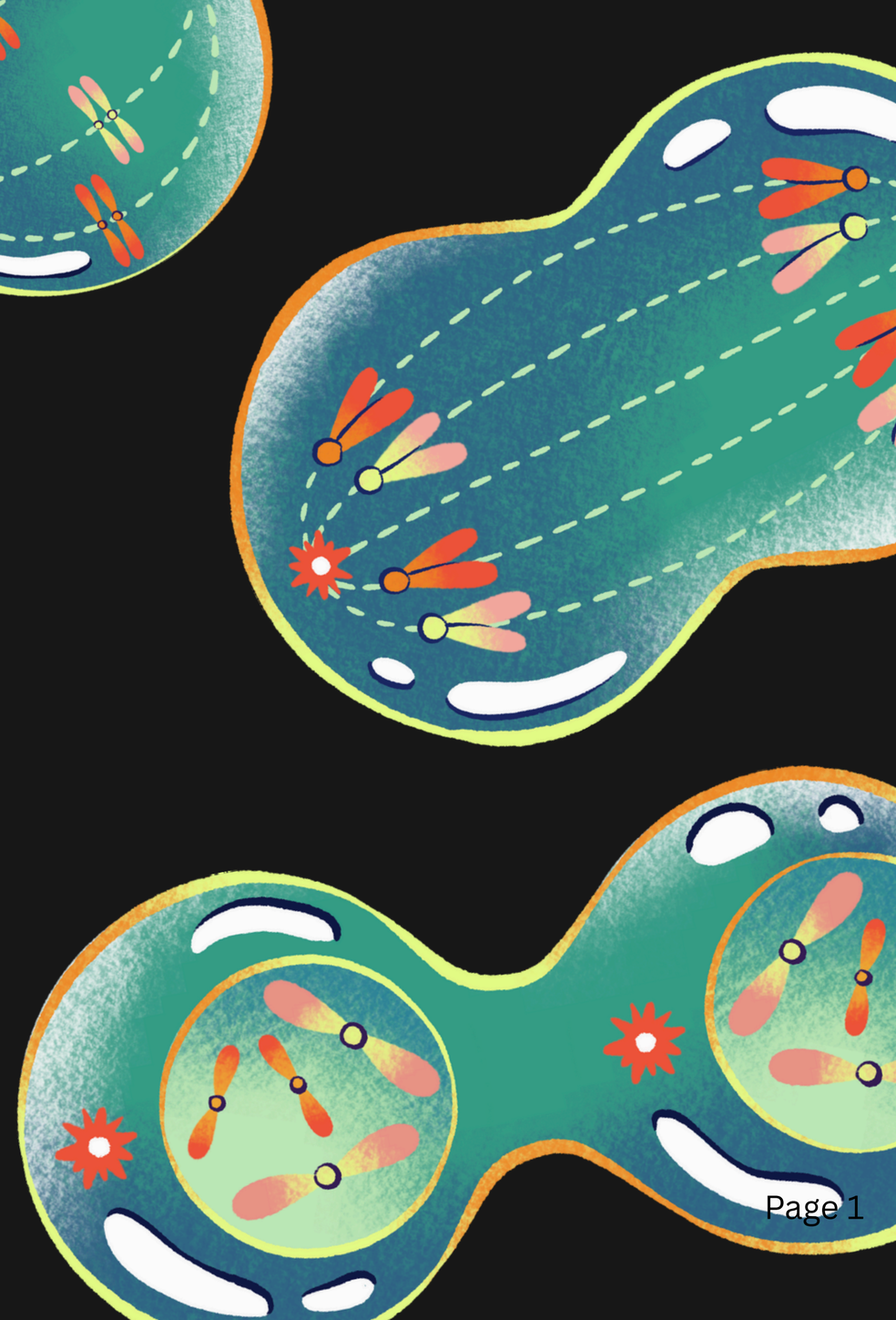
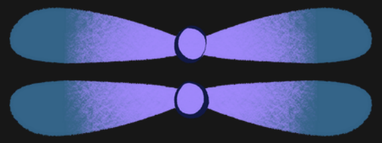
Bioinformatics for Omics Sciences - 2025-1

SEQUENCE TRIMMING WORKSHOP

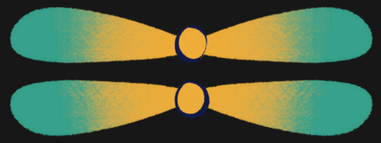
Francisco J. Salamanca
Elsy Carvajal



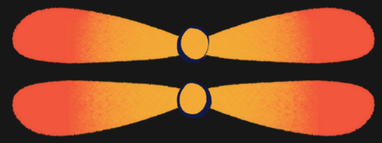
LESSON OBJECTIVES



Comprehend and utilize sequencing data cleaning tools (Trimmomatic and fastp).



Assess the quality of both raw and processed reads using FastQC.



Compare quality metrics prior to and following trimming.

BIOLOGICAL FRAMEWORK

Article

A new coronavirus associated with human respiratory disease in China

https://doi.org/10.1038/s41586-020-2008-3

Received: 7 January 2020

Accepted: 28 January 2020

Published online: 3 February 2020

Open access

Check for updates

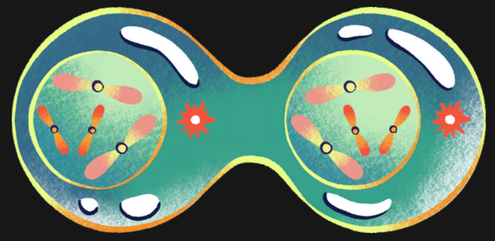
Fan Wu[1,7], Su Zhao[2,7], Bin Yu[3,7], Yan-Mei Chen[1,7], Wen Wang[4,7], Zhi-Gang Song[1,7], Yi Hu[2,7], Zhao-Wu Tao[2], Jun-Hua Tian[3], Yuan-Yuan Pei[1], Ming-Li Yuan[2], Yu-Ling Zhang[1], Fa-Hui Dai[1], Yi Liu[1], Qi-Min Wang[1], Jiao-Jiao Zheng[1], Lin Xu[1], Edward C. Holmes[1,5] & Yong-Zhen Zhang[1,4,6]

Emerging infectious diseases, such as severe acute respiratory syndrome (SARS) and Zika virus disease, present a major threat to public health[1–3]. Despite intense research efforts, how, when and where new diseases appear are still a source of considerable uncertainty. A severe respiratory disease was recently reported in Wuhan, Hubei province, China. As of 25 January 2020, at least 1,975 cases had been reported since the first patient was hospitalized on 12 December 2019. Epidemiological investigations have suggested that the outbreak was associated with a seafood market in Wuhan. Here we study a single patient who was a worker at the market and who was admitted to the Central Hospital of Wuhan on 26 December 2019 while experiencing a severe respiratory syndrome that included fever, dizziness and a cough. Metagenomic RNA sequencing[4] of a sample of bronchoalveolar lavage fluid from the patient identified a new RNA virus strain from the family *Coronaviridae*, which is designated here 'WH-Human 1' coronavirus (and has also been referred to as '2019-nCoV'). Phylogenetic analysis of the complete viral genome (29,903 nucleotides) revealed that the virus was most closely related (89.1% nucleotide similarity) to a group of SARS-like coronaviruses (genus Betacoronavirus, subgenus Sarbecovirus) that had previously been found in bats in China[5]. This outbreak highlights the ongoing ability of viral spill-over from animals to cause severe disease in humans.

Sample type: SARS-CoV-2 in humans.

Platform: Illumina MiniSeq (paired-end sequencing).

ORIGINAL SEQUENCES: INFO



WH_R1.fastq

WH_R2.fastq

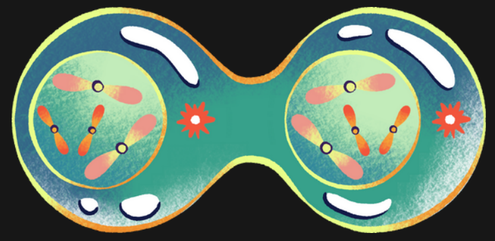
Basic Statistics

Measure	Value
Filename	WH_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	28282964
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	47

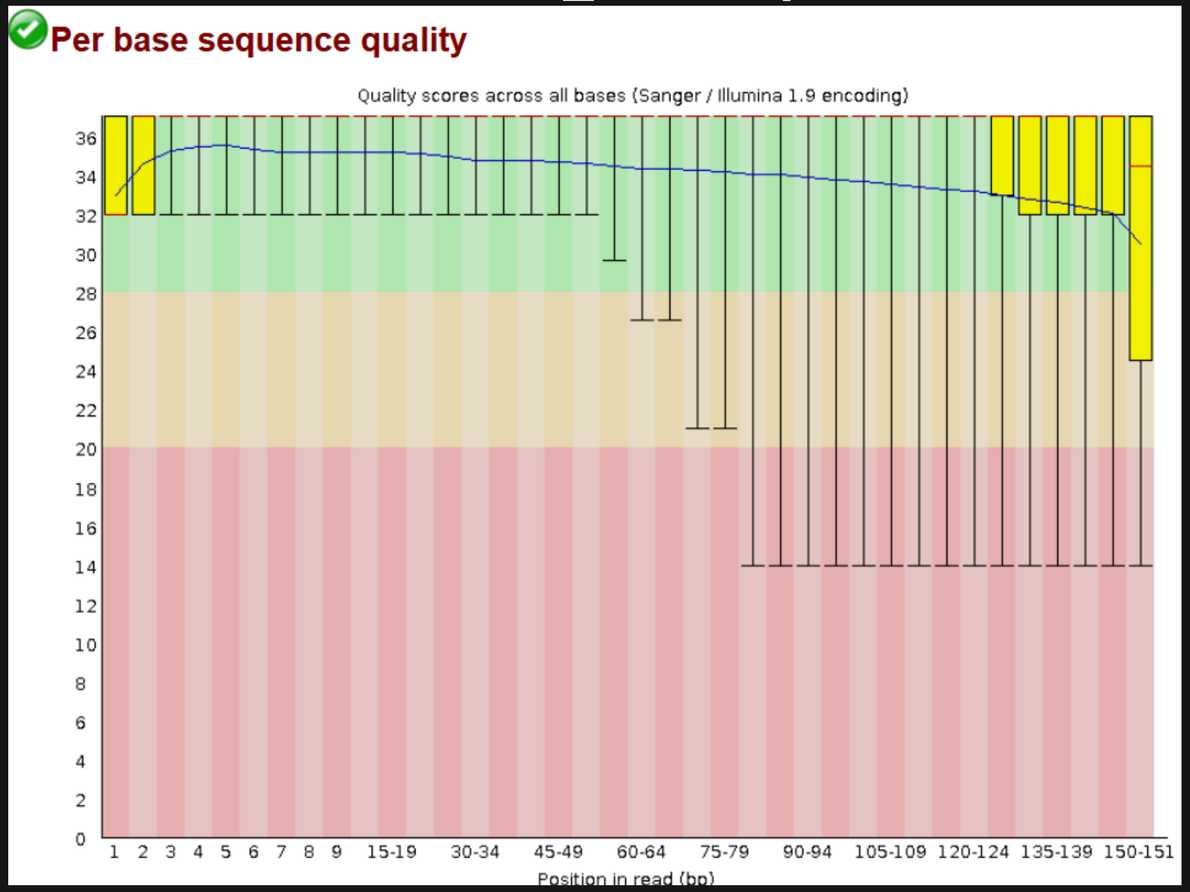
Basic Statistics

Measure	Value
Filename	WH_R2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	28282964
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	46

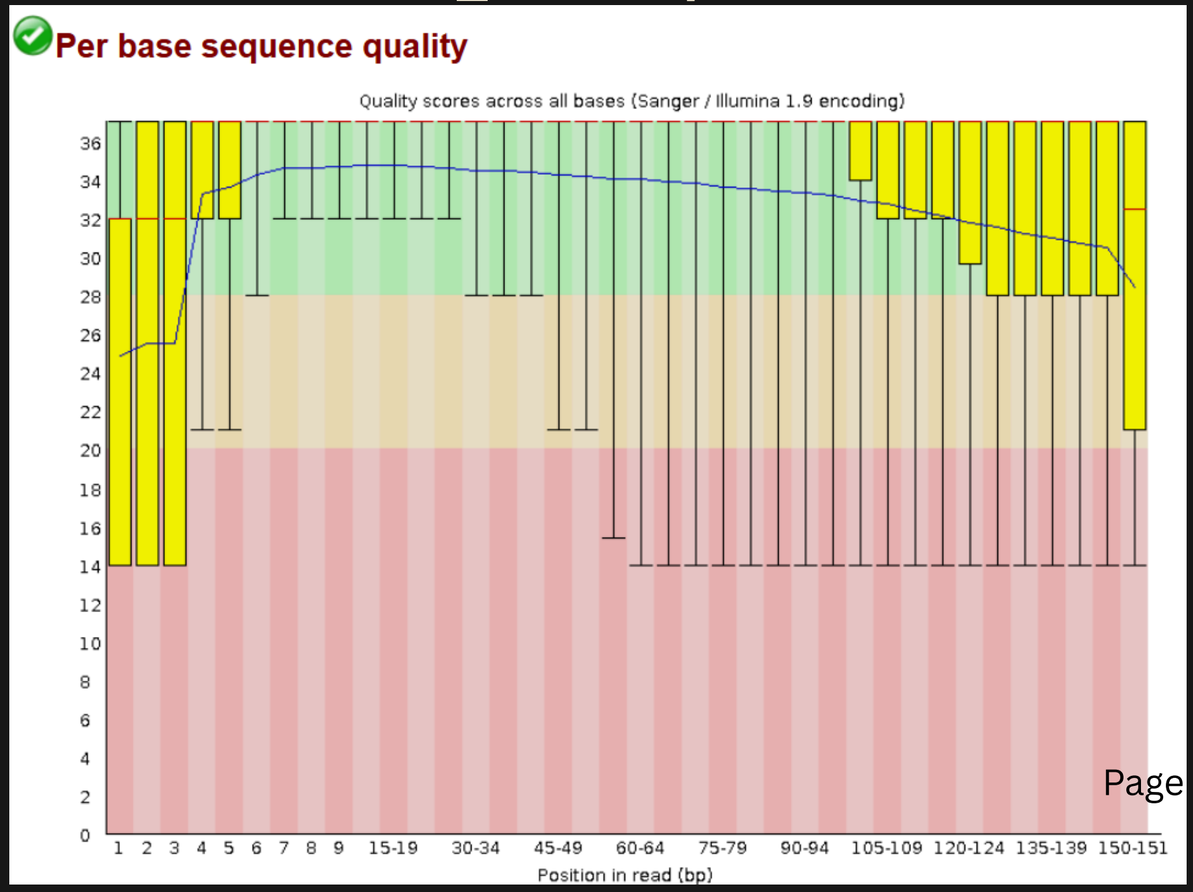
ORIGINAL SEQUENCES: QUALITY



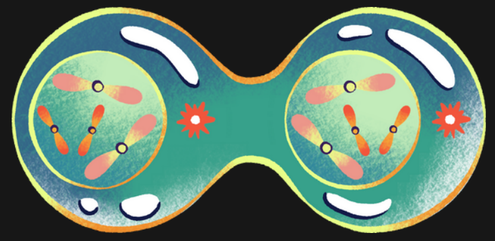
WH_R1.fastq



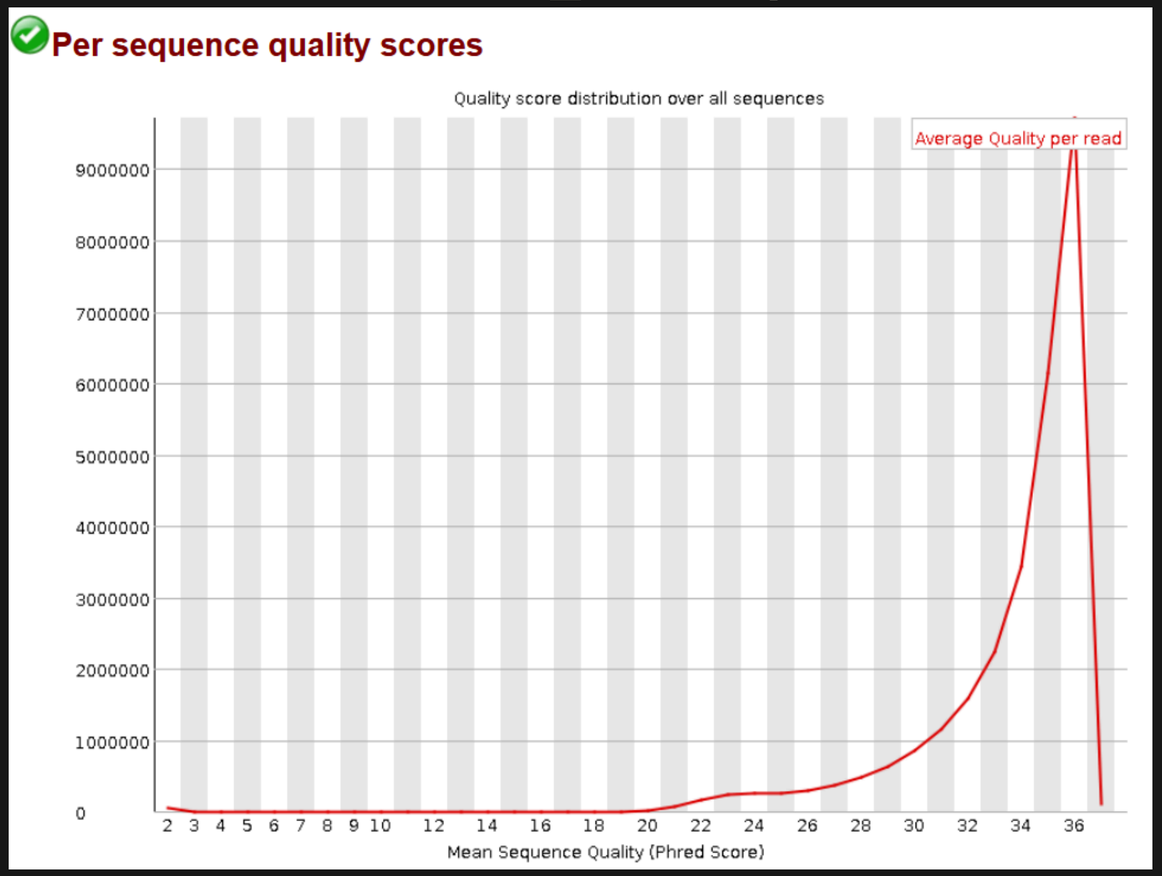
WH_R2.fastq



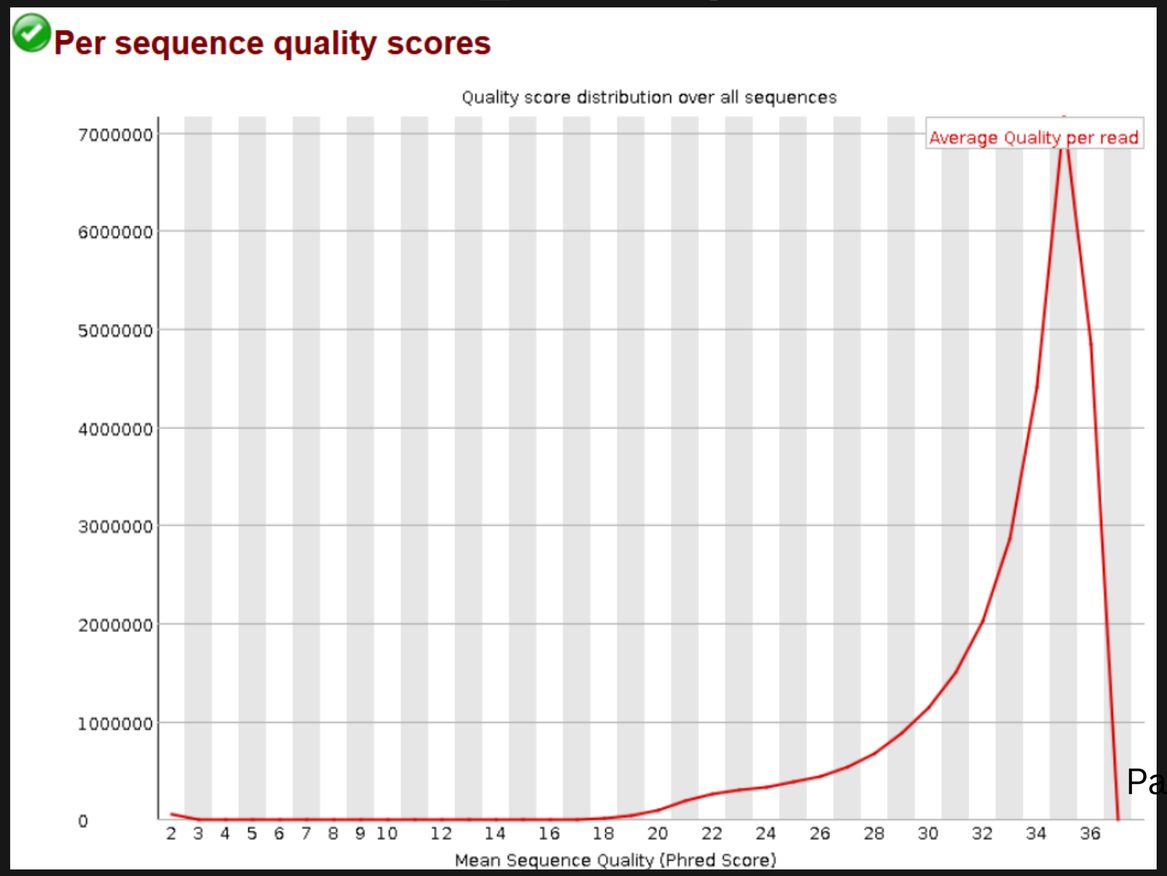
ORIGINAL SEQUENCES: QUALITY



WH_R1.fastq



WH_R2.fastq



TRIMMING PROCESS

HEADCROP:3 LEADING:3 TRAILING:3 MAXINFO:50:0.5 MINLEN:31

- **HEADCROP**

  Cut the specified number of bases from the start of the read

- **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

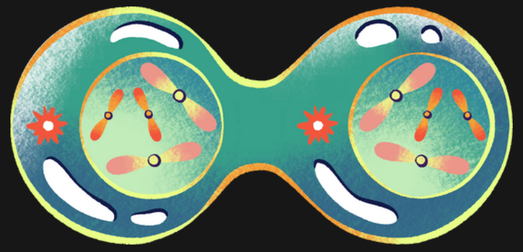
- **MAXINFO**

  Trims the end of reads by balancing length and quality

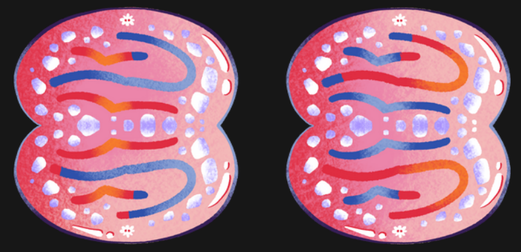
- **MINLEN**

  Drop the read if it is below a specified length

POST-PROCESSING SEQUENCES



WH_R1.fastq



WH_R2.fastq

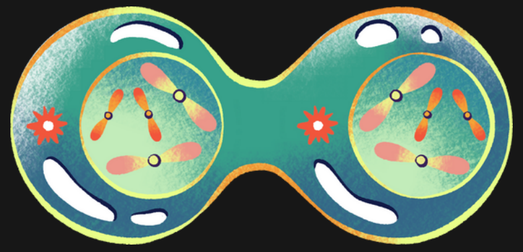
Basic Statistics

Measure	Value
Filename	R1_clean_MX3.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	28214303
Sequences flagged as poor quality	0
Sequence length	32-148
%GC	47

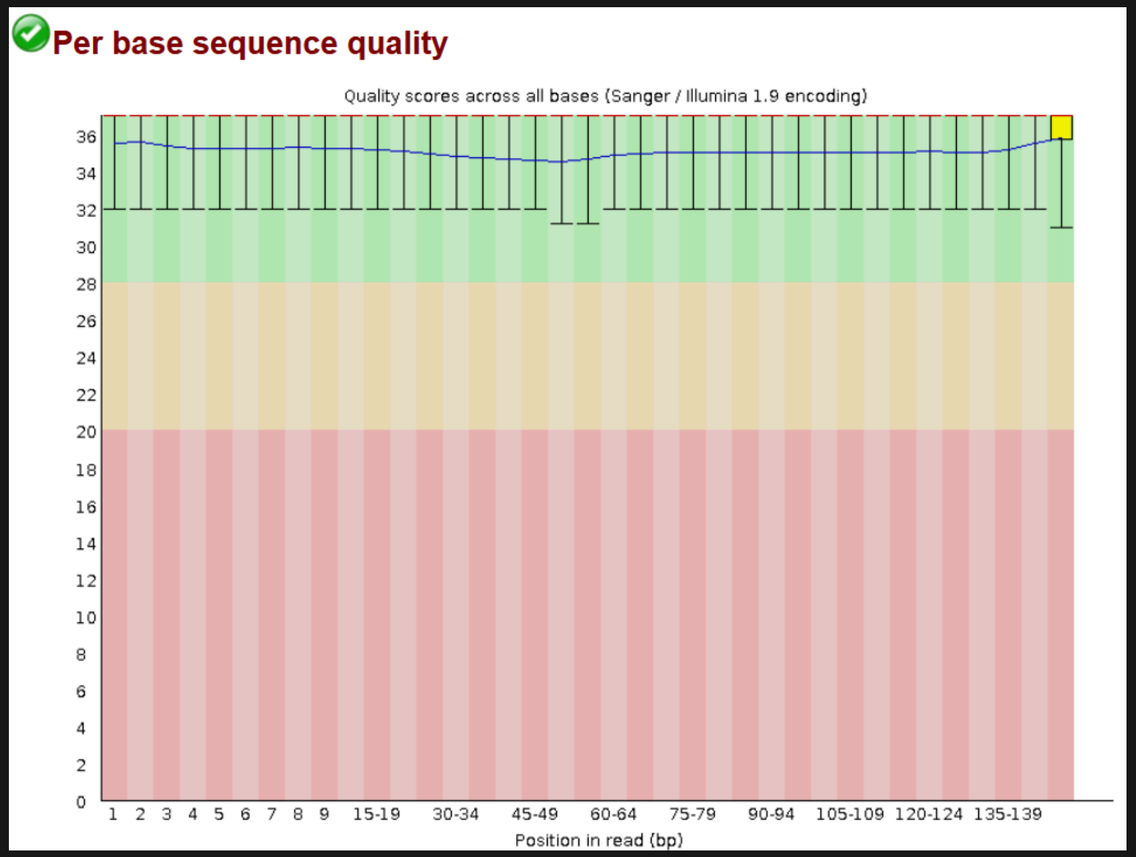
Basic Statistics

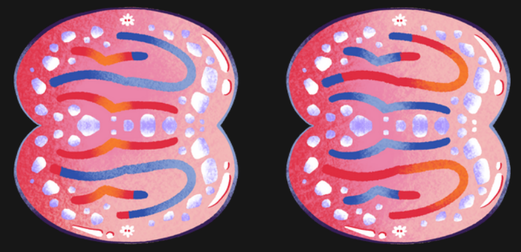
Measure	Value
Filename	R2_clean_MX3.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	28214303
Sequences flagged as poor quality	0
Sequence length	31-148
%GC	46

POST-PROCESSING SEQUENCES

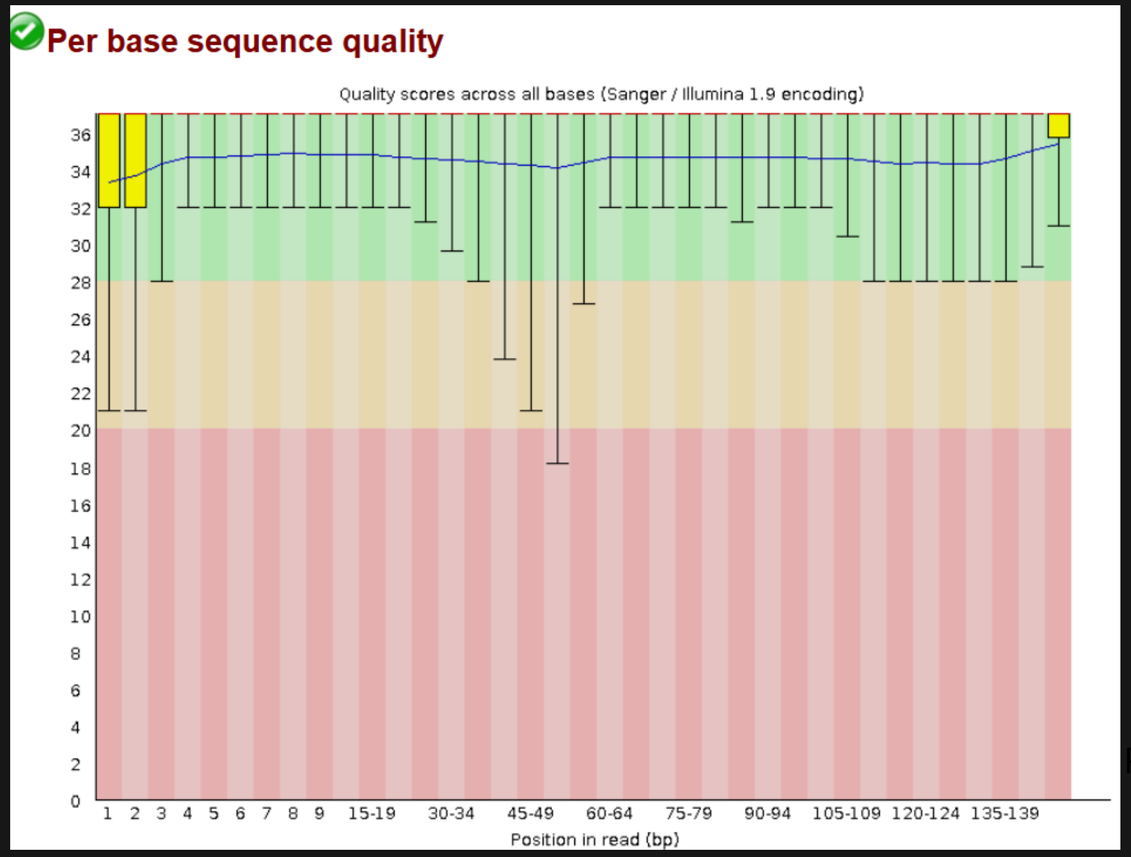


WH_R1.fastq

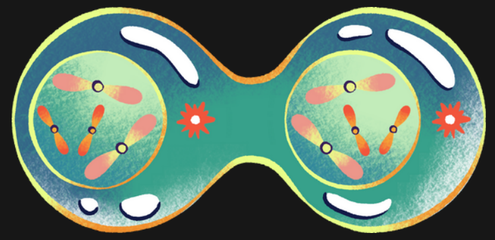




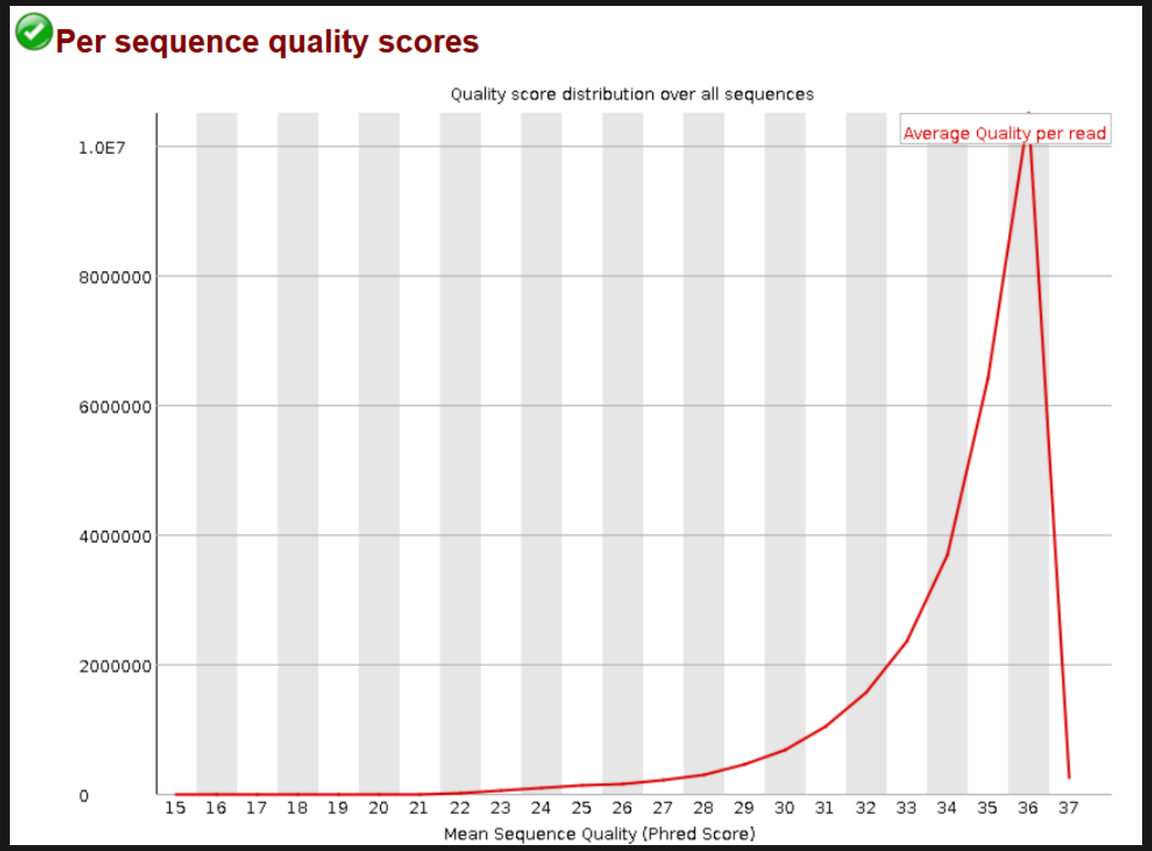
WH_R2.fastq



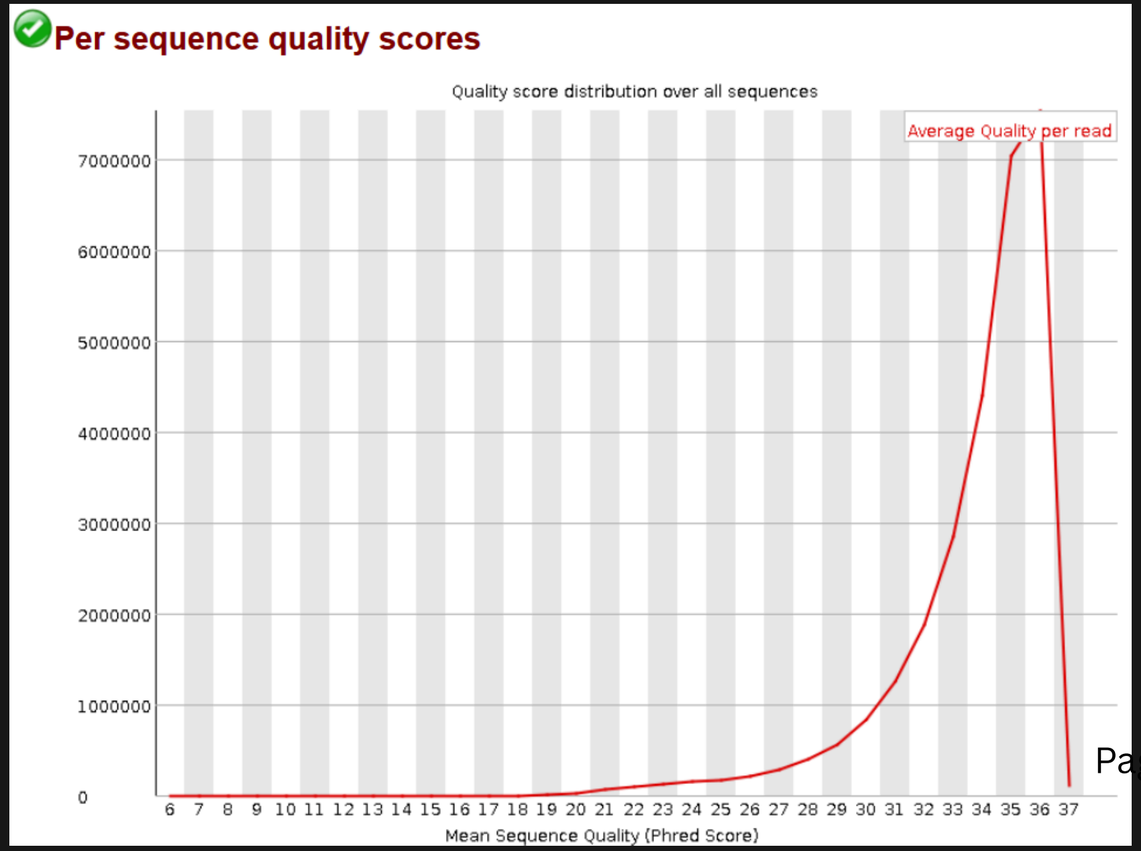
POST-PROCESSING SEQUENCES



WH_R1.fastq



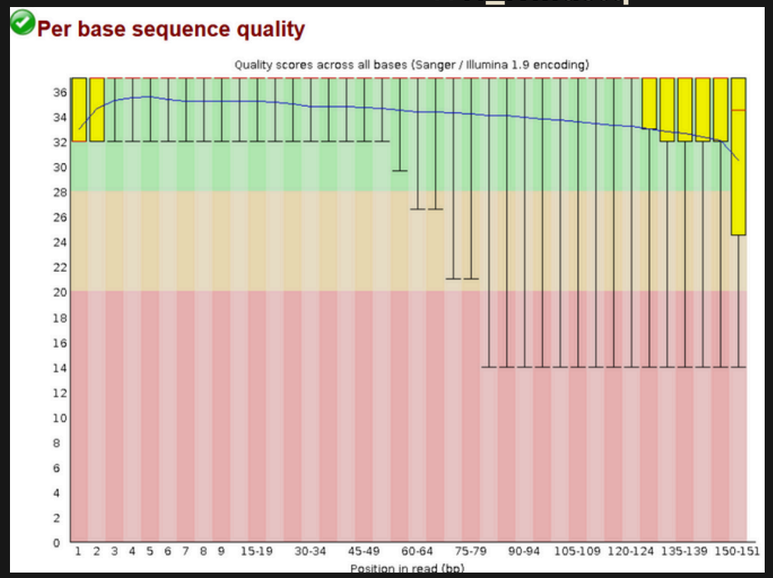
WH_R2.fastq



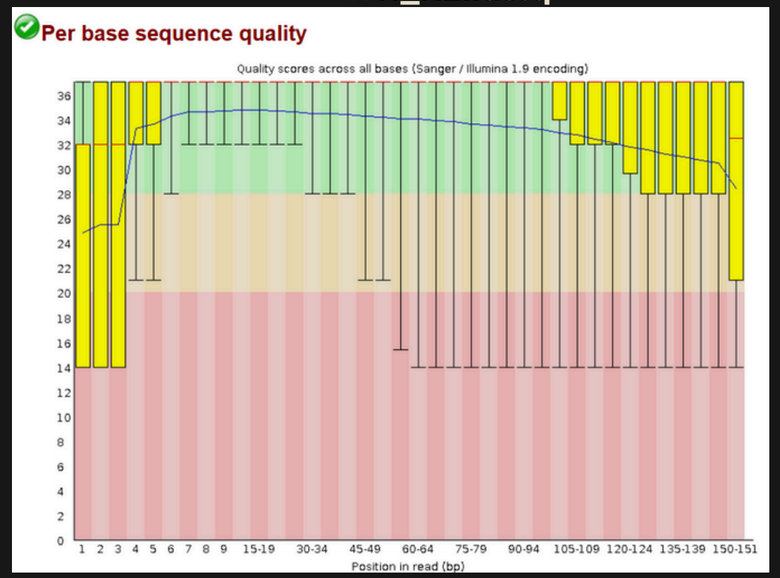
BEFORE AND AFTER

BEFORE

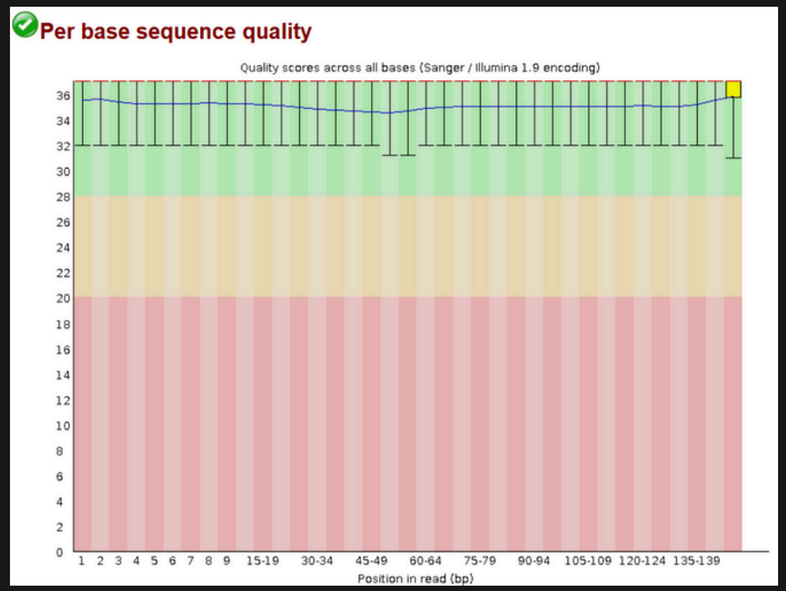
WH_R1.fastq

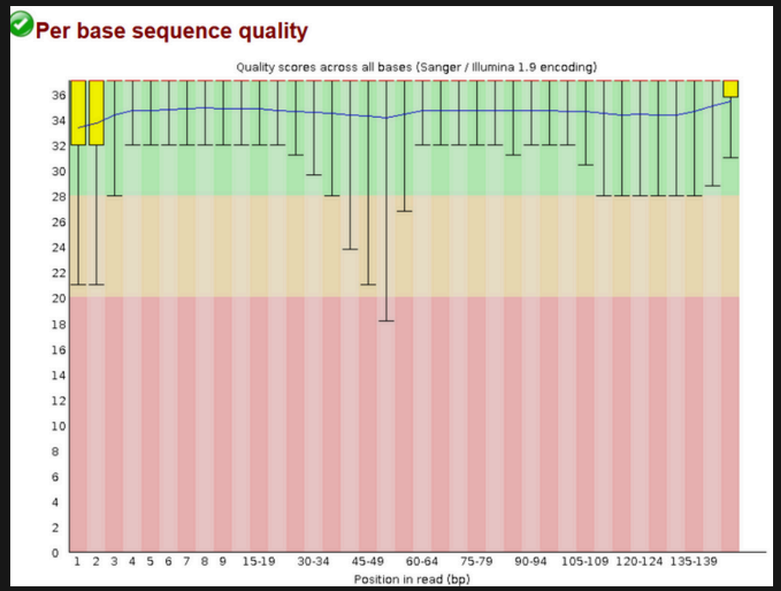


WH_R2.fastq



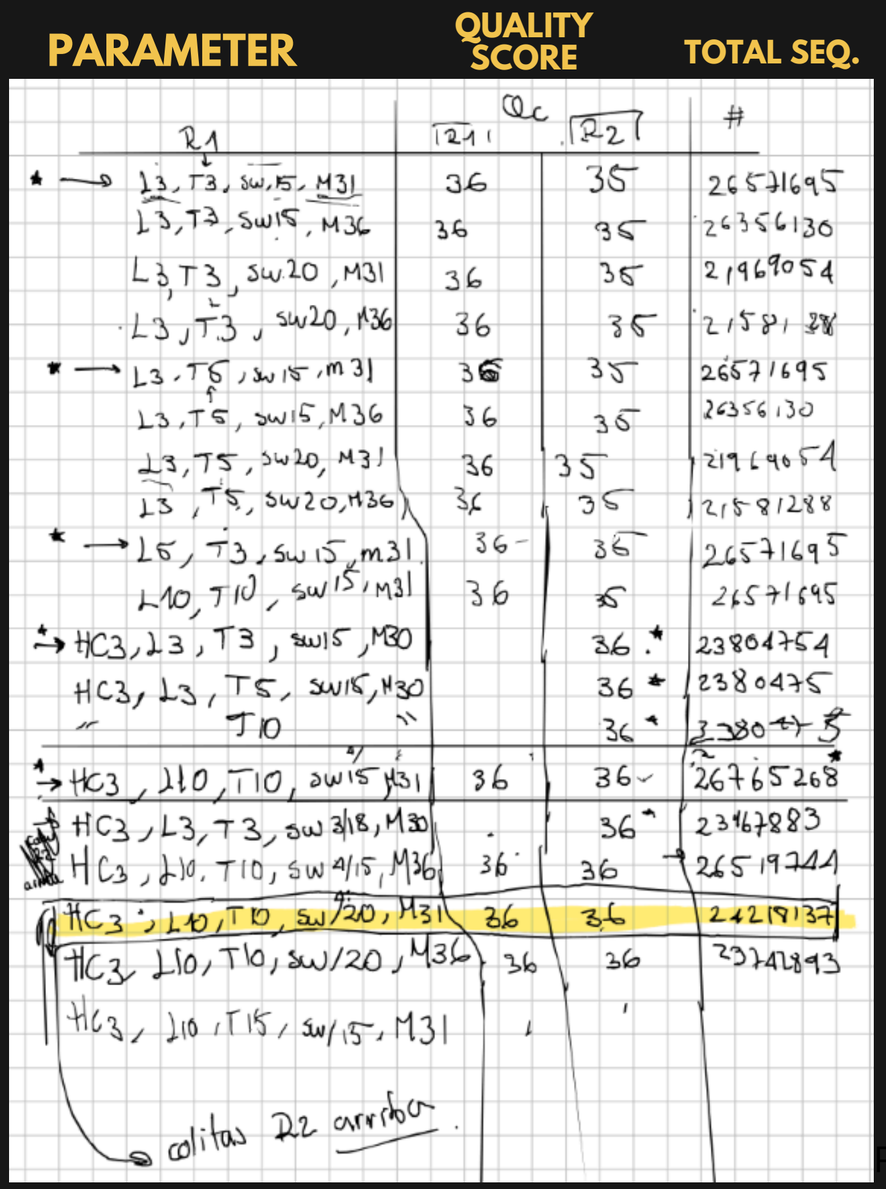
AFTER





ANOTHER COMPARISSONS WITH DIFF. PARAMETERS

```
# Archivos de entrada
R1="WH_R1.fastq"
R2="WH_R2.fastq"

# Archivo resumen
SUMMARY="summary_fastqc_results.csv"
echo "ID,Total Sequences,R1_Q30_Percent,R2_Q30_Percent,R1_path,R2_path" > "$SUMMARY"


# Iteraciones
for HEADCR in 3 ; do
        for LEAD in 10 15; do
          for TRAIL in 10 15; do
                for SW in 15 20; do
                for MINL in 31 36; do

                        ID="L${LEAD}_T${TRAIL}_SW${SW}_M${MINL}"
                        OUTDIR="results/${ID}"
                        mkdir -p "$OUTDIR/trimmed" "$OUTDIR/fastqc"

                        OUT_R1="${OUTDIR}/trimmed/R1_clean.fastq"
                        OUT_R2="${OUTDIR}/trimmed/R2_clean.fastq"
                        UNP_R1="${OUTDIR}/trimmed/R1_unp.fastq"
                    UNP_R2="${OUTDIR}/trimmed/R2_unp.fastq"

                    echo "🔧 Trimming: $ID"
                        TrimmomaticPE -phred33 "$R1" "$R2" \
                 "$OUT_R1" "$UNP_R1" "$OUT_R2" "$UNP_R2" \
                  HEADCROP:$HEADCR LEADING:$LEAD TRAILING:$TRAIL SLIDINGWINDOW:4:$SW MINLEN:$MINL

                echo "🔍 Running FastQC..."
                fastqc "$OUT_R1" "$OUT_R2" -o "$OUTDIR/fastqc"
```



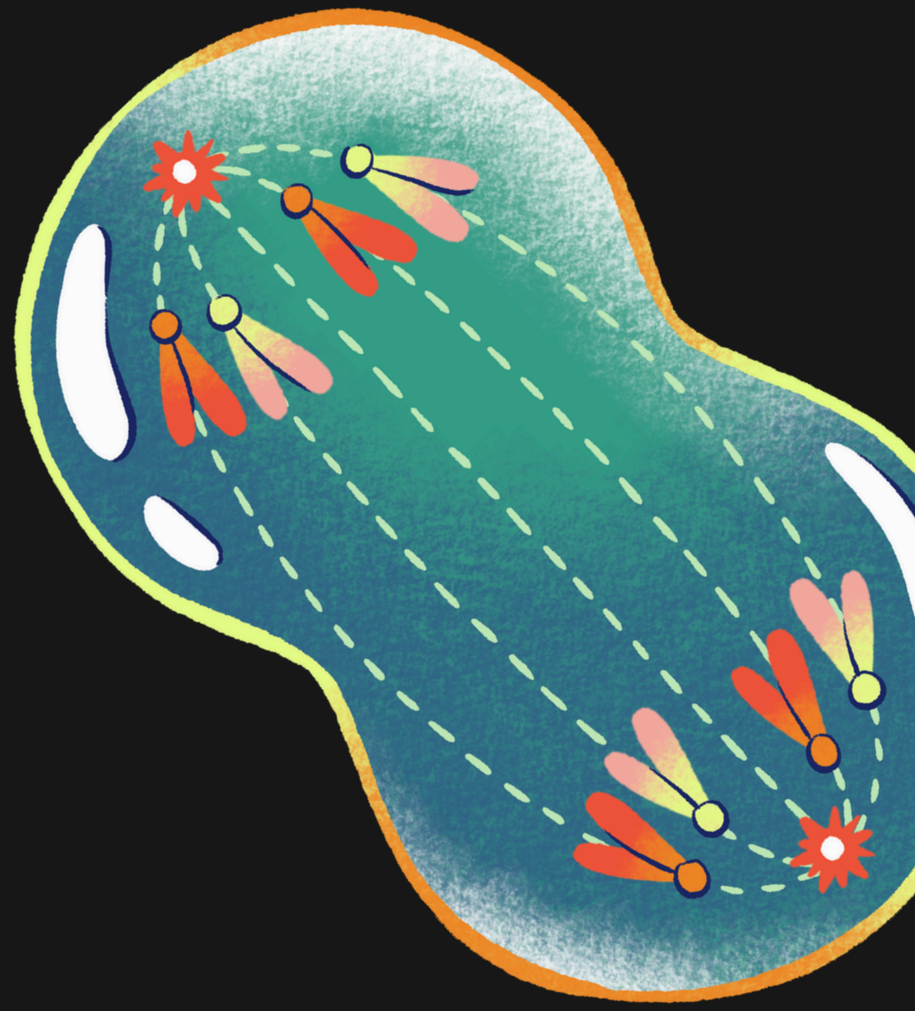
CONCLUSIONS

What modifications were accomplished with Trimmomatic?

- The Quality of R1 and R2 improves,
- The total of reading sequences removed was 0.24%

What tools can be utilized to automate or enhance the pipeline?

- Nextflow or SnakeMake for pipelines with multiple readings



BONUS: FASTP

```
fastp -i WH_R1.fastq -I WH_R2.fastq \
-o R1_clean.fastq -O R2_clean.fastq \
--detect_adapter_for_pe \
--cut_front \
--cut_tail \
--cut_window_size 4 \
--average_qual 20 \
--cut_mean_quality 25 \
--length_required 36 \
--correction \
--thread 4
```

✂️ Trimming y filtrado de calidad

Parámetro	Descripción
`--cut_by_quality5`	Corta bases de baja calidad al inicio (5') de la lectura.
`--cut_by_quality3`	Corta bases de baja calidad al final (3') de la lectura.
`--cut_window_size`	Tamaño de ventana para corte por calidad.
`--cut_mean_quality`	Calidad mínima promedio en la ventana para hacer el corte.
`--trim_front1`	Corta N bases al inicio de R1.
`--trim_front2`	Corta N bases al inicio de R2.
`--trim_tail1`	Corta N bases al final de R1.
`--trim_tail2`	Corta N bases al final de R2.
`--max_len1`	Longitud máxima permitida en R1.
`--length_required`	Longitud mínima para mantener la lectura (default = 15).
`--qualified_quality_phred`	Umbral de calidad phred para contar una base como "buena".
`--unqualified_percent_limit`	Máximo % de bases malas antes de descartar la lectura.