32.0    32.0    32.0    32.0
2        31.26387664111815       32.0    32.0    32.0    32.0    32.0
3        31.305511916113986      32.0    32.0    32.0    32.0    32.0
4        31.31268871098888       32.0    32.0    32.0    32.0    32.0
5        31.295367806724546      32.0    32.0    32.0    32.0    32.0
6        34.93476007867835       36.0    36.0    36.0    36.0    36.0
7        34.90135491559175       36.0    36.0    36.0    36.0    36.0
8        34.877736693273064      36.0    36.0    36.0    36.0    36.0
9        34.85289903177069       36.0    36.0    36.0    36.0    36.0
10-14    34.794267078722804      36.0    36.0    36.0    32.8    36.0
15-19    34.63174441262762       36.0    36.0    36.0    32.0    36.0
20-24    34.39053213106741       36.0    36.0    36.0    32.0    36.0
25-29    34.078852798899455      36.0    36.0    36.0    32.0    36.0
30-34    33.72509700346935       36.0    36.0    36.0    28.8    36.0
35-39    33.393023262114426      36.0    36.0    36.0    19.6    36.0
40-44    33.03631369384756       36.0    36.0    36.0    14.0    36.0
45-49    32.6878573088964        36.0    36.0    36.0    14.0    36.0
50-54    32.34977135408522       36.0    36.0    36.0    14.0    36.0
55-59    32.01107153097464       36.0    35.2    36.0    14.0    36.0
60-64    31.676836247091565      36.0    32.0    36.0    14.0    36.0
65-69    31.352852296050163      36.0    32.0    36.0    14.0    36.0
70-74    31.055488956038232      36.0    32.0    36.0    14.0    36.0
75-79    30.772721117913836      36.0    32.0    36.0    14.0    36.0
80-84    30.534869291979238      36.0    28.0    36.0    14.0    36.0
85-89    30.33230997410834       36.0    27.0    36.0    14.0    36.0
90-94    30.14002404468541       36.0    27.0    36.0    14.0    36.0
95-99    29.98056355801439       36.0    27.0    36.0    14.0    36.0
100-104  29.822685376161463      36.0    25.8    36.0    14.0    36.0
105-109  29.604385744467827      36.0    21.0    36.0    14.0    36.0
110-114  29.306667221007736      36.0    15.4    36.0    14.0    36.0
115-119  29.025009216521045      36.0    14.0    36.0    14.0    36.0
120-124  28.74652928129292       36.0    14.0    36.0    14.0    36.0
125-129  28.46743150604746       33.6    14.0    36.0    14.0    36.0
130-134  28.100615125803927      32.0    14.0    36.0    14.0    36.0
135-139  27.686450329081328      32.0    14.0    36.0    14.0    36.0
140-144  27.269882646255432      32.0    14.0    36.0    14.0    36.0
145-149  26.87824931942991       32.0    14.0    35.2    14.0    36.0
150-151  25.35226424829505       29.5    14.0    32.0    14.0    34.0
>>END_MODULE
```

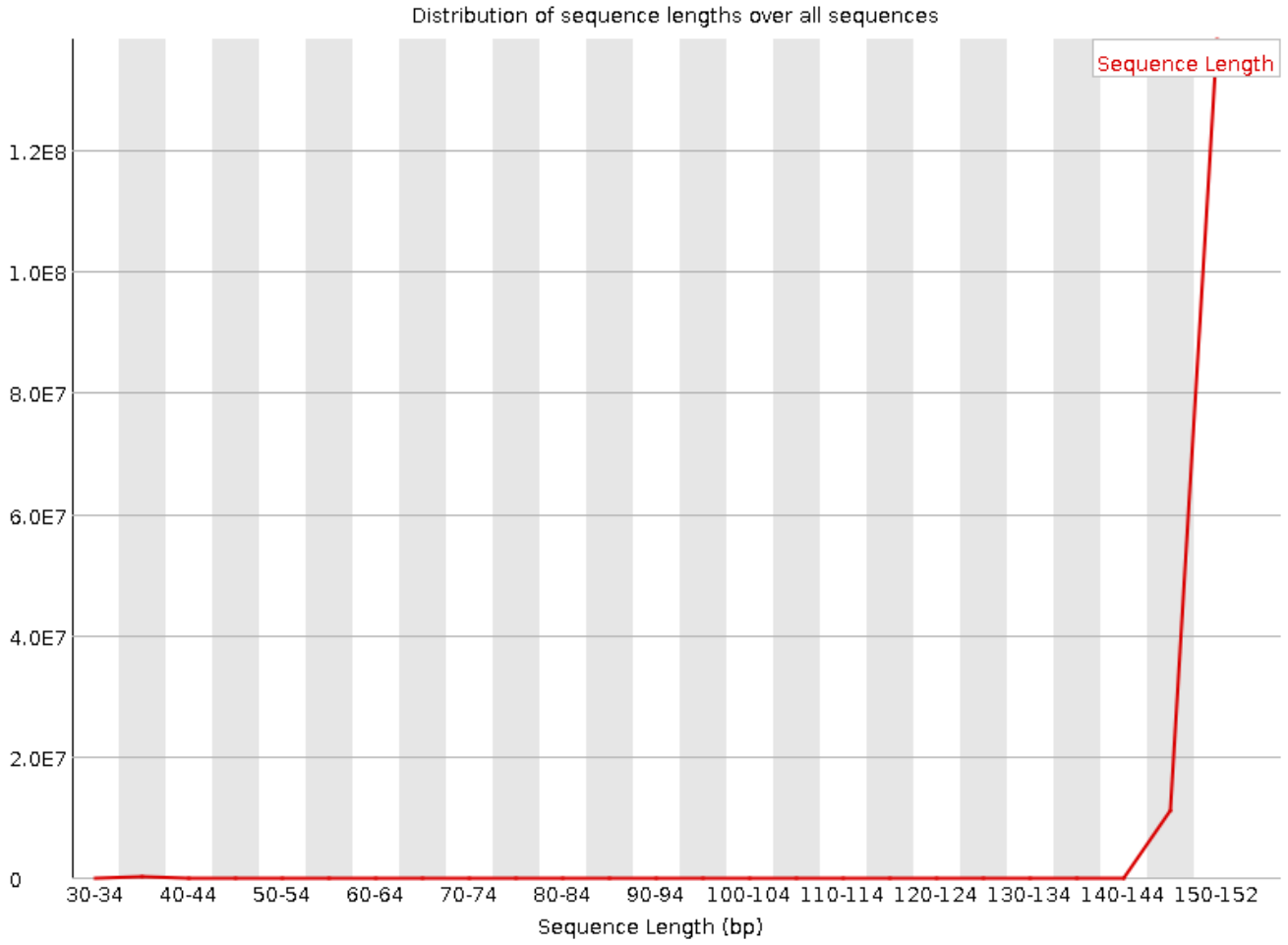
# SCORE DE CALIDAD POR SECUENCIA.

## Per sequence quality scores



# HISTOGRAMA DE TAMAÑO DE LAS SECUENCIAS

## Sequence Length Distribution



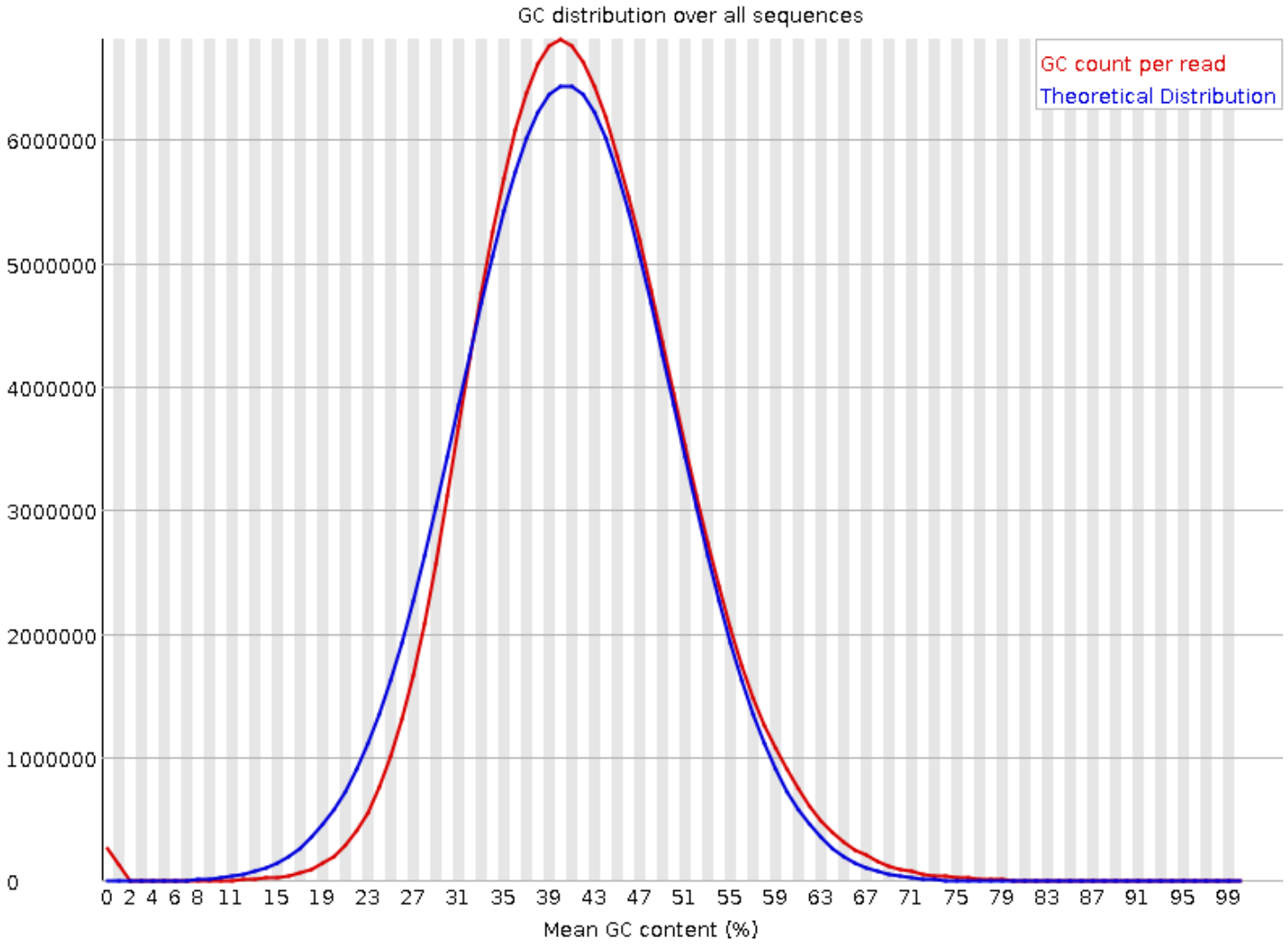
# FastQC REPORT: módulos 3 y 7

```
>>Per sequence quality scores   pass
#Quality        Count
2       170551.0
3       3312.0
4       2835.0
5       1759.0
6       1404.0
7       1712.0
8       2103.0
9       2014.0
10      807.0
11      718.0
12      1048.0
13      1394.0
14      9745.0
15      77727.0
16      175672.0
17      224176.0
18      269380.0
19      393376.0
20      764701.0
21      1737801.0
22      3643357.0
23      5833862.0
24      6777435.0
25      6041079.0
26      4552845.0
27      3618313.0
28      3496625.0
29      3993808.0
30      5141117.0
31      7356270.0
32      1.1959821E7
33      2.229884E7
34      3.5427351E7
35      9340624.0
>>END_MODULE
```

```
>>Sequence Length Distribution  warn
#Length Count
35-39   223104.0
40-44   47028.0
45-49   40416.0
50-54   32463.0
55-59   28845.0
60-64   26129.0
65-69   23353.0
70-74   22425.0
75-79   19658.0
80-84   19359.0
85-89   26150.0
90-94   30137.0
95-99   22920.0
100-104 23492.0
105-109 26746.0
110-114 29487.0
115-119 31243.0
120-124 39270.0
125-129 38557.0
130-134 40962.0
135-139 51273.0
140-144 67505.0
145-149 1.1180246E7
150-152 1.21232814E8
>>END_MODULE
```

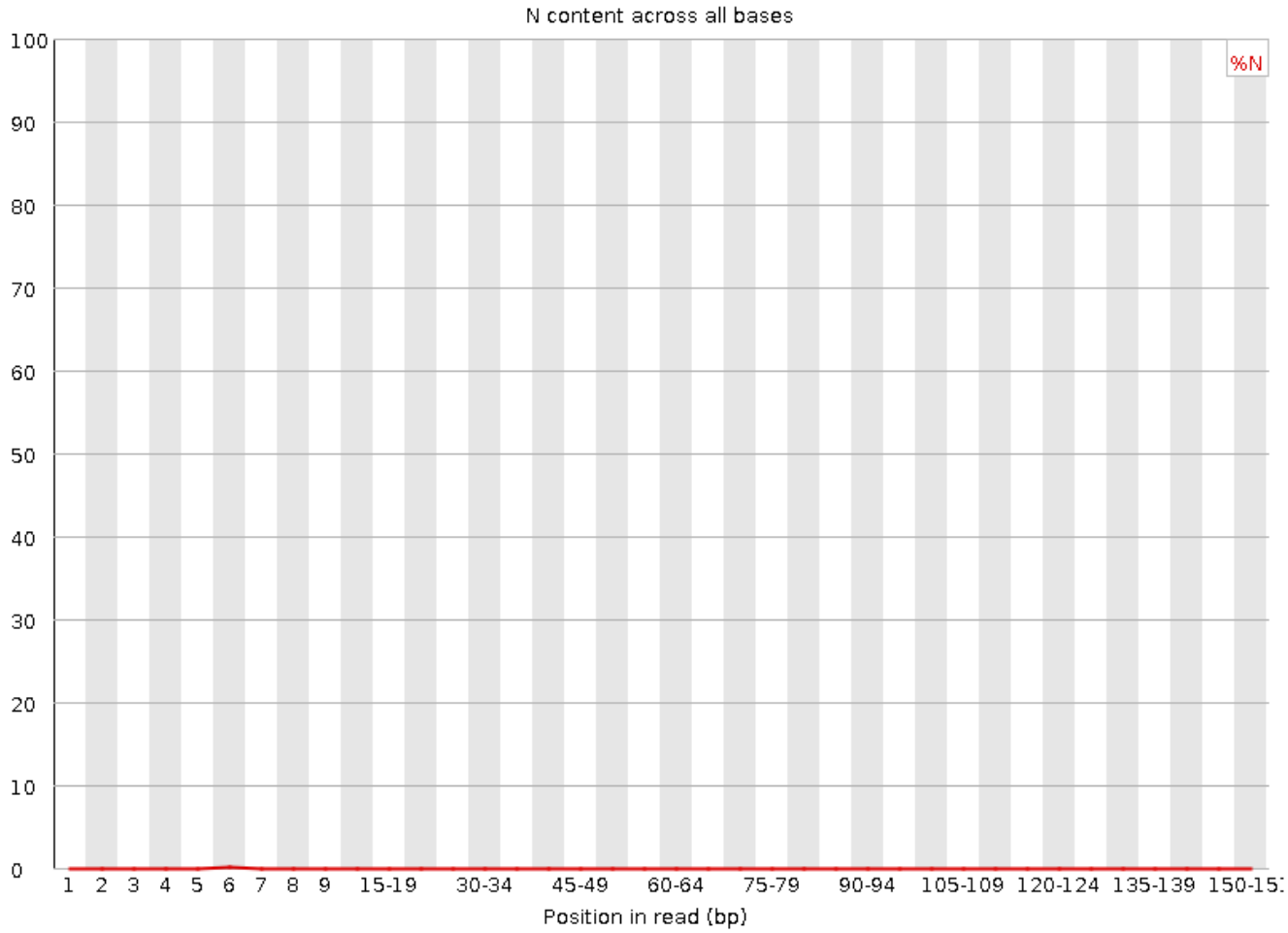
# CONTENIDO CG

## Per sequence GC content



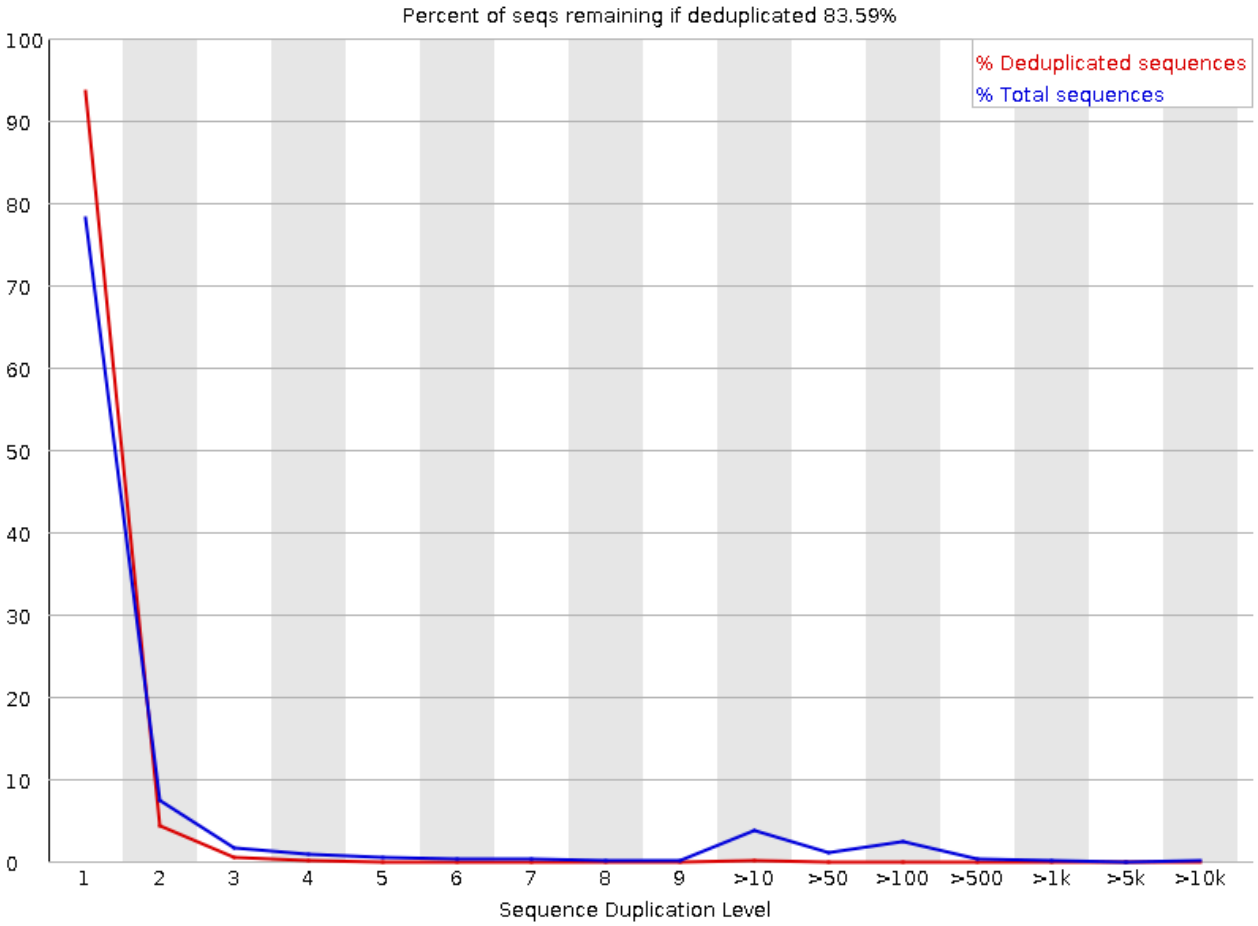
# CONTENIDO N (CUALQUIER NUCLEÓTIDO)

## Per base N content



# NIVEL DE SECUENCIAS DUPLICADAS

## Sequence Duplication Levels

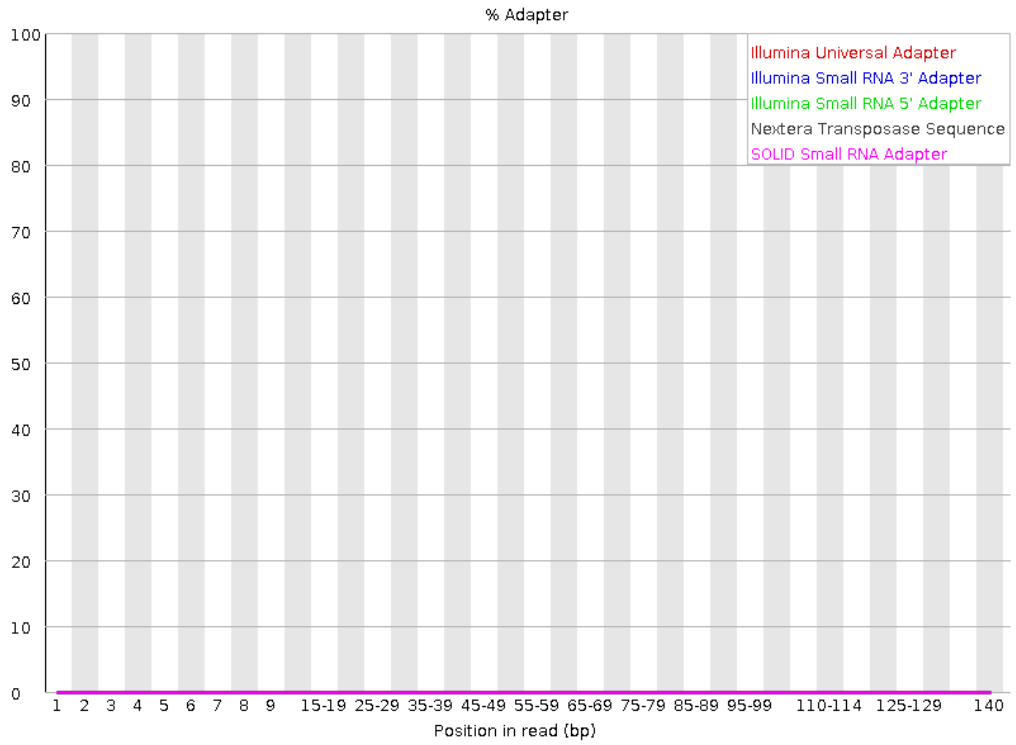


# SECUENCIAS SOBREREPRESENTADAS Y ADAPTADORES

## Overrepresented sequences

| Sequence | Count | Percentage | Possible Source |
|---|---|---|---|
| NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN | 266246 | 0.1768232677121315 | No Hit |

## Adapter Content



# Filtrado y poda: Software Trimmomatic

**Trimmomatic:** Software flexible que permite filtrar y podar secuencias NGS, así como remover adaptadores.

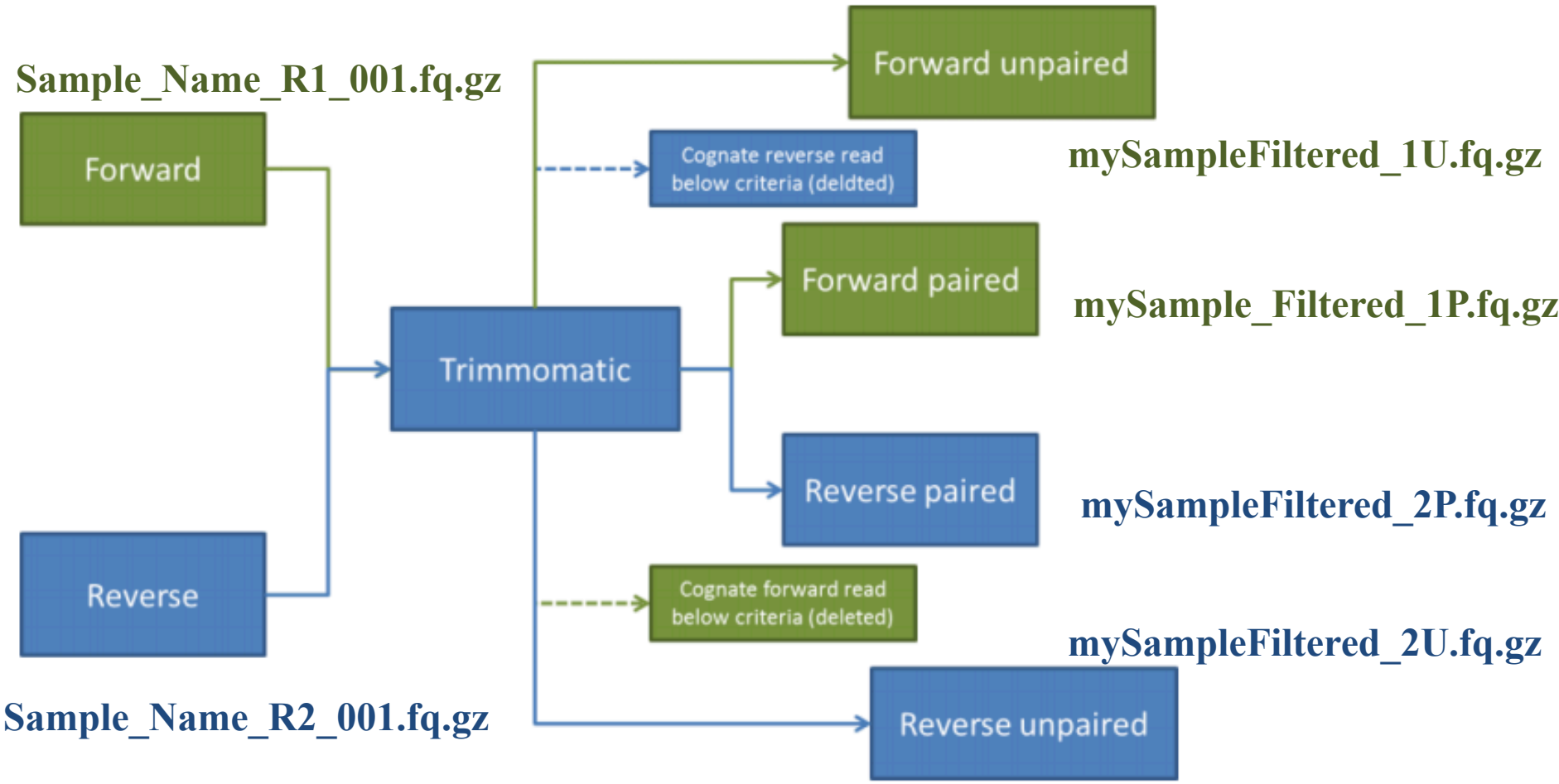
**El software trabaja en dos modos diferentes:** 1) Paired –end: para secuencias pareadas, 2) Single-end como para secuencias únicas.

**Compresión:** Permite el trabajo de secuencias comprimidas en formatos gzip (.gz).

**Atención:** El orden de los comandos determina el orden del proceso. Se sugiere eliminar adaptadores al inicio del proceso y no al final.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England), 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

# Flujo de trabajo secuencias Paired-end



# Software Trimmomatic: comandos y argumentos

The current trimming steps are:

**ILLUMINACLIP:** Cut adapter and other illumina-specific sequences from the read.

**SLIDINGWINDOW:** Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.

**LEADING:** Cut bases off the start of a read, if below a threshold quality

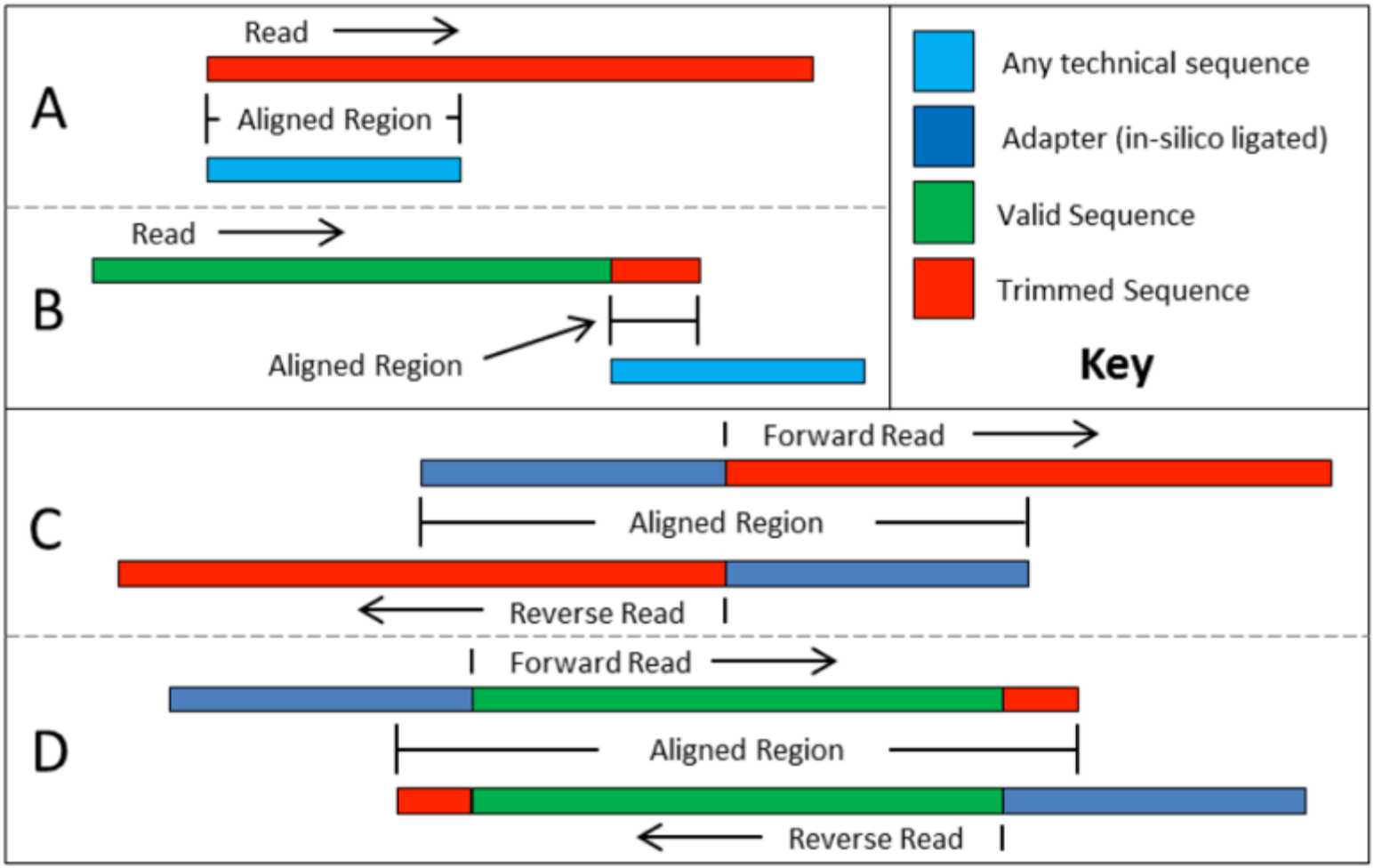
**TRAILING:** Cut bases off the end of a read, if below a threshold quality

**CROP:** Cut the read to a specified length

**HEADCROP:** Cut the specified number of bases from the start of the read

**MINLEN:** Drop the read if it is below a specified length

# Ejemplo eliminación de adaptadores



# SECUENCIAS RECOMENDADAS PARA REMOVER ADAPTADORES DESDE UNA LIBRERÍA TruSEq-PE

**TruSeq LT and TruSeq HT-based kits:**

Read 1: AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
Read 2: AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

# Práctica Análisis de secuencias NGS

Análisis de secuencias FASTQ y control de calidad.

# OBJETIVOS DEL TRABAJO PRÁCTICO

Esta práctica tiene como propósito:

1) Comprobación de integridad de descarga de archivos usando md5sum o similar.
2) Análisis de control de calidad.
3) Filtrado y poda de secuencias.
4) Transferir archivos mediante protocolo FTP desde Servidor a Cliente.